Deep Clustering of Cooperative Multi-Agent Reinforcement Learning to Optimize Multi Chiller HVAC Systems for Smart Buildings Energy Management

Raad Z. Homod[1*], Zaher Mundher Yaseen[2], Ahmed Kadhim Hussein[3], Amjad Almusaed[4], Omer A. Alawi[5], Mayadah W. Falah[6], Ali H. Abdelrazek[7], Waqar Ahmed[7], Mahmoud Eltaweel[8]

[1]Department of Oil and Gas Engineering, Basrah University for Oil and Gas, Basra, Iraq; raadahmood@yahoo.com

[2]Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia; z.yaseen@kfupm.edu.isa

[3]Department of Mechanical Engineering, University of Babylon, Babylon city, Iraq; ahmedkadhim7474@gmail.com

[4]Jonkoping University, Department of Construction Engineering and lighting science, Sweden; amjad.al-musaed@ju.se

[5]Department of Thermofluids, School of Mechanical Engineering, Universiti Teknolog Malaysia, 81310 UTM Skudai, Johor Bahru, Malaysia; omeralawi@utm.my

[6]Building and construction techniques engineering department, AL-Mustaqbal University College, Hillah 51001, Iraq; mayadahwaheed@mustaqbal-college.edu.iq

[7]Takasago i-Kohza, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia; ali_hassan80@siswa.um.edu.my & ahmed.waqar@utm.my

[8]School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, AL10 9AB, United Kingdom; m.eltaweel@herts.ac.uk

[*]Corresponding: raadahmood@yahoo.com (Raad Z. Homod)

Abstract

Chillers are responsible for almost half of the total energy demand in buildings. Hence, the obligation of control systems of multi-chiller due to changes indoor environments is one of the most significant parts of a smart building. Such a controller is described as a nonlinear and multi-objective algorithm, and its fabrication is crucial to achieving the optimal balance between indoor thermal comfort and running a minimum number of chillers. This work proposes deep clustering of cooperative multi-agent reinforcement learning (DCCMARL) as well-suited to such system control, which supports centralized control by learning of agents. In MARL, since the learning of agents is based on discrete sets of actions and stats, this drawback significantly affects the model of agents for representing their actions with efficient performance. This drawback becomes considerably worse when increasing the number of agents, due to the increased complexity of solving MARL, which makes modeling policy very challenging. Therefore, the DCCMARL of multi-objective reinforcement learning is leveraging powerful frameworks of a hybrid clustering algorithm to deal with complexity and uncertainty, which is a critical factor that influences to the achievement of high levels of a performance action. The results showed that the ability of agents to manipulate the behavior of the smart building could improve indoor thermal conditions, as well as save energy up to 44.5% compared to conventional methods. It seems reasonable to conclude that agents' performance is influenced by what type of model structure.

Keywords: Optimal chiller sequencing control (OCSC); multi-unit residential buildings; Clustering of multi-agent reinforcement learning (MARL) policy; Hybrid layer model; Takagi–Sugeno Fuzzy (TSF) identification; Multi-objective reinforcement learning (MORL).

Nomenclature

Symbols

A, a : set of action, agents' action

E_t, E_b, ED: peak total, diffuse, and direct irradiance, W/m^2

pr: pre-cooling coil valve position (open%)

OA, RA: the damper position in the AHU, (open%)

OF$_t$, OF$_b$, OF$_r$: opaque-surface cooling factors

Ma: main cooling coil valve position (open%)

Cp: specific heat, $J/kg.\ ^oC$

$\dot{m}$: mass flow rate, kg$/s$

Mcp: heat capacitance, $J/\ ^oC$

T: temperature, ^oC

□: thermal resistance, $^oC/W$

N_{oc}: number of occupants

N_{br}: number of bedrooms

α_{roof}: roof solar absorbance

τ: time constant, s

I: infiltration coefficient

$\Delta\omega$: indoor-outdoor humidity ratio difference, kg$_w$/kg$_{da}$

Subscripts

m: air in a mixing box

r: room/ return

ω: humidity ratio, kg_w/kg_{da}

□: latent heat/ heat transfer coefficient, J/kg, $W/(m^2.\ ^oC)$

$\dot{Q}$: cooling load, W

CF: surface cooling factor, W/m^2

U: construction U-factor, $W/(m^2.\ ^oC)$

ΔT: cooling design temperature difference, ^oC

DR: cooling daily range, ^oC

CF_{fen}: surface cooling factor, W/m^2

U_{NFRC}: fenestration U-factor, $W/(m^2.\ ^oC)$

PXI: peak exterior irradiance, W/m^2

SHGC: Solar heat gain coefficient

IAC: interior shading attenuation coefficient

FF_s: fenestration solar load factor

T_X: transmission of the exterior attachment

F_{shd}: fraction of the fenestration shaded by overhangs or fins

s: State

SLF: shade line factor

D_{oh}: depth of the overhang, m

X_{oh}: vertical distance from the top of the fenestration to the overhang, m

F_{cl}: shade fraction closed (0 to 1)

K: turbulent kinetic energy (m^2/s^2)

RNG: Renormalization group

ψ: exposure (surface azimuth), measured as degrees from the south

$\dot{V}$: volumetric flow rate, L/s

DF: infiltration driving force, $L/(s.\ cm^2)$

out: outside

os: outside supply

i: inside

He: heat exchanger

a: the air

w: water

aHe: air in the heat exchanger

L: leakage

W_{in}: water input

W_{out}: water output

Wl: wall

room: inside room

out: outside room

g: glass

fg: the heat of vaporization

Opq: opaque

inf: infiltration

fen: fenestration

t: time slot index

flue*:* flue effective

es: exposed

ul: unit leakage

ig: internal gains

l: latent

fur: furniture

cl: closed

1. Introduction

The growing rate of urbanization and the exponential increase of populations in the next couple of decades leads to an expansion in the building sector, particularly residential buildings [1]. The energy consumed by the global buildings sector is a public concern because of shows a significant amount of high growth in the usage rate over the years, which could cause severe effects on the environment [2]. The analysis of the residential building sector can be taken alone, where it is responsible for around 17% of global carbon emissions and 22% of energy consumption [3, 4]. In developed countries, the residential building consumes 48% of the total energy generated in such countries, and most of this energy is consumed by heating, ventilation, and air-conditioning (HVAC) systems [5, 6]. It is well known that chiller plants consume more than half of the total energy used in the HVAC system [7]. The highest coefficient of performance (COP) of chillers occurs at a full load, whereas the cooling load of the building changes hourly from time to time, so the single chiller system needs to operate at part load with a reduced COP for most of the time [8]. Therefore, the chiller's operation strategies of HVAC systems are highly effective in energy-saving based on their configuration, single or multi-chiller, and whether the multi-chiller is equal in size (symmetric) or different sizes (sequential) [9]. In general, the capacities of chillers in multiple-chiller systems are identical due to the convenience of control, installation, and maintenance [10]. The sequential or different sizes of multi-chillers are the best energy-efficient system due to allowing more alternative scenarios of choosing the appropriate chiller size to meet the current cooling load [11, 12]. The staging up and down of sequential chillers is essential to running an optimal number of chillers to meet the cooling load. Accordingly, optimal chiller sequencing control (OCSC) provides evidence of potential energy savings and reduced emissions to protect the environment [13]. A programmable logic controller (PLC) is a pioneer in managing the sequencing of the multi-chiller to control chilled water temperature [14]. The previous studies reported that the chiller sequencing control by PLC needs a state machine algorithm to define which chillers are online or offline, according to four methods to indicate the cooling load: return chilled water temperature, direct power, bypass flow, and cooling load [15, 16]. In addition, the transition state to reducing or adding the number of operating chillers by the state machine needs a dead band to reduce the switching frequency around which the chillers

be online or offline. The drawbacks above are among the significant challenges facing implementing chiller sequencing control. Furthermore, the conventional multiple chiller plants can be managed by the sequencing control method, which has limitations such as not providing the optimal load distribution and running an inappropriate number of operating chillers [17, 18]. Some researchers are intended to address these limitations by implementing the model predictive control (MPC) [18]. Although the PMC is robust to both disturbances and time-varying parameters and coefficient of performance (COP) improvements, it needs to identify the system's proper model, which can be challenging [19]. Thus, the MPC involves a complex model to get high-quality nonlinear model fit and expends a lot of time for more calculations [20].

Intelligent systems, such as fuzzy logic control, can be used to achieve optimal chiller sequence operation [21]. Such optimization is of great importance for the operation of the multi-chiller plants to ensure a reduction in operation time and cooling energy generation and thus should force the sequencing control to save energy. The Mamdani fuzzy inference system (MFIS) can satisfy these competing goals [22]. In this regard, Chen et al. [23] verified the effectiveness of the MFIS in causing a significant reduction of total energy consumption compared to the PLC conventional system. However, obtaining such a fuzzy inference system is challenging, particularly for antecedent and consequent parameters of rules, since the dynamic process of HVAC systems of such systems is complex [24]. Another challenge with MFIS for ensuring multiple ambiguous crisp outputs is the associated parameter tuning during the training process [25].

The challenges for conventional OCSC are to provide optimal load distribution over all chillers and set an inappropriate number of operating chillers [26]. In recent years reinforcement learning (RL) has emerged as one of the most techniques to deal with such challenges and optimization methods to reduce the energy consumption of the plant chiller. The RL is demonstrated to be an alternative to rule-based decision-making in energy management and does not require a tuple fuzzy linguistic representation model [27]. RL, a subfield of machine learning, aims to use collected data to determine the best course of action. When developing a policy model, RL can be divided into two subfields of unsupervised machine learning: model-based and model-free [28]. Although the broad classification of model-based, and model-free show unique advantages, both are needed for offline learning into the optimization phase, which is quite different from traditional MPC. The model-free combines the information from the environment with previous estimates or beliefs about state values. Therefore, it is less statistical efficiency than the model-based, which is used the information directly [29].

One significant challenge when using model-based RL is representing the states of the environment defined by dynamic models to obtain datasets passed to the training process [30]. During the initial training period, the agent can improve its policy on the tasks even if the stats are held constant and this period can be extended when increasing the number of states [31]. Regardless of the training period and time-consuming, it is essential to robustly solve challenging decision problems during the training and operation periods. Obtaining tabular model data sets of each agent's policy requires a model of policy whereby the agent's action decisions are represented. Several technologies can represent the policy model, some of which have demonstrated exemplary performance on simulated policy data [29].

The neural network deep learning is one of the significant literary involved in the black-box model to represent an agent's policy [32]. Such a model is called deep reinforcement learning (DRL), which comes to handling large state spaces [33]. The DRL is a category of advanced machine learning techniques that enable agents to learn from past actions, which are generated based on rewards or penalizes. The combined deep learning (DL) with RL provided high-dimensional action spaces or continuous actions, making a new structure able to solve much more challenging problems [34]. And such a model agent structure has shown great potential for sequential decision-making. Thus, enabling the algorithm to react and change the staging of sequential chillers profitably to follow the fluctuating cooling loads of the building. The application of DRL that runs on OCSC shows effectiveness in performance compared to conventional controllers [35].

The multi-input multi-output (MIMO) system uses many agents to interact with each other and the different operating environments. The numerous mechanical and electrical systems in buildings can be controlled by multi-agent. Integrating multi-agent in the building is linked to the increase of energy efficiency and cooperated agent of chillers is one of the most important goals for achieving energy saving [36]. This will disable the DRLs while the number of agents is increased due to increased data of the environment and the consequent degradation of agents' performances [37]. The best representation learning model is called the clustering method, where no new technique can outperform the representation of agents' policies in the clustering approach [30].

From the literature can be concluded that the case of running an HVAC system under part-load conditions needs sequential features for types of the multi chiller to increase operational feasibility and COP by using multi-objective reinforcement learning (MORL) [38]. In addition, to achieve energy-saving and improve occupants' satisfaction in intelligent buildings, the multi-agent system approach is essential for tackling the complexity of control buildings. Many studies were carried out to solve intelligent building operations by subdividing them into smaller tasks that are controlled autonomously entities, known as agents [39]. Based on how agents are incentivized, there are three broad categories of multiagent systems (MAS): cooperative, competitive, and mixed (agents compromise with their reward by enabling both sides to cooperative and competitive mode based on conditions). When developing a policy model, RL may be broken down into two subfields of unsupervised machine learning: model-based and

model-free [40]. The recent work by Pinto et al. [41] proves that harmonious architecture outperforms others in terms of energy saving and cost.

In this study, the integration of deep clustering into RL for the cooperative multi-agent policy is applied to deal with a large amount of training data when state and action spaces are continuous. The objective of the proposed deep clustering of cooperative MARL (DCCMARL) is to train hybrid layers of multiple agents based on a clustering algorithm to obtain the redoubled goal. The proposed clustering structure of a cooperative of various agents can handle both ultra-large data volumes of sets and actions due to such design allowing the distribution of more spaces and flexibility to the storage of learning operations than traditional DRL. The proposed strategy DCCMARL has adopted by using the varied capacity of multiple-chiller to make energy-saving the main objective. Furthermore, the DCCMARL tackles two tricky problems associated with conventional controllers: First, decoupling chillers operation, where the water loop is decoupled into a primary chiller loop and secondary air handling unit (AHU) loop by a bypass line; second, the conventional controller needs to tune the dead-band to avoid chattering of the switch on/off as shown in Fig. 1. It should be noted that the dependent variable of thresholds is the time axis in Fig. 1 which is a function of the instantaneous cooling load of the building and thus leads to switching on/off of the chillers.

The main contribution of the proposed DCCMARL in the study of RL is to handle massive amounts of data sets and reduce the convergence time. This is achieved by implementing a TSF-based multi-agent policy for the sequential control of three sequential chiller techniques and other building appliances, a difficult task for conventional controllers to handle. Furthermore, adopting a hybrid clustering algorithm based on the reward of the Markov Decision Process (MDP) makes it focused on fast and accurate decisions. Moreover, the optimization algorithm of nonlinear least squares regression (NLSR) significantly helped to cooperate effectively on developing policies of heterogeneous MARL (each agent has different input and output target). So, the main contributions of this work can be summarized as follows:

1- The HVAC control system is formulated based on MDP, with definitions of RL terms (state, action, and reward) to be data-driven, which requires learning a DCCMARL.
2- The structure of a design technique of DCCMARL is strongly emphasized in the adoption of capturing a large number of datasets, which allows the multi-agent to handle dynamic large-scale domains of multi-action.
3- The NLSR is used to optimize the parameters and weights in a hybrid network structure to represent agents' policies.
4- The multi-agent systems (MAS) are integrated into a cluster model-based to save energy for the HVAC system's AHUs and multi-chiller.

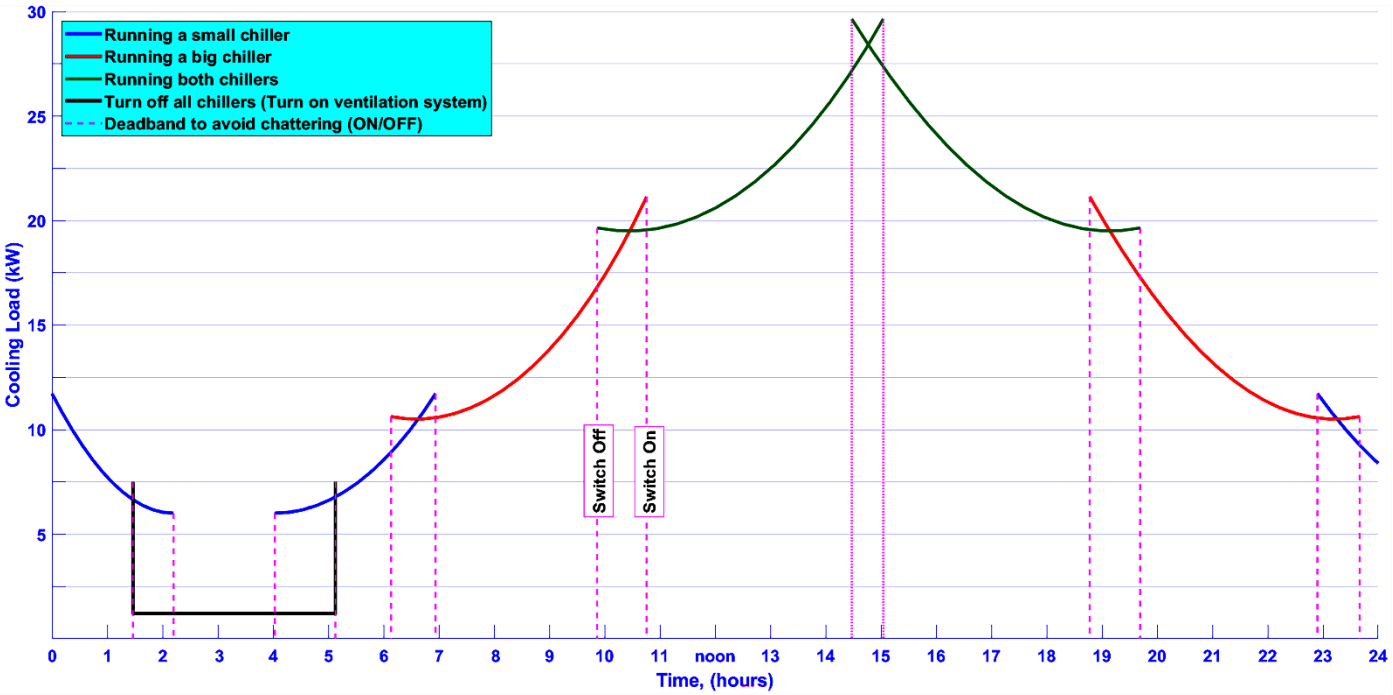

Fig. 1. Multiple-chiller sequencing principle for a decoupled loop.

2. Parametric Design of Modelling and Problem

Following a problem statement that is well-structured based on the Markov decision process (MDP), this part attempts to specify the combination of AHUs and multi-chiller plants to construct an HVAC system hygrothermal model that is easily combined with the building model.

2.1. *Whole-building Hygrothermal Modeling*

The hygrothermal design model of the HVAC systems and building can assess precisely the response of indoor conditions in terms of thermal comfort and moisture loads to provide a set of RL environment states. The dynamic environment states (state transition) are essential to keep active learning in model-based reinforcement learning (MBRL). The dynamic states of whole-building (chillers, AHUs, and building structure) are affected by internal and external environmental factors. In order to face unexpected disturbances in both factors, the previous works divided the hygrothermal model into six major sub-models due to a large number of variables being required to represent such a model, as described by Eq. (1) [42-46].

$$\overbrace{[\dot{X}] = f(X,t) + G(X,t)\begin{bmatrix} k_2 \\ T_o(\mathrm{t}) \\ \mathrm{T_r} \\ f_{DR} \\ A_{slab} \\ k_3 \\ f_4 \\ \omega_o(t) \\ \dot{Q}_{ig,l} \end{bmatrix}}^{The\ Building\ Structures\ State\ Equation} \quad \overbrace{[\dot{X}] = f(X,t) + G(X,t)\begin{bmatrix} \dot{m}_{os} \\ \dot{m}_r \\ \dot{\mathrm{m}}_{\mathrm{w}} \\ T_o \\ \dot{\mathrm{m}}_{\mathrm{mw}} \\ \omega_o \\ T_r \\ k_2 \\ A_{Slab} \\ F_{Dr} \\ f_4 \\ \dot{Q}_{ial} \end{bmatrix}}^{The\ Whole\ HVAC\ State\ Equation} \quad \overbrace{\begin{bmatrix} T_{rr}(t) \\ T_r(t) \\ RH_r(t) \\ v_{ar}(t) \\ T_{o\,chil}(t) \end{bmatrix} = h(X,t)}^{Output\ Equation} \quad (1)$$

where the vector of $X \in R^n$ is assigned to the states, $T_r(t)$ and $T_{rr}(t)$ are the model outputs variables of indoor temperature and indoor radiant temperature respectively, ${}^oC$, $RH_r(t)$ is the model output of indoor relative humidity, %, $v_{ar}(t)$ is the model output of indoor air velocity, m/s, $T_{o\,chil}(t)$ is the model output representing cooling water output temperature, ${}^oC$, and the inputs and disturbance variables are $T_o(t)$ is outside a building temperature, ${}^oC$, $\mathrm{k}_2 = \frac{\sum_j A_{w_j} U_j OF_b + \sum_j A_{w_j} U_j OF_r DR}{\sum_j A_{w_j} U_j OF_t + \sum_j A_{w_j} h_{i_j}}$: disturbance factor because of variation in the incident solar radiation influence on the building and thermal resistance of the walls and ceiling against heat conduction, $\mathrm{k}_3 = \frac{\sum_j A_{w_j} h_{i_j}}{\sum_j A_{w_j} U_j OF_t + \sum_j A_{w_j} h_{i_j}}$: the disturbance factor of dynamic convection heat transfer and thermal resistance to heat flow by convection, A_{slab} is an influencing factor in slab floor area, $f_3 = C_s \times A_L \times IDF + \dot{m}_{ven} cp_a\ (W/k)$: the ventilation and infiltration factors of sensible cooling load added to the indoor space by disturbance factor of outdoor temperature variation, f_{DR} is a location factor that affects climate by focusing on existing wind patterns, topography, deforestations, elevation/altitude, etc., ω_o : the disturbance factor in the surrounding environment because of the humidity of the air outdoor ($\frac{Kg_w}{Kg_{da}}$), $f_4 = f_{fen} + 136 + 2.2A_{cf} + 22N_{oc}(W)$: the disturbance factor from an indoor generated gain related to sensible cooling load sources such as radiation emitted by the occupants, illumination, and indoor equipment, and $\dot{Q}_{ig,l}$: the disturbance factor of latent heat emitted by indoor sources (people and equipment), (W).

The overall transfer function for a system composed of subsystems in relating inputs to outputs is the product of the transfer functions of the individual subsystems elements, where the product variables (outputs) are inputs to the predicted mean vote (PMV), as shown in Fig.2. The PMV enables to evaluate of indoor conditions, from the outputs of the overall transfer function in Fig.2, it is possible to obtain $RH_r(t)$ since $\omega_r(s)$ is a function of time for output equation $h(X,t)$ by $RH_r(t)$. Therefore, the output equation $h(X,t)$ signal passed through the inputs of the PMV transfer function can be expressed as Equation (2) derived by [47, 48].

$$PMV = TF\begin{pmatrix} T_{rr}(t) \\ T_r(t) \\ RH_r(t) \\ v_{ar}(t) \end{pmatrix} \quad (2)$$

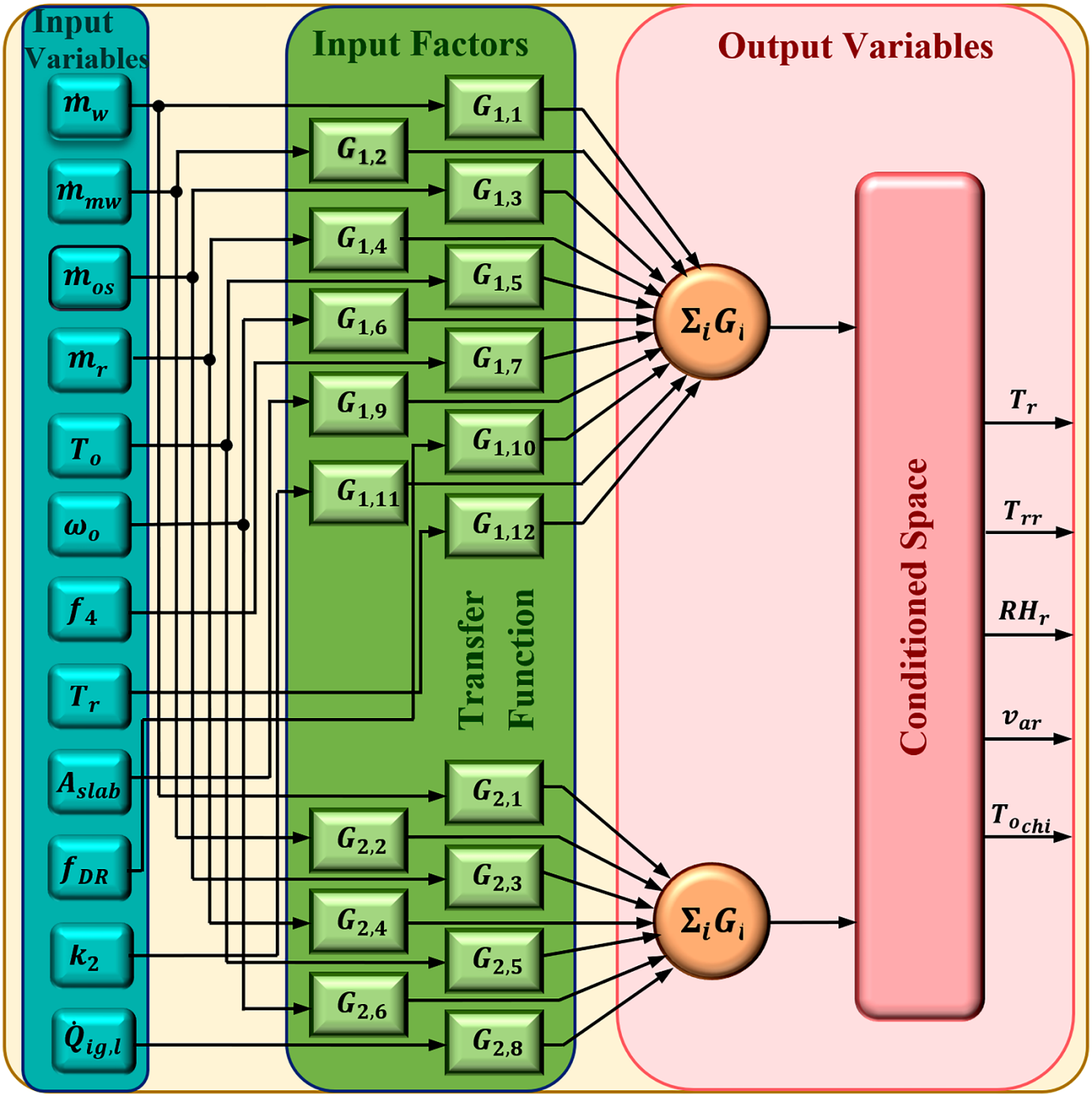

Fig. 2. Illustrates the schematic representation of the interactions between sub-systems of the whole building (chillers, AHUs and building structure).

2.2. *The framework of multi-agent RL*

The multi-agent of a building learns the values of actions in sequential decision-making by interacting with its environment to save energy while maintaining the desired indoor conditions. The decision of an intelligent agent can be predicted based on addressing the dynamic scheduling (exploration) which is directly influenced by current building stats and various factors that interfere with environmental disturbance. The following building stats (room humidity and temperature and outlet chilled water temperature) are possibly obtained after carrying out a step in the experience batch (tuples) sampled during MDP training, then will pass into successor states. In the MDP framework, transition probabilities to the following states (s') depend only on the current state (s) and not on the history of the past changes for states (stationary policies) due to each current state at the experience tuple is collected all relevant data on the previous state. The transition probability matrix (P) defined in Eq. 3 has the probability distribution based on a discrete mathematical formula, which can represent each given condition of whole building stats. In this

matrix, the model has systemized mapping from state-action pairs to the next states ($S_{t+1} = P(\mu_{st,at},\ \theta_{st,at})$), where $\mu_{st,at}$= set of transition function parameters and $\theta_{st,at}$= mean function.

$$p_{SS'} = \begin{array}{c} \\ S_{1,t} \\ S_{2,t} \\ \vdots \\ S_{n,t} \end{array} \begin{array}{c} \begin{array}{cccc} S_{1,t+1} & S_{2,t+1} & \cdots & S_{n,t+1} \end{array} \\ \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,n} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ P_{n,1} & P_{n,2} & \cdots & P_{n,n} \end{bmatrix} \end{array} \tag{3}$$

where $p_{SS'}$ is the probabilities matrix of transition from one state to another, S is denoted by S_t and S' is denoted by S_{t+1} where S and S' represent the current state and the next state, respectively.

The unsupervised control of chillers, AHUs, and building structure can be modeled as an MDP with an effective total rewards strategy (under the energy-saving criterion and recommended indoor conditions) and solved through machine learning techniques. Thus, the dynamic RL algorithm is formulated as MDPs of sequential decision-making to enable agents in such domains of the whole building to evaluate action policies. In general, the finite MDP is clearly described by an ordered repeating set consisting of 5-tuple (S, A, P, R, γ) specifies what actions should be taken from each state, where $s_t \in S$ is the states of inside/outside air condition and chillers outlet temperature at time-dependent t, $a_t \in A$ is a finite set of actions available in each state at time-dependent t, P ($s_{t+1}|s_t, a_t$) is called the transition probability of the process moving from state (s) to state (s') after the agent observes current states and picks actions, r (s_t, a_t) $\in R$ is a reward obtained by the agent from that particular state at time step t, and $\gamma \in [0,1]$ is a notion of the discount factor.

The agent in MDP needs to have the ability to get its optimal policy; RL is one of the most important tools of machine learning used to create an optimal policy model based on data gathered from an environment. The agents' objective is to identify actions that maximize the discounted sum of future rewards by accumulating over time, this return estimated by the value function of a state $V(s)$. To solve the approximate $V(s)$ in an iterative algorithm, Bellman's equation is the most suitable and crucial tool for obtaining the optimal value of the $V(s)$. To formulate a recursive and explicit function of arithmetic sequences, we need to consider two factors: the first represents an accumulated reward over its sequence of actions, and the second is related to the value of the following state-action value. Researchers can express this in two ways: an algebraic equation (4) or a square matrix equation (5).

$$V(s) = R_s + \gamma \sum_{s' \in S} p_{SS'} V(S') \tag{4}$$

where R_s is a scalar reward representing the actual return to an agent while transitioning from state (s) to state (s'), $V(S')$ is a value function of the next state, γ is a discounted future reward.

$$\begin{bmatrix} V(1) \\ V(2) \\ \vdots \\ V(n) \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{bmatrix} + \gamma \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,n} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ P_{n,1} & P_{n,2} & \cdots & P_{n,n} \end{bmatrix} \begin{bmatrix} V(1) \\ V(2) \\ \vdots \\ V(n) \end{bmatrix} \tag{5}$$

The cooperative multi-agent systems (CMAS) are interacted with each other and distribute decision policies for agents according to the current states to coordinate their actions; these manipulative actions techniques involve changes in the building's ventilation, indoor temperature, OCSC, lighting, and so on. The main parameters that dynamically affect manipulating stats are close/opening windows/doors, dimming/Brightening or turning On/Off lights, OCSC, adjusting the chilled water valve position, etc. By such sets of actions, the learning agent happens during a series of discrete-time. At every timestep t, the agents send these actions $\{U = u^1_{closWin}, u^2_{openWin}, u^3_{onLigh}, \dots u^m_{openVal}\}$. In physical space, the agents perform their actions in sets $A = \{a^1,\ a^2, \dots, a^p\}$ of the OCSC, including all manipulated parameters to achieve energy saving by learning policies of MAS using the multi-objective optimization algorithm, i.e., $p = m^{no.Equ}$. The equation p (a finite set of sampled MAS actions) is polynomially growing up rapidly according to the number of state variables. This suggests that the performance of agent actions degrades significantly as the size of the stats dataset increases due to the exponential growth of the mathematical space of action sets A. Although RL is one of the most suited methods for resolving MDP and making optimum decisions based on data sampled from the environment, multi-agent actions deteriorate their effectiveness as the nonlinearity of state space in an environment and the number of states increase. Furthermore, the Bellman optimality equation provides a recursive formulation to obtain the optimal value function in an MDP, applying a Bellman equation in the optimal $V^*(s)$ given by Eq.6.

$$V^*(s) = \max_a \left(R_s^a + \gamma \sum_{s' \in S} p_{SS'} V^*(S') \right) \quad (6)$$

The scalar reward (or cost) functions that species the reinforcement signal generated from the environment in each state to evaluate the multi-policy scenario of the MDP. The mathematical framework for modeling decision-making is long-term described in a tuple of five elements (S, A, P, R, γ) based on the reward of MDP, which can be expressed as connection weights adjusted according to time, where the highest optimal value of the rewards is the incumbent solution at time t. The tuple of MDP adopts a MORL algorithm that needs to tune its elements by training the multi-agent according to the reward function, which enforces a priority level based on the agent in order, thus achieving a sufficient level of precision in path-finding. After describing the relationship of the reward function and then balancing the trade-off between energy consumption and thermal comfort, the following Eqs. 7, 8 described the feedback reward of the agent of the chilled water valve.

$$ON = \begin{cases} 0, & if\ T_o \leq \left(\frac{T_{Min}^{set} + T_{Max}^{set}}{2} \right) \\ 1, & otherwise \end{cases} \quad (7)$$

$$R(ON, \dot{m}_{ch}, T_r, RH_r) = -ON * \gamma \left[\beta \dot{m}_{ch} + \tau \left(\frac{2T_r - T_{Min}^{set} - T_{Max}^{set}}{2} \right)^2 + \omega \left(\frac{2RH_r - RH_{Min}^{set} - RH_{Max}^{set}}{2} \right)^2 \right] \quad (8)$$

where the $ON(T_o)$ the signal is a function of the outdoor temperature used to turn off all chillers ($ON = 0$); thus, when the outdoor temperature is less than the upper dead-band desired temperature, the agents are given a reward $R(ON, \dot{m}_{ch}, T_r, RH_r)$ by the environment, based on conditions states (T_r, RH_r, T_o) and energy consumption state ($\dot{m}_{ch}$) which is the chilled water flow rate (kg/s), γ : is a discounted reward over time ∈ [0, 1] was taken 0.99 in this work, $\omega, \tau,$ and β: are the weight of indoor thermal conditions and power consumption, were taken 0.3, 0.3, and 0.4 respectively, the set points of upper and lower dead-band condition operation temperatures and relative humidity are: T_{Min}^{set}= 20°C, T_{Max}^{set}= 24°C, RH_{Min}^{set}= 45%, and RH_{Max}^{set}= 55%.
The goal of agents is to collect the return (rewards) over an infinite horizon when the following policy π, the optimal policy is achieved when iterative learning seeks to maximize value for states, and both V and π are marked by an asterisk (*). The convergence of iterative RL algorithms to an optimal value function $V^* = V^{\pi^*}$ depends on the strategy for the transitions; this leads to getting optimal action that an optimal policy $a^* = \pi^*$. In other words, the RL algorithm for finding the optimal policy of an MDP leads to maximizing the v value over time, as defined in Eq. 9.

$$\pi^*(a|s) = \underset{a \in A}{\operatorname{argmax}} V^*(S') \quad (9)$$

The sequence of policy updates is carried out using a whole range of Eqs. 6, 9. Subsequently, the generalization of two fundamental operations in an iterative manner is policy improvement and policy evaluation, as illustrated updates trajectory in Fig. 3.

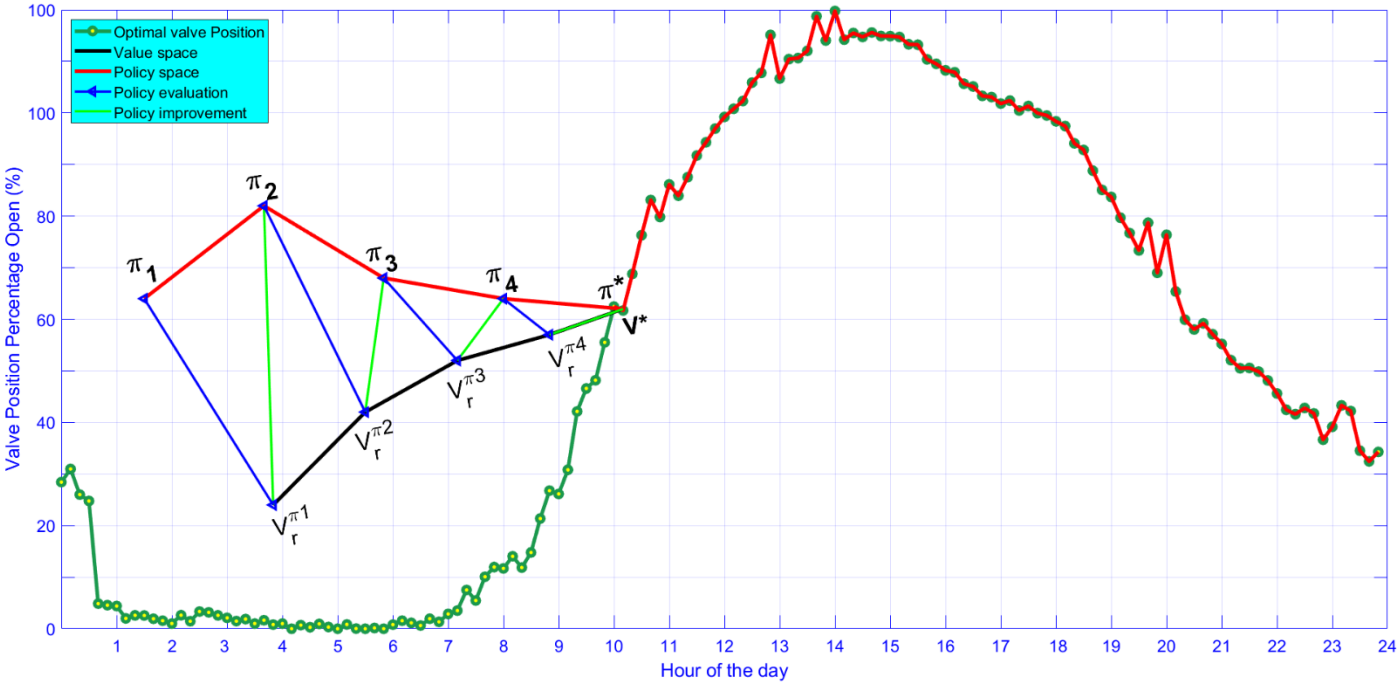

Fig. 3. The trajectory of policy updates for both the policy and *V* value spaces.

3. Description of Design DCCMARL

The multi-agent policy contains a large amount of data in the form of tuples, and managing that big data has become increasingly challenging; it's better to design a new structure with the most important characteristic of applying to storing extensive data and enabling it to deal with large-scale domains. Therefore, the proposed structure DCCMARL of MORL is powerful enough to store or process data efficiently, allowing the multi-agent to handle dynamic large-scale domains of multi-action, which is in contrast to DRL. The goals for a variety of cooperative multi-agent control are to achieve energy-saving potential due to compromising indoor thermal comfort; such individual goals are generated long-time horizons of operation of chillers and AHUs. In this section, a cluster of multi-agent decisions fabricated by using TSF rules to specify the parameters of hybrid layers of each agent's action policy that provides a much satisfying position of drives and actuators is used to ensure systematic coverage of DCCMARL assurance methods (e.g., the flow rate of chilled water for both primary and main air cooling heat exchanger, the actuator of OCSC, position of motor driven both dampers fresh and returned air, and varied the fan speed for manipulating air flow rates). The signals of these positions of drives and actuators are adjusted through the agent learns by interacting with the environment of the whole building until a arrive at the agreed time and in acceptable condition of indoor PMV set-point value $[T_r(t) \quad RH_r(t) \quad T_{rr}(t) \quad v_{ar}(t)]_{DES}^T$. This PMV vector value is acquired by implementing the reward mechanism that takes action according to its policy of model-based. Thus, being related to the indoor thermal comfort conditions, the state error between the desired value and the actual value ($e = [T_r(t) \quad RH_r(t) \quad T_{rr}(t) \quad v_{ar}(t)]^T - [T_r(t) \quad RH_r(t) \quad T_{rr}(t) \quad v_{ar}(t)]_{DES}^T$) affected by specifying for estimating the reward function.
The conceptual design of DCCMARL needs to be fabricated efficiently for multiple hybrid layers supported by capable of handling the problems with multi-agent of significant state and action spaces. The updated formula for the parameters of the hybrid layer (physical parameters and neural networks weight) adopted the coverage ratio of setpoint and reward quality. Each agent's (RL policy-represented) cluster center of action route regresses as a result of the adjusted or revised parameters, as seen in Fig. 4. When considering the DCCMARL's emphasis on observation and the multi-agent integration, it becomes apparent that this structure may be constructed in two phases (offline cluster fabricating and online updating stage). However, the well-structured hybrid layers can prove valuable in modelling applications. Before these stages, three sequential steps of creativity shall be implemented prior to the beginning of the two tuning stages. The sequential steps are the initial mapping step of the RL policy, the clustering step, and the TSF identification step.

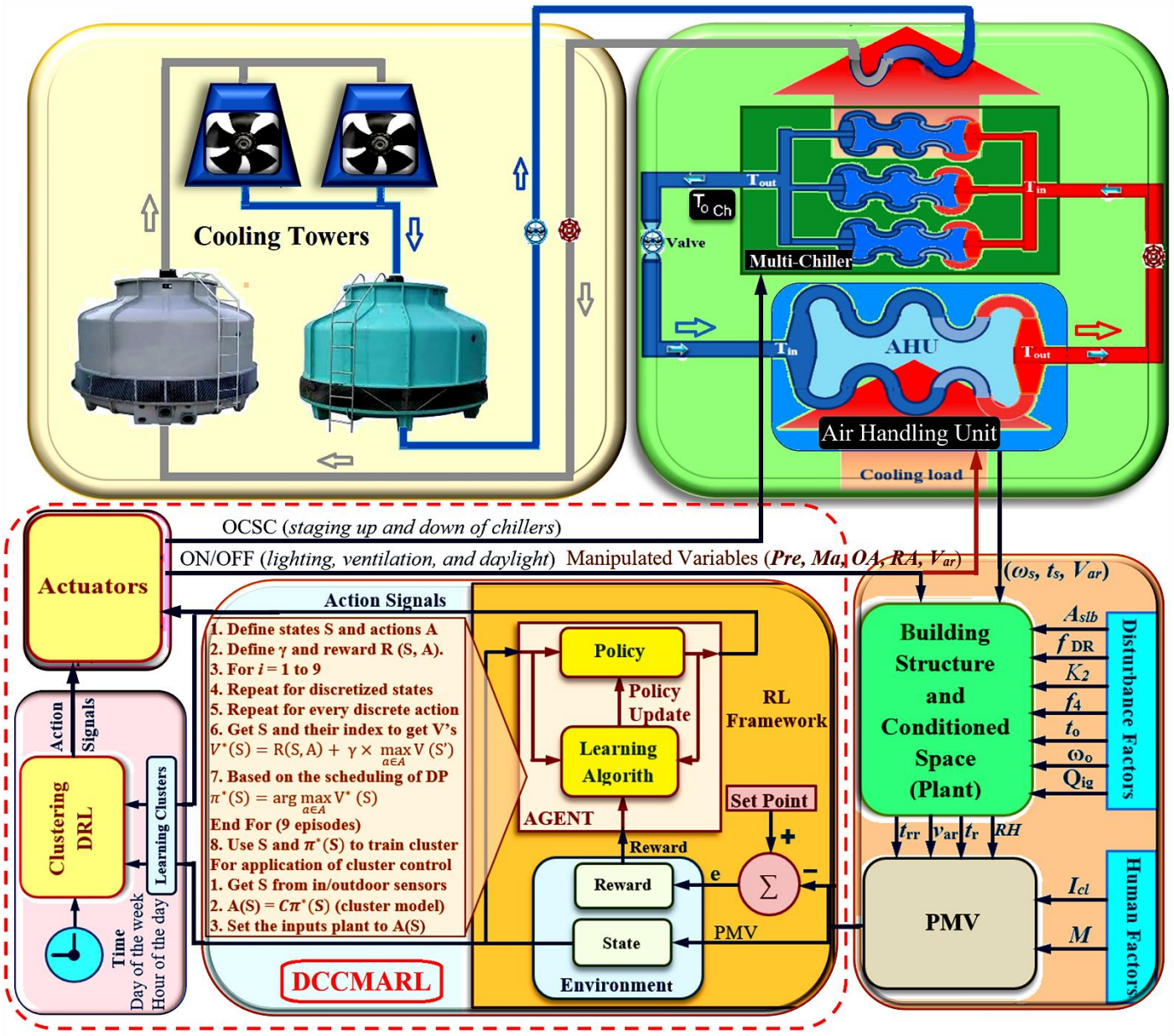

Fig. 4. Schematic representation of new framework composed of DCCMARL based on OCSC and environment sensors.

3.1. Mapping step of RL policy

The optimal RL dynamic model of multi-agent policy is mapping from states to actions with the complete environment data set and, afterwards, this model is converted into clustering techniques using the form of TSF. The policy model dictates the agents' actions to manipulate the whole building's multivariate controllability, such as the input parameters of AHUs, chillers, ventilation, lighting, etc. Each multivariate controllable is translated into an agent-based decision algorithm that is assigned to manage changes in the value of the actions (a), such as manipulating air and chilled water flow rates for both primary and main cooling coils, On/Off chillers, and On/Off light open/close windows (ventilation). Then, the weight of reward (r) received by the agent for each action it takes, based on well-defined (p) transition probability (agent moves from one state to the other), and states (s), the goal states of energy-saving (chilled water flow at AHUs) and thermal comfort can be fulfilled when value function (agents map state-action pairs) V^* optimized iteratively by applying the Eq. (6) of Bellman expectation. The set of environment states is defined as a set of dependent variables for actions; therefore, the trained multi-agent is devoted to the variation of forcing variables states such as chilled water temperatures (leaving from the chiller), indoor temperature and RH, which are assigned to be the states of the environment. Since the states of inside air conditions (temperature and RH) are close to room thermal sensation, which are important to aid estimation and interpretation of the *PMV* using the empirical Fanger equation for achieving compliance with either *ISO*7730 or *ASHRAE* Standard 55-2010 [49]. Therefore, the lower and upper bound of the indoor comfort ranges are potentially incorporated into the reward function (Eq. 8) for RL agents to avoid the excessive cooling and wasting energy.

Optimizing the compromise between energy saving and room condition states is carried out by sequencing the probability of actions to the multivariate controllable (chillers, AHUs systems, etc.). The implementation of cooperative discrete-time MAS makes the conducting of indoor conditions states suitable for six agent-controlled signals (agents of lighting, chilled water valve position (AHUs), windows open position, chillers (OCSC), return-air damper position and fresh-air damper position). The iterative learning of agents under state transfers are benefited from each agent's actions being discretely divided into time intervals (time slot t). This leads to obtaining a dynamic response of the states with time to eliminate indoor air temperature swings. The pseudo-code listed in Table 1 summarizes the significant steps of the conceptual idea behind the DCCMARL and how its agent manipulates a starting state (MDP environment) through a set of actions to achieve the goal of states. The agent actions of the multi-chiller ON/OFF sequencing approaches are representative samples to check the performance responses of a controller according to the chiller sequencing scheme.

Table 1 Algorithm of the DCCMARL regarding OCSC

Algorithm: Pseudo code of the DCCMARL regarding OCSC

1.Initialization:
- 1.1. Setting the policy model parameters; (e.g., environment, states, agents and reward)
- 1.2. Setting controlled states (S); (e.g., indoor temperature and RH, lighting, outlet chilled water temperature, etc.)
- 1.3. Create an initial population of multi-agent actions A_n(n = 1, 2, …, N)
- 1.4. Setting the set of each agent action index $A = a_1, a_2 \dots a_m$

2.Iterate over 24-hour episodes:
- 2.1. For: Start with an initial each state (1 to 9)
- 2.2. Repeat for discretized states
- 2.3. Repeat for every discrete action
  - 2.3.1. Get S for each action
  - 2.3.2. Select the action with the maximum function value
  - 2.3.3. Get Max S and their index to get v*
- 2.4. $V^*(s) = R(S,A) + \gamma \times \max_{a \epsilon A} V(S')$
- 2.5. $\pi^*(a|s) = \underset{a \in A}{\operatorname{argmax}} V^*(S')$
- 2.6. End For (9 episodes)
- 2.7. Use S and π^* to train clustering model

3.For the application of cluster control
- 3.1. Get S from indoor/outdoor and chillers sensors
- 3.2. cluster model → A(S)= $C\pi^*(a|s)$
- 3.3. Set the inputs plant to A(s)

3.2. Clustering step

The next step after using the RL algorithm to generate optimal policy is building a clustering model based on deep learning and unsupervised neural networks to be trained by the optimal policy. Unsupervised data clustering learning may be divided into three types based on network initialization: clustering loss, deep neural network (DNN), and network loss. The proposed model used hybrid neural network clustering based on a deep learning architecture. The dataset learning of RL policy is not systemized as distinct clusters, allowing multi-agent actions to be clustered into groups of similar data attributed together without providing policy guidance or supervision. A clustering dimension is an essential and appropriate concept for comparing different model structures. Each agent's action space in a certain state determines its center location by using the fuzzy c-means clustering algorithm to assign each set of action points $\{Y_i^c\}$in a particular cluster for multi-dimensional clusters. The partitions of the dataset into the c cluster are generated by the TSF, where each center point coordinate is assigned using its mean values. Accordingly, the RL policy task is based on a c discretization of the dataset action for each agent $\{U\}$, $U = \{u_1, u_2, \dots u_K\}$ formed into a family of cohesive clusters $\{C_i\}$, i=1,2 ... c, $[2 \le c \le K]$, the equations formula of set-theoretic description for various cluster parameter sets typical is:

$$Y_{i=1}^{c} C_i = U \tag{10}$$

A framework for establishing and tuning a string of clusters is set up and integrated with two-parameter types to be hybrid layers obtained by converting the TSF rules into basis and membership function (MSF), these parameters (weights of neural network and physical) tabulated in chronological order. The hybrid parameters are listed in the memory cells under dynamic and static characteristics. The dynamic part is related to neural weights, and the other is related to physical parameters. The TSF inference system technique created a systematized hybrid layers concept from the collection of if-then rules and adapted it to fit each signal

of the agent's action. [50]. The signal set of actions that the agent takes is divided into sub-clusters based on the number of *c* which is related to static and dynamic parameters (physical and neural weights) of hybrid layers. The variables of hybrid parameters are initiated primarily by TSF sets from the given training (offline learning) data collected from an RL agent's policy learning; then, the parameter errors are corrected by tuning processes using the nonlinear least-squares regression (NLSR). The online continuously refined its weights is necessary to adjust the controlled states by rectifying the agent policy. Whereas the steps for making fine adjustments to fit the model (hybrid deep clustering layer) signal to agents' action according to current states are implemented online or adaptive fine-tuning to updating model parameters by using NLSR-based iterative algorithm, thus providing a much more robust and accurate actions response. As illustrated in Fig. 5, the model's ability to depict multi-agent behaviors regardless of the size of the dataset or the number of tuple states in it is perhaps its most impressive feature. By making the independent variables four dimensions (4D), the dimensionality of the resultant hybrid model network structure is enhanced by one set of hybrid layers. Accordingly, the independent variables are assigned to be the states/observation of the environment/system; as for the conditions states, it can be assigned x_1, x_2, x_3, and *time* to indoor temperature and RH, outdoor temperature, and time respectively. Furthermore, time is a crucial independent factor because its values are strongly related to heating/cooling load; thus, it is essential to be the leading independent input variable of the proposed model, with the configuration and arrangement independent variables of the hybrid deep clustering approach depicted in Fig. 5.

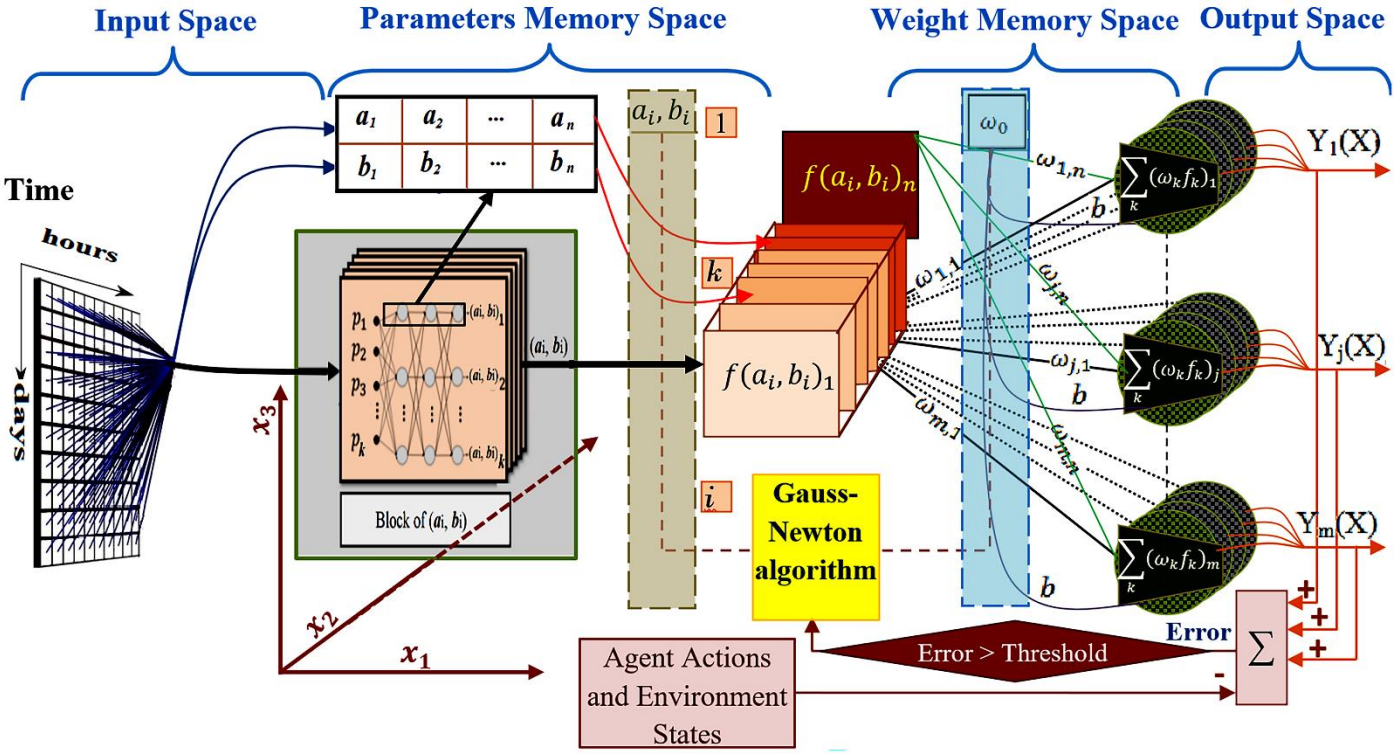

Fig. 5. The network architecture for the hybrid scheme of DCCMARL. The (a_i, b_i) denote the physical parameters and (□) the weights of layers

To better tune and manipulate the action of AHUs and chillers parameters based on the dynamic cooling loads impact to keep the thermal comfort level within the standard, a hierarchical structure may be created by using the clustering approach on a range of each agent's action. The iteration adjustment based on the independent variable of the environment (set of states) is to formalized profile action to match an optimized value of RL policy. The action path for each agent (represented by RL policy), might get its cluster by interpreting the input and output vectors (policy of states, action) to match from the antecedent to the consequent using a rule-based fuzzy inference system. The TSF inference approach produces a fuzzy singleton (polynomial equation) or crisp output vector with their hybrid parameters (parameters of weight and physical value) of layers. The inference starts with random initial weights, which are adjusted by tunable physical parameters; their values have been released through offline learning. These hybrid parameter values are essential to reduce the online learning time due to starting near the target value.

The most effective way to ensure fair regulation of the controlled parameters of HVAC systems, according to achieving energy saving and indoor thermal comfort, is by improving the precise action of multi-agent. The proposed clustering structure allows the DCCMARL to manipulate all building variables, including obvious and easily manipulated variables like building fixtures (turning on and off lights and opening and closing windows) and HVAC systems (valves position for each AHU on each coil, optimal chiller sequencing control (OCSC), the position of motor driven both dampers fresh and returned air, and fan speed for manipulating air flow rates). According to controlled variables, the DCCMARL is a collection of autonomous agents, and the action sequences of each agent are divided into seven groups of piecewise-continuous clusters by leveraging the hyper-ellipsoid

technique (basis and MSFs), as shown at the bottom in Fig. 6. Finally, there are one step ahead of creating the clusters, the dataset of policies is normalized by its maximum value to the range [0, 1] interval as defined in Eq. (11).

$$Norm(xi) = \frac{x_i - x_{max}}{x_{max} - x_{min}} \tag{11}$$

The minimum and maximum (x_{min} and x_{max}) values are denoted by the greatest and least elements in the policy set of agents, and x_i is a current value in the dataset of policy.

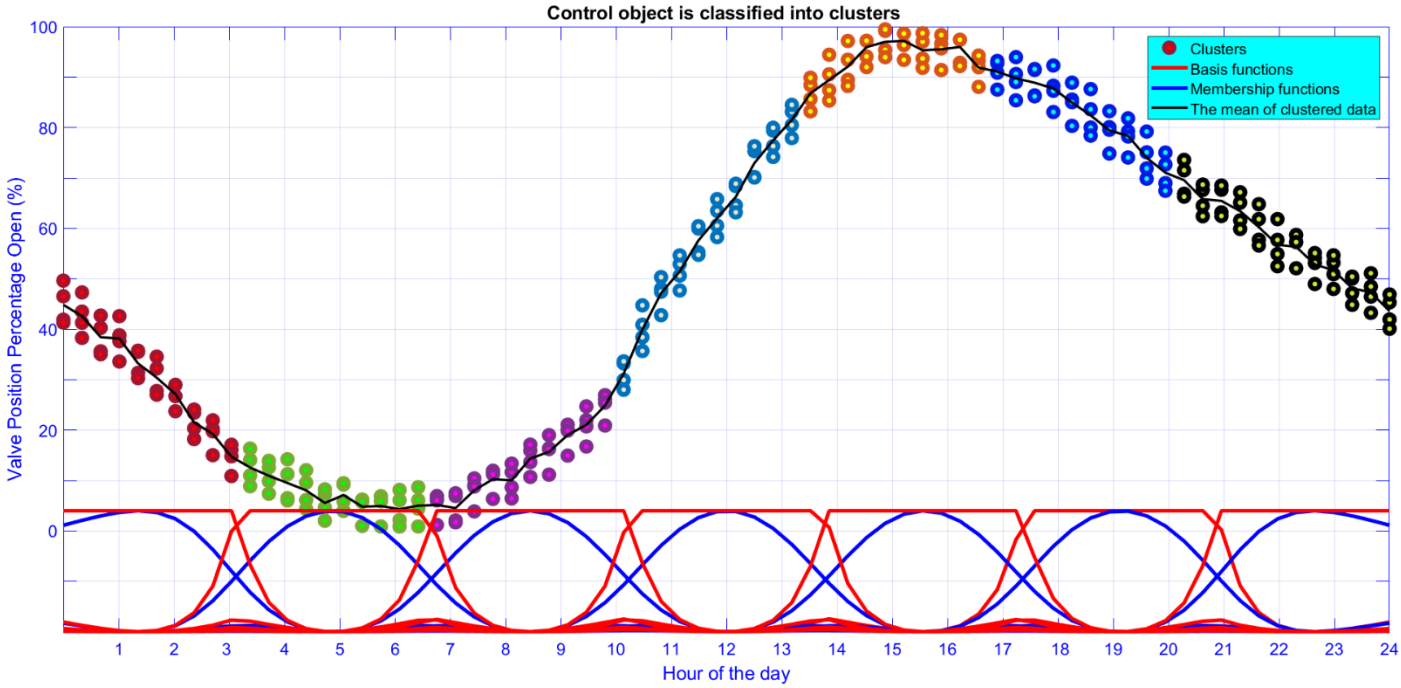

Fig. 6. The policy profile represented by the various positions of the supply chilled water flow rates is broken up on basis of the MSFs and antecedent in TSF.

3.3. TSF identification step

Since the HVAC and building consist of multiple subsystems, which leads to high-dimensional action spaces, it needs to build the multi-hybrid layers structure. The DCCMARL uses the TSF technique for fuzzification to build the multi-hybrid layers, converting the antecedent (IF component of an association rule) and the consequent (THEN component) of TSF into a white-box model (memory hybrid layers parameters). The consequent parameters of the TSF polynomial function generated the hybrid parameters of weight and physical value, which are tuned by fitting the output of the TSF function into action policy mappings. Thus, offline parameters recognition is to design a better-regularized of cluster classification. Where each cluster can be transformed from the consequent part by a polynomial equation of the TSF rules into a crisp output, the formulated equation based on cooling load variation with time is defined by Eq.12.

$$\textit{Cluster}\ (\square_i)\text{:}\quad \text{if}\ \ x_1 \text{ is } M_i^{\,d(x_1)} \text{ and } x_2 \text{ is } M_i^{\,d(x_2)} \cdots \text{ and } x_m \text{ is } M_i^{\,d(x_m)} \text{ then } Y_i(X) = F_i(x_1, x_2 \cdots x_m;\ a_i, b_i) = \omega_i y_i \tag{12}$$

The conceptual clustering techniques for independent learning in a *TSF* function formula, the universe of discourse (UoD) for the input of membership functions (*MFs*) is used for ties for minimum distance between clusters at each of its center and action policy signal. Where UoD represents the environment at each time step ($X = [x_1, x_2 \ldots x_m, time]^T$), the consequent part in Eq. 12 shows the *time* factor revealed the main effects of orientation of the target of the fitting. The sequence dataset of each agent policy action with its dependent variables (states and time) is used to be training data (inputs/outputs) for cluster function to calibrate the hybrid parameters and systemize them in the hybrid layers framework. The dependent variables *m* takes an active part in the framework to accommodate correlated equilibrium for each rule, represented by subscript *i*. Applied inputs/outputs are fuzzified by entering linguistic values *d(xi)*, …, *d(xm)* of fuzzy sets. The *TSF* function refined its parameters of hybrid layers by learning from RL

policy better to fit the consequent part to the agent action. Then, the refined regularization parameters of physical arguments (a_i and b_i) and weights (and bias) (ω_i) are used to verify the matching of the nonlinear consequent function $F_i(x_1, x_2 \cdots x_m; a_i, b_i)$.

The fuzzy inference engine rule basis (three states with time) in each cluster implements a specific type of fuzzy logic based on RL policy. The regression process of continually changing these hybrid arguments (a_i, b_i, and ω_i) is referred to as learning. It is terminated when the outputs of the TSF model error function are minimized, obtaining the starting values of the arguments. Due to TSF correlated coefficients for the dynamic model between the basis and premise FMs, and clusters, the ideas are generated by the consequent part. Furthermore, the periodically re-calibrated parameter's value can find the center of the cluster, which is significantly related to both the estimated number (design decisions) of FMS based on $\mu\tilde{A}(x)$ and truncates the fuzzy set of the consequent part as illustrated in Fig. 6. The TSF inference process takes crisp independent input of MFs, and for the i^{th} rule, adopted three Gaussian functions as MF form because they are smooth and non-zero at all points. The MFs have evaluated the rules using fuzzy reasoning depending on the basis that comes out of the fuzzy logic operation. Fuzzy logic sets effectively quantize their antecedent part or input, which can be defined as Z: zero, P: positive, and N: negative, thus leading to great estimators for the ambiguities and uncertainties associated with the action of the agent. So, the correlation between states with time and agent action can be represented by antecedent and consequent parts for each rule; the following equation illustrates this:

$$\Re_i^{\mathrm{TS}}: \text{ if } \text{chilw}(n-1) \text{ is } (N, Z, P) \text{ and } \text{light}(n-1) \text{ is } (N, Z, P) \text{ and } \mathrm{T}_r\,(n) \text{ is } (N, Z, P) \cdots \text{ and } T_{out}(n) \text{ is } (N, Z, P)$$
$$\text{then } Y_i^{\,n}(X) = \omega_i\left(b_1 + a_{i1}x_1 + \cdots + a_{ij}x_j + \cdots + a_{im}x_m + a_{i0}\right) \tag{13}$$

For a supervised learning data set of policy-based RL (states, time, and agents' action), up to four process steps are needed to perform the TSF inference to generate clusters: the first one is fuzzification of the input variables, the second one is its rule evaluation, the third one shows the aggregation of the rule outputs, and finally the defuzzification obtained. Fig. 7 depicts the general sketch of the TSF clustering and its premise of the rules, while the singleton value forms the corresponding consequences of the rules. The learning dataset is composed of the premise and consequence parts. These are $\mathrm{y}_1(n-1)$ ... $\mathrm{y}_4(n-1)$ represent the previous values of policy and $\left(\mathrm{x}_1(n) \ \ldots \ \mathrm{x}_4(n), \text{ and Time } (n)\right)$ are the states and time values.

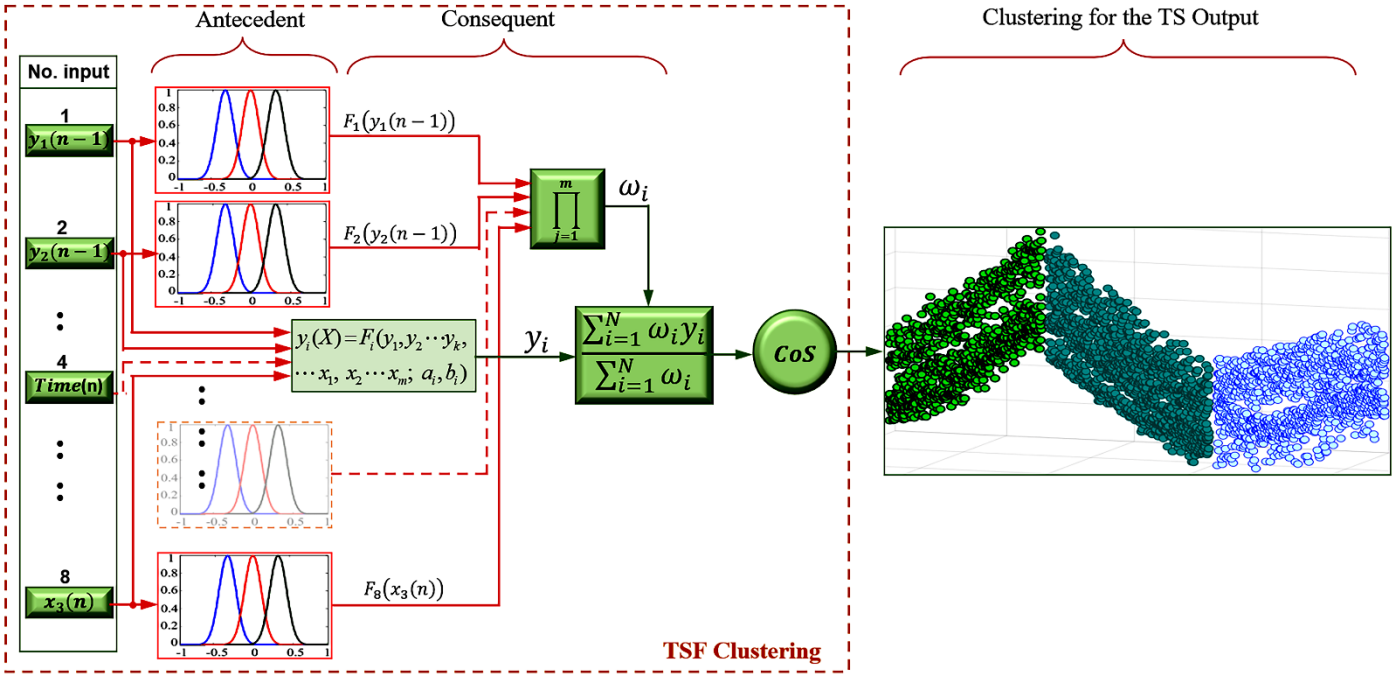

Fig. 7 Three Gaussian functions as MF of inference are used as guidance of TSF to classify clustering data and then calibrate (ω_i, y_i) of each cluster.

In the third process aggregation of the rule outputs, the customary conditions for inference of this process by combining antecedent are shown as a singleton function which can be represented in Boolean values for fuzzy sets [0, 1]. The output function

of TSF is a singleton (crisp) calculated by the weighted average of each of its depending hybrid parameters, which are calibrated by replicates for the data set of RL policy for each rule, as defined in the following equation:

$$Cluster\ (\square_i):\ \text{if}\ \ x_1 \text{ is } M_i^{d(x_1)} \text{ and } x_2 \text{ is } M_i^{d(x_2)} \quad \cdots \quad \text{and } x_m \text{ is } M_i^{d(x_m)} \text{ then } Y(X) = \sum_j Y_i\,(X) \ = \sum_{i=1}^{N} \omega_i y_i \tag{14}$$

In the consequent part of function 14, the *j* value denotes the number of clusters, and the number of clusters is optimized by periodically re-fitting each cluster's centroid into captured optimal agent's action behavior. Finally, in the fourth process, defuzzification of the consequent part, the centroid method or center of gravity (COG) is used to aggregation of all rules to convert the TSF function into crisp/numerical values. The aggregation of all rules generates the numerical values of sampling can be formulated by the following expression:

$$Y(X) = \frac{\sum_{i=1}^{N} \beta_i(u) y_i}{\sum_{i=1}^{N} \beta_i(u)} \tag{15}$$

According to equation (15) for i^{th} individual, the contribution concept is weighted by the degree of activation $\beta_i(u)$ corresponding to the product wight $\prod_{j=1}^{m} \beta_{i_j}(u_j)$, and according to both a linguistic term set and an interval-valued hesitant fuzzy linguistic set (IVHFLS), which provides a computational basis and denotes the possible degrees of the linguistic variable *N*. The nonlinear activation that follows is to accumulate the following terms:

$$\beta_i(u) = \mu M_i^{d(x_1)}(x_1) \wedge \mu M_i^{d(x2)}(x_2) \wedge \ \cdot \ \ \cdot \ \ \cdot \wedge \mu M_i^{d(x_m)}(x_m), \quad 1 \le i \le N \tag{16}$$

The COG defuzzification output is used by the TSF inference approach for singleton type to improve the efficiency of the fuzzified process by drastically reducing the number of defuzzification iterations required to reach the outputs of Eq. 15. According to this technique, Equation15 tends to act like obtaining an average weight value, requiring fewer time iterations than other types of defuzzification and rapidly clustering fitness increases with an increased accurate weights value [51]. Moreover, the consequent element formulation of the nonlinear equation can be expressed in the standard singleton form.

$$Y(X) = \frac{\sum_{i=1}^{N} \beta_i(u) y_i}{\sum_{i=1}^{N} \beta_i(u)} = \sum_{i=1}^{N} \omega_i y_i \qquad \omega_i = \frac{\beta_i(u)}{\sum_{i=1}^{N} \beta_i(u)} \qquad \beta_i(u) = \prod_{j=1}^{m} \beta_{i_j}(u_j) \tag{17}$$

In general, it is possible to calibrate the hybrid parameters of the singleton output Eq. 15 by easing out using an exponential equation with a base of "e", where its parameters ω_i, a_i and b_i are changed over time. The value of parameters is updated based on the change in the agent's action using an equation (18) derived by [52]:

$$Y(X) = \sum_{i=1}^{N} \omega_i a_i \left(1 - \ell^{-b_i\, x}\right) \tag{18}$$

It's obvious, that Eq. 18 is expressed in terms of two types of independent variables or parameters, such as weights (ω_i) and physical (a_i and b_i), these values can be specified accurately when the equation of the consequent part at Eq. 12 passes into the points of policy data set. In principle, the consequent part of Eq. 12 represents the agents' outputs (defuzzification of the action) depending on the states of the environment, where the inputs (states) and the outputs (actions) are correlated by looking at how hybrid parameters (physical and weight values) calibrate their average return value. The correlation coefficients (hybrid parameters) is assigned to cells formed (hybrid layer) at the core of each memory defined by their axes, as illustrated in Fig. 5.

The correlation coefficients in the hybrid layer structure are systemized based on four independents, which are named as (Time; $x_1, x_2,$ and x_3). According to such structure of the model, and its consequent parameters are organized as hybrid layers, which easily lead to these parameters values can be found by crossover from the four dimensions (4D) to the two dimensions (2D), as shown in Fig. 8. The values of parameters $a(x_1, x_2)$ are tabulated to be related to states x_1 and x_2, as illustrated in Fig. 8a whereas Fig. 8b shows the values of parameters $b(x_1, x_2)$ related to different niche axes of states x_1 and x_2.

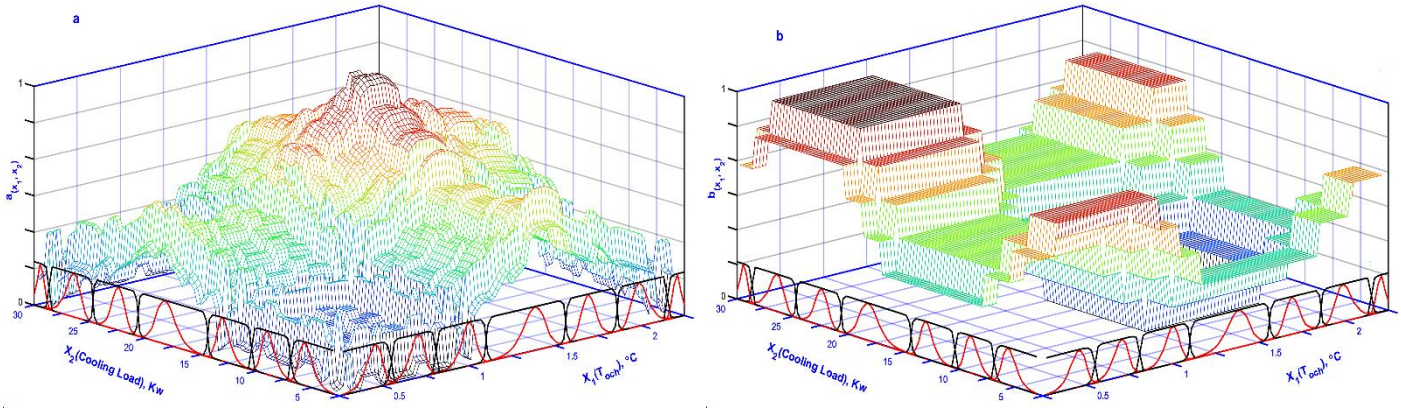

Fig. 8. The different behaviour of physical parameters according to variation of independent variables x_1 and x_2: (a) represents parameter a_i and (b) represents parameter bi.

4. Optimization

The hybrid parameters of the TSF consequent require nonlinear optimization to achieve accurate actions. There are many different ways to optimize parameters and reduce the overall errors in the consequent part, but the most popular way to do it is known as the nonlinear least-squares regression (NLSR). However, two optimization techniques (preoperative offline optimization and postoperative online optimization) are implemented in this section to refine the hybrid parameters.

4.1. Preoperative offline optimization

The consequent part of TSF has segmented each agent's action into seven clusters according to the environment feedback from the HVAC system and occupants. From Eq. (18), the nonlinear functions through the piecewise affine the approximation of outputs that be expressed in an exponential function as follows:

$$\textit{Cluster } (\square_i)\text{: } \textit{if } x_1 \text{ is } M_i^{d(x_1)} \textit{ and } x_2 \text{ is } M_i^{d(x_2)} \quad \cdots \quad \textit{and } x_m \text{ is } M_i^{d(x_m)} \textit{ then } Y(X) = \sum_{i=1}^{N} \omega_i a_i \left(1 - \ell^{-b_i\, x}\right) \tag{19}$$

The TSF uses an extension of the piecewise parabolic function Y(X) of multi-agent action by the clustering technique to break up the policy data into seven segments. Then the optimization tunes the hybrid parameters to accommodate and align the center of seven segments based on the behavior of the expected values of each action in every state. To adjust the parameters, it needs to run the iterative of NLSR until getting a fairly fitting representation of the clusters. The iteration of NLSR provides refined physical parameter values of a_i and b_i for the convergence of optimal action by taking the errors to be used as calibrating factors to update them. In this case, the following equation can express the mistake of the consequent part of the TSF.

$$y_i = \omega_i f(x_i;\ a, b) + E_i \tag{20}$$

The physical variables (a_i and b_i) are the constants of the consequent part (exponential) in Eq. (18), and the residual at each iteration can be expressed as follows.

$$E_i = Y_i(X) - \sum_{i=1}^{N} \omega_i a\left(1 - \ell^{-b\, x}\right) \tag{21}$$

In statistics, the residual sum of squares (RSS) measures the level of variance in the error term of a regression correlation, which it can be found using the formula below:

$$S_r = \sum_{i=1}^{N} E_i^{\,2} = \sum_{i=1}^{N} \left(y_i - \omega_i a\left(1 - \ell^{-b\,x_i}\right)\right)^2 \tag{22}$$

The value of *a* and *b* can be estimated by minimizing the residual sum of squares (RSS) by differentiating with respect to *a* and *b* and equating the resulting equations to zero.

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^{N} 2\left(y_i - \omega_i a\left(1 - \ell^{-b_i\,x_i}\right)\right)\left(-\omega_i(1 - \ell^{-b\,x_i})\right) = 0 \tag{23}$$

However, the dependent variable *a* can be written explicitly in terms of the independent variable *b* as

$$a = \frac{-\sum_{i=1}^{N}\left(-y_i\omega_i(1 - \ell^{-b\,x_i})\right)}{\sum_{i=1}^{N}\left(\omega_i(1 - \ell^{-b\,x_i})\right)^2} \tag{24}$$

Proceeding in the same manner as in the previous, the nonlinear equation in *b* can be derived to obtain

$$-a\sum_{i=1}^{N} y_i\omega_i x_i \ell^{-b\,x_i} + a^2\sum_{i=1}^{N} \omega_i(\omega_i x_i \ell^{-b\,x_i} - \omega_i a x_i \ell^{-2b\,x_i}) = 0 \tag{25}$$

By substituting Equation (24) into Equation (25) and the readjusting the obtained equation in terms of b.

$$\frac{\sum_{i=1}^{N}\left(y_i\omega_i\left(1 - \ell^{-b\,x_i}\right)\right)}{\sum_{i=1}^{N}\left(\omega_i\left(1 - \ell^{-b\,x_i}\right)\right)^2}\sum_{i=1}^{N} y_i\omega_i x_i \ell^{-b\,x_i} = \left(\frac{\sum_{i=1}^{N}\left(y_i\omega_i(1 - \ell^{-b\,x_i})\right)}{\sum_{i=1}^{N}\left(\omega_i(1 - \ell^{-b\,x_i})\right)^2}\right)^2 \sum_{i=1}^{N} \omega_i(\omega_i x_i \ell^{-b\,x_i} - \omega_i a x_i \ell^{-2b\,x_i}) \tag{26}$$

This Equation (26) can be solved by numerical iterative schemes (such as the secant or bisection method) for estimating the root of exponential equations. The hybrid parameters (ω_i, a and b) varied with the time of day as the inputs of the environment states changed or independent variables x_i, they revealed a significant relationship between their deviation values and output TSF error E_i. Based on the consequent partition of the TSF, the parameters can be adjusted using the Taylor expansion method by differentiating the singleton function at the center point of the cluster. In the context of the numerical solution of the exponential partial differential equation using Taylor approximation as

$$f(x_i)_{q+1} = f(x_i)_q + \frac{\partial f(x_i)_q}{\partial a}\Delta a + \frac{\partial f(x_i)_q}{\partial b}\Delta b + \frac{\partial f(x_i)_q}{\partial \omega}\Delta\omega \tag{27}$$

In offline learning, the hybrid parameters are initially calibrated to help the clusters to seek the goal of fitting toward the agent's action policy. This needs to introduce errors in agents' actions by subtracting Eq. (27) from Eq. (20).

$$y_i - f(x_i)_q = \frac{\partial f(x_i)_q}{\partial a}\Delta a + \frac{\partial f(x_i)_q}{\partial b}\Delta b + \frac{\partial f(x_i)_q}{\partial \omega}\Delta\omega + E_i \tag{28}$$

Then, by using the NLSR approach, the Jacobian matrix is formed from partial derivatives of the hybrid parameters such a matrix scheme is considered to be a transfer function to evaluate the updating values according to the Jacobian matrix $[Z_q]$ and its gradients. For example, the matrix form for the iterative Eq. (28) is given by the following equation [53]:

$$\{D\} = \left[Z_q\right]\{\Delta A\} + \{E\} \tag{29}$$

The vector of the dependent variable on the left-hand side of Eq. (29) $\{D\}^T = [d_1 \quad d_2 \quad \cdots \quad d_n]$ can be a measure of inequality or difference between the current policy and the optimal policy of the agent's action. The D vector is helpful for evaluating the tuning values that be used to update hybrid parameters $\{\Delta A\}^T = [\Delta a_0 \quad \Delta a_1 \quad \cdots \quad \Delta a_m]$, at each iteration for each parameter, the difference between values in time series data are $\Delta\omega = \omega_{q+1} - \omega_q$, $\Delta a = a_{q+1} - a_q$ and $\Delta b = b_{q+1} - b_q$. The target of NLSR algorithm is to reduce residual error as much as possible by iteration process to generate the regression fitting error curve $\{E\}^T = [e_1 \quad e_2 \quad \cdots \quad e_n]$.

The inverse matrix $[Z_q]^{-1}$in Eq. (30) allows NLSR to find the values of the hybrid parameters $\{\Delta A\}$ that yield the curve of agent action closest to the optimal policy; eliminated errors are an indicator that measures the closeness between the agent's action and optimal policy.

$$\Delta A = \left[[Z_q]^T [Z_q]\right]^{-1} \left\{ [Z_q]^T \{D\} \right\} \tag{30}$$

The obtaining value of parameters' vector {ΔA}by offline learning using the NLSR algorithm is crucial to reducing the execution time of online learning.

4.2. postoperative online optimization

Since the offline-learned coefficients served as a starting point for the online iteration of NLSR, it provided continually fine-tuned values of ai and bi for the convergence of optimum action by using the output errors as calibrating factors to update them. The online learning of NLSR is designed to eliminate the error entirely by repetitive error detection in agents' actions. The error is usually generated due to the system uncertainty, which leads to the mapping value for the path parameter not being correct at current values, so they need to online update the mapping of the values of the hybrid layer $\{A\}$.

The shortest step of least-squares target ΔS between n_1 and n_2 of each repetitive cluster center point is represented in the output $\partial Y(t(q);\ \omega_i,\ a_i, b_i)$ to be the guidance for getting fitness to policy. The ΔS is adopted to seek the right values to update the coefficient vector of hybrid layers, as follows.

$$A_i(q+1) = A_i(q) + \Delta S_i(q) \cdot \frac{\partial Y(t(q);\ \omega_i, a_i, b_i)}{\partial A_i(q)} \qquad (i = 1,\ \cdots, n) \tag{31}$$

In each time interval, the parameters continuously refined their values by learning from the environment feedback, which is a function of the time slot value of q. The adopting a time slot as an input to the outputs of $(Y(t(q);\ \omega_i,\ a_i, b_i))$, there are errors associated with the outputs due to the uncertainty of predictions center point of the cluster. Therefore, dynamic iteratively online learning is crucial to recalibrating the hybrid parameters $[\omega_i\ a_i\ b_i]^{\mathrm{T}}$, and full convergence occurs when the error goes to zero over time, and iteration schemes can be presented as in Eq. (32).

$$\nabla e_{q+1} = \left[\frac{\partial}{\partial\omega_i} Y(t_{q+1};\ \omega_i,\ a_i, b_i) \quad \frac{\partial}{\partial a_i} Y(t_{q+1};\ \omega_i,\ a_i, b_i) \quad \frac{\partial}{\partial b_i} Y(t_{q+1};\ \omega_i,\ a_i, b_i)\right]^T \tag{32}$$

The parameters $[\omega_i\ a_i\ b_i]^{\mathrm{T}}$ of the consequent function are also periodically re-calibrated with the latest environment operating data, as follows.

$$\begin{bmatrix} \omega_i \\ a_i \\ b \end{bmatrix}_{q+1} = \begin{bmatrix} \omega_i \\ a_i \\ b \end{bmatrix}_q + \begin{bmatrix} \Delta\omega_i \\ \Delta a_i \\ \Delta b \end{bmatrix}_q \tag{33}$$

In the NLSR technique, the online weight update value $\begin{bmatrix} \Delta\omega_i \\ \Delta a_i \\ \Delta b \end{bmatrix}_q$ is computed by learning from the environment feedback from actions in which refined coefficients become available in sequential order based on the descent gradient of the error ∇e_{q+1}.

$$\begin{bmatrix} \Delta\omega_i \\ \Delta a_i \\ \Delta b \end{bmatrix}_q = \Delta S_q \nabla e_{q+1} \tag{34}$$

Each agent in MAS makes individual policy making, but all contribute globally to the HVAC system evolution when, taking a look at the AHU agent, its action manipulated indoor conditions. Therefore, to rectify the error of indoor PMV needs to refine agent action; once an agent acts, the environment (states) reveals its response error due to the uncalibrated hybrid coefficients, as described by Equation (35).

$$e_{q+1} = e_q + \Delta e = mpv\ (t_{q+1}) - S.\ P._q = \sum_{i=1}^{N} \omega_i f(t_{q+1};\ a_i, b_i)\ - S.\ P._q = Y(t_{q+1};\ \omega_i,\ a_i, b_i)\ - S.\ P._q \tag{35}$$

In the next iteration, the indoor thermal comfort is denoted by *pmv* (t_{q+1}), where t_{q+1} represent next time, the other term in Eq. (35) is $S.\ P._q$ denoted to set-point of indoor condition at iteration q and N specifying the sequence number of cluster to ensure that its centre has activated.

The training samples of the updated value vector $\begin{bmatrix} \Delta\omega_i \\ \Delta a_i \\ \Delta b \end{bmatrix}_q$ is calculated by using Eq. (33), such a design problem can be carried out by applying multivariable Taylor series coefficients. So, this optimization is based on the truncated Taylor series of the integrated hybrid parameters, as defined by equation (36).

$$e_{q+1} = e_q + \Delta e = Y(t_q;\ \omega_i,\ a_i, b_i) + \frac{\partial}{\partial \omega_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta\omega_i\ + \frac{\partial}{\partial a_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta a_i + \frac{\partial}{\partial b_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta b_i - S.\ P._q \tag{36}$$

$$e_{q+1} = e_q + \frac{\partial}{\partial \omega_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta\omega_i + \frac{\partial}{\partial a_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta a_i + \frac{\partial}{\partial b_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta b_i \tag{37}$$

In online learning, the time needed to repeat a set of tasks to refine hybrid parameters, thus due to the environment of offline learning, is changed over time. As long as the highly varied environmental conditions, the step-size ΔS at each iteration requires significantly more computation by NLSR according to the error value. For NLSR, an approximation of linearization can be constructed by using the step-size ΔS as a tangent line (shortest path) of the regression, which leads to obtaining the relative errors by substituting the values of the parameter's variation $[\Delta\omega_i\ \Delta a_i\ \Delta b_i]^T$ of Eq. (36) into Eq. (37). The following deriving steps describe the conduction of the mechanism that is taken to improve the iteration process over time.

$$e_{q+1} = e_q + \nabla\left(\frac{\partial}{\partial \omega_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q + \frac{\partial}{\partial a_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q + \frac{\partial}{\partial b_i} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q\right) \tag{38}$$

$$e_{q+1} = e_q + \frac{\partial^2}{\partial \omega^2} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q + \frac{\partial^2}{\partial a^2} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q + \frac{\partial^2}{\partial b^2} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q \tag{39}$$

$$e_q + \frac{\partial^2}{\partial \omega^2} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q + \frac{\partial^2}{\partial a^2} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q + \frac{\partial^2}{\partial b^2} Y(t_q;\ \omega_i,\ a_i, b_i)\ \Delta S_q = 0 \tag{40}$$

To simplify an equation (40) and reduce the expression to its simplest form in terms of step length ΔS_q, and completing the simplification yields Eq. (41).

$$\Delta S_q = \frac{-e_q}{\frac{\partial^2}{\partial \omega^2} Y(t_q;\ \omega_i,\ a_i, b_i) + \frac{\partial^2}{\partial a^2} Y(t_q;\ \omega_i,\ a_i, b_i) + \frac{\partial^2}{\partial b^2} Y(t_q;\ \omega_i,\ a_i, b_i)} \tag{41}$$

$$= \frac{S.\ P_{.q} - \omega_i(q)\ a_i(q)\ \left(1 - \ell^{-b_i(q)\ t_i}\right)}{\omega_i(q)\ a_i(q)\ \ell^{-b_i(q)\ t_i}} = \frac{S.\ P_{.q} - \omega_i(q)\ a_i(q)}{\omega_i(q)\ a_i(q)\ \ell^{-b_i(q)\ t_i}} + 1$$

Therefore, initiating the ΔS_0 based on Eq. (41) is necessary to calculate the updating weight of hybrid coefficients using Eq. (33).

$$\begin{bmatrix} \omega_i \\ a_i \\ b \end{bmatrix}_{q+1} = \begin{bmatrix} \omega_i \\ a_i \\ b \end{bmatrix}_q + \left(\frac{S.\ P_{.q} - \omega_i(q)\ a_i(q)}{\omega_i(q)\ a_i(q)\ \ell^{-b_i(q)\ t_i}} + 1 \right) \nabla e_{q+1} \tag{42}$$

The gradient descent of the least error in Eq. (42) for each set of coefficients is eliminated from the parameter update by an iterative process [54, 55].

5. Physical characteristics of building and chillers

The building descriptions are key to identifying the style of chillers and their capacities by calculating the cooling load of the building. The proposed modeling procedure uses the previous work of white-box building thermal modeling within real environmental conditions to calculate the dynamic residential indoor conditions [42]. Table 1 summarizes the physical specifications for the building envelope in sufficient detail to allow assessment of compliance with the hygrothermal transfer model requirements. Therefore, such a building needs to suitably specify the chillers system by a qualified cooling load demands for comfort. The sequential or different capacity sizes of multi chillers is implemented as the best energy-efficient system due to allowing more alternative scenarios of choosing the appropriate chiller size to meet the current cooling load. Accordingly, two different capacity chillers (10 *Kw* and 20*Kw*) are integrated into the optimal chiller sequencing control (OCSC) system. Thus, it can be provided with four different capacities (0*Kw*, 10*Kw*, 20*Kw* and 30*Kw*) by running/off chillers to meet the current cooling load. The chilled water circulated between chillers and the cooling coils in the AHU based on variable air volume (VAV) boxes. The VAV system delivers ventilation and recirculated cooled air with significant energy-saving and comfortable indoor conditions by reducing the supply airflow rates into an indoor space.

Table 1. The properties of the materials that are used for the residential building construction

Component	Quantity	Variables	Notes
The overall area	870 m^2		
Area of conditioned space	683 m^2 x 2 story		The area excluding unconditioned space (attached garage and storage)
The gross area of exposed windows and walls	441.7 m^2		The net wall exterior area is 379.75 m^2
The overall house volume	1,640.45 m^3 x 2		The house volume excluding the garage and storage
Ceiling	700 m^2	U= 0.18 W/ (m^2. K) Light α_{roof} = 0.4	Overall area less courtyard area (22.6 × 11) – (7.3 × 7.3)
Doors	13.3 m^2	U= 2.3 W/ (m^2. K)	2 (each 0.9 by 2.1 m)
Windows	48.65 m^2	*Fixed: U = 2.84* W/ (m^2. K)*; SHGC=0.67 Operable: U = 2.87* W/ (m^2. K)*; SHGC=0.57 T_x = 0.64; IAC_{cl} = 0.6*	
Walls, exposed exterior	441.7 m^2 gross, 379.75 m^2 net	U= 0.7 W/ (m^2. K)	Wall height = 2.4 m
Walls, garage	122.5 m^2	U= 0.7 W/ (m^2. K)	
Floor area	333.55 m^2	R_{cvr} *= 0.21* W/ (m^2. K)	
Floor perimeter	235.2 m		Include perimeter adjacent to the courtyard
Total exposed surface	235.2 m^2		Wall gross area (including courtyard wall) plus ceiling area

6. Results and discussions

The body of the results section has been conveniently grouped into two primary classifications: evaluating overall chiller plant and whole-building performance.

6.1. The analysis and discussion of chillers' results

This subsection deals with performance aspects of the action of chiller sequencing for DCCMARL duties under varied conditions for plant chillers.

6.1.1. Performance analysis of DCCMARL model

The initial assessment is to identify the optimal value function learnt by a pure RL and the proposed DCCMARL controller. The precisely define optimal value function is required to speculate an optimal policy. The mathematical framework for modelling value function structure is an extension of the Markov chain concept, which is a sequential decision based on the dynamic environment states at instant *t*, which are effective on the uncertainty of agent action, in order to reject disturbance control for uncertain needs to render the modeling value function optimally. The environment training data are generated value function by purely RL iteration to conduct the shortest path counting method to identify the surface of the optimal value function. The defectives of pure RL for the surface of optimal value are clearly shown in Fig. 9a, thus leading to more oscillations and errors inside the dead band set-point compared to the DCCMARL approach. The main concept of DCCMARL is to find the trade-off relation between chillers' energy consumption and the indoor thermal comfort level. Therefore, the value function sufficiently refined its surface by learning (using the NLSR algorithm) from the stats environment feedback to rectify a fault of surface structure. So, the clustering feedback data set of the value function facilitates refined surface tenure and elicits confidence. The DCCMARL rendering value function is much smoother than the pure RL results, as its smoothness is illustrated in Fig. 9b, thus leading to generating the optimal policy. Because both pure RL and the proposed DCCMARL are initiated based on MDP, there is a high degree of similarity between the two values, particularly the max and min values which are specified by using two stats (leaving chilled water temperature of the chillers and outdoor temperature) of each, as both states are represented in Fig.9a and b. As can be seen from both Figs. 9a and b, the variance in leaving chilled water temperature of the chillers ($T_o$ch) are more sensitive than the outdoor temperature (T_{out}) due to the effect of thermal inertia (thermal flywheel effect) against the outdoor temperature effect as well as the priority given to the main task of achieving indoor thermal comfort.

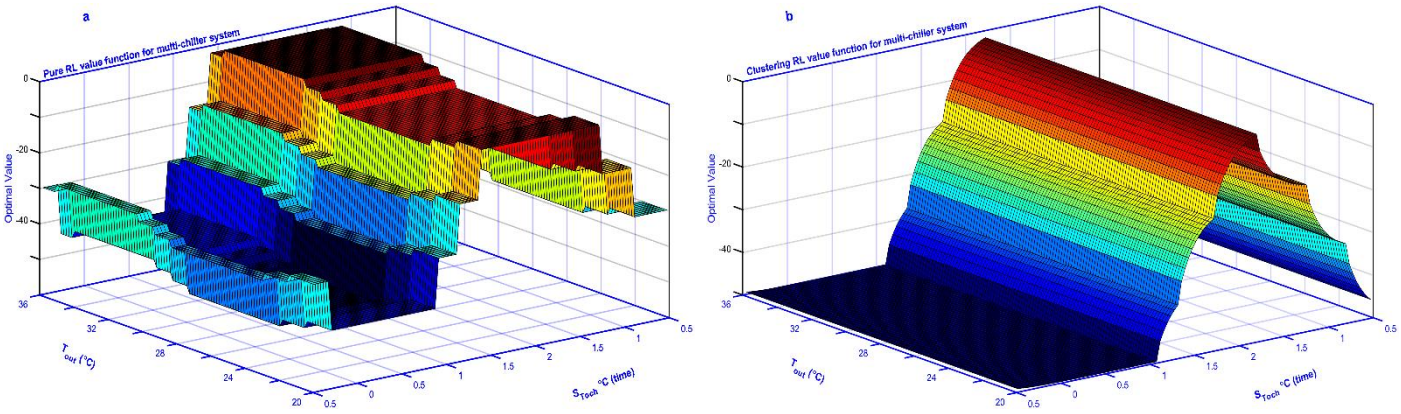

Fig. 9. The two cases of the pure RL and the proposed DCCMARL show difference in the value function as illustrated in (a) and (b), respectively.

The main objective of this analysis is to derive implications regarding the best policy for each method. In other words, the well-systematized way to represent the value function is provided a smooth surface that incorporates precise sequential decisions (optimal policy), leading to a balance of both exploration and exploitation of the RL agent and well-trained. The RL agents learn optimal policies by the iterating manner of Eq. (9) over the states and domain of actions where the value function represents its independent part. The two most effective states represent the best approach of the proposed DCCMARL as independent variables: the $T_o$ch and T_{out}. As shown in Fig. 10, the agent actions of the multi-chiller ON/OFF sequencing techniques for multi-objective are used as samples to demonstrate the dependability of control action by regularly re-calibrating agent parameters using the most recent states' operating data. The best policy of operating sequencing for chillers plant is represented by agent signal action to target optimal decisions based on independent variables, which are the states ($T_o$ch and T_{out}) of the environment, as shown in Fig. 10. It is obvious from Fig. 10 that the policy of the chiller sequencing is operated in four stepladders of the intended move toward the full load according to two sequential types of chillers adopted in this study. The chiller sequencing follows the building's instantaneous

cooling load, as shown in Fig. 10, and includes four actions: at the first action, all chillers are turned off when the outdoor temperature is dropped at late night and early in the morning; at the second and third actions, only one chiller is running; and at the fourth action, all chillers are running. There is a large disparity between the action times of different policies for pure RL and the proposed DCCMARL, as shown in Fig. 10a and b, where the time of all chillers in the run state is longest at pure RL and the opposite is true when all chillers are turned off, as the stepladder action presented there.

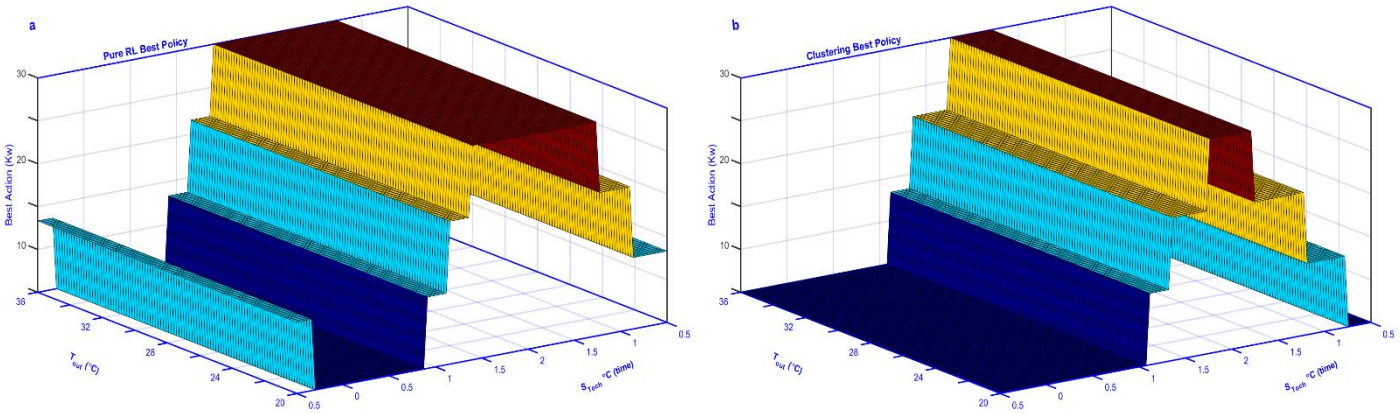

Fig. 10. The difference in the best policy action of two cases: the pure RL and proposed DCCMARL are illustrated in (a) and (b) respectively.

6.1.2. Comparison of chilled water temperature with supervised and pure methods

In this subsection, to be fair to evaluate the performance of the proposed DCCMARL, the comparison test of three algorithms (conventional PID controller, pure RL, and proposed DCCMARL) is conducted under the same conditions. The leaving chilled water temperature of the chiller plant ($T_o$ch) is adopted to be a reference or feedback setpoint for the PID controller and as a reward criterion for the RL agent. Therefore, the $T_o$ch is used to investigate the performance of agents' actions on chiller ON/OFF sequencing. The agent policy interacts with an environment ($T_o$ch) by acting of switching ON/OFF chillers according to the building's instantaneous cooling load to follow the recommended $T_o$ch; its range is defined from 0 °C to 2 °C within 24 h of a day. It is evident from Fig. 11 that the $T_o$ch of the proposed DCCMARL exhibits robust and smooth behavior within a recommended set band, At the same time, the pure RL shows roughness profiles of $T_o$ch, and it can be noticed that the PID is the highest degree of roughness with violating of the recommended band. The proposed DCCMARL policy explores the main components of the building cooling load as a passive and active building thermal storage involved in varying outdoor temperatures. In the passive case, the agent policy utilizes the outdoor temperature drops to the thermal comfort range at night by increasing passive storage using pre-cooling via ventilation. During the passive period (1 to 7 h), the goal of all agents is to achieve as much indoor thermal comfort as possible while turning off all chillers; as shown in Fig. 11, the proposed $T_o$ch is slightly influenced by the chilled water uncirculated through AHU, whereas the conventional PID operates in the opposite direction.

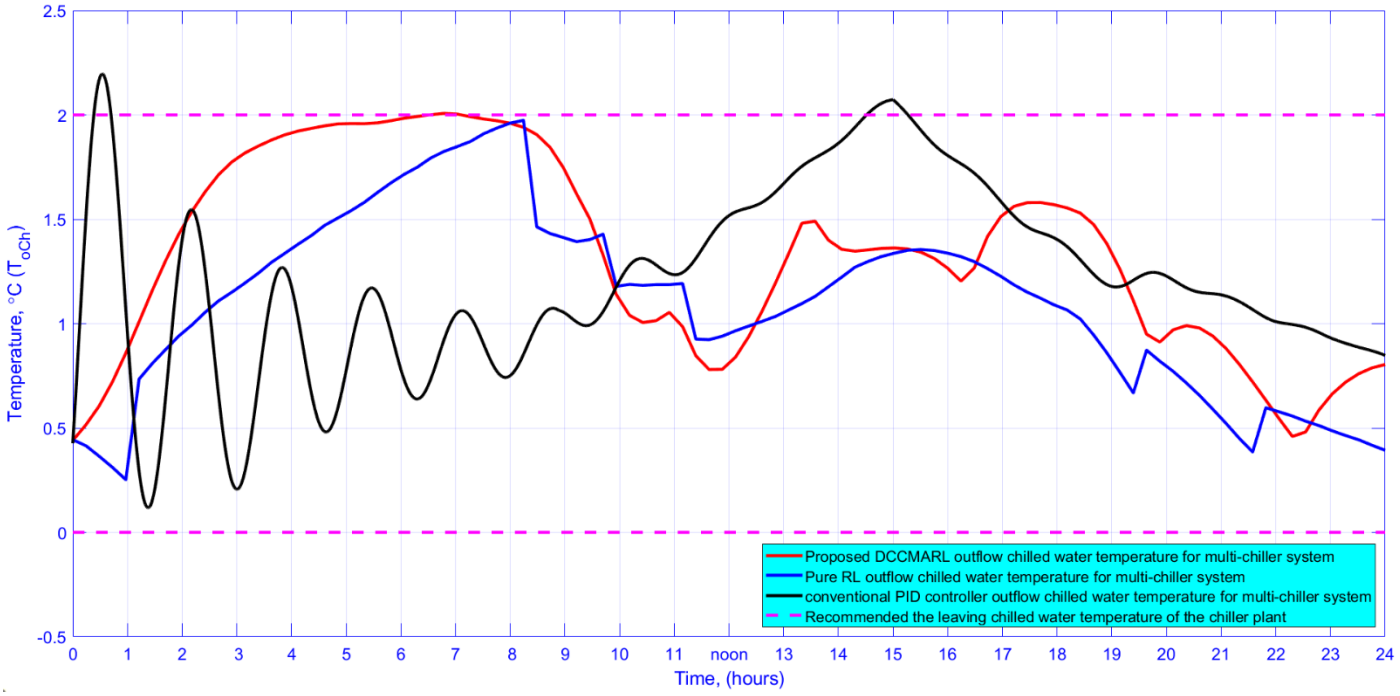

Fig. 11. A comparison between the DCCMARL and other conventional controllers of the $T_o$ch.

6.1.3. A comparative analysis of energy efficiency of chiller plant system

The agent action on chiller sequencing control under dynamic loads has a significant potential for energy saving. From the comparison between the DCCMARL and other conventional controller strategies in the previous subsections 6.1.1 and 6.1.2, the proposed algorithm yields the best performance due to its ability to identify the continuous state-action spaces and massive data training in the shortest time. It minimizes the average run time of chillers and helps to reduce the amount of energy required to provide indoor thermal comfort [56]. Thus, the MAS takes advantage of the thermal mass inertia to synergy to Support two different trends, saving energy at the chiller plant and maintaining the $T_o$ch within the specified stander upper and lower bounds for set-point tracking. To avoid excessive reduction of the $T_o$ch and wasting energy by chillers, the penalty of cooling loads of the building and the violation penalty of dead-band set-point of the $T_o$ch is included in the reward function to optimize energy consumption and keep indoor conditions within comfort limits. The cooling load dynamic characteristics impact the chiller sequencing RL agent action. By closely inspecting Fig. 12, it is evident that the proposed DCCMARL action of chiller sequencing employs passive cooling strategies such as ambient cooling of the building envelope. As long as the cooling coil load is relevant to the agent action of chiller sequencing, which is related to response to the dynamic flow rate of chilled water, the action signal of the agent is used as the main supply component of power for the chiller plant, in addition to the energy of the pumps and fans, to calculate the total cooling coil load. The result of the cooling load in Fig. 12 is a rather remarkable achievement by the strategy of DCCMARL due to it being the best performer out of both PID and pure RL strategies. Meanwhile, the mean cooling coil load values are shown in Fig. 12. The proposed controller is the lowest mean cooling coil load value (power) among the two conventional controller types (the DCCMARL = 10.94 kW, pure RL = 16.02 kW, traditional PID = 21.3 kW, and optimal calculated value = 10.1 kW), so it can be concluded that the performance of chiller sequencing control using DCCMARL policy is recommendable.

The cumulative power consumption profile accounting for daily fluctuations in cooling demand for the chiller plant is the following key criterion to be relevant and necessary in assessing the energy-saving of the proposed DCCMARL. This criterion may be evaluated in terms of cumulative power consumption over time and utilized to accomplish the control performance objective on the chiller plant by modifying the hybrid layer coefficients. The compressors are the primary energy consumers in a chiller plant; thus, understanding the relative advantages and constraints of sequence control is critical for an agent to make the proper action selections. When the chiller plant system is fully operational, the compressors and pumps account for a significant portion of daily energy consumption. Therefore, the agent is molded by feeding the training datasets (e.g., compressors and pumps) into the model, which is completely dependent on the type of clustering learning task to represent the policy framework. The energy-saving policy of the agent demonstrated its profile in Fig. 13 and clearly shows its ability to achieve the goal of optimal cumulative power. The trend of energy-saving by DCCMARL policy saves more than 49.3% in energy a day when compared to conventional PID as illustrated in Fig. 13.

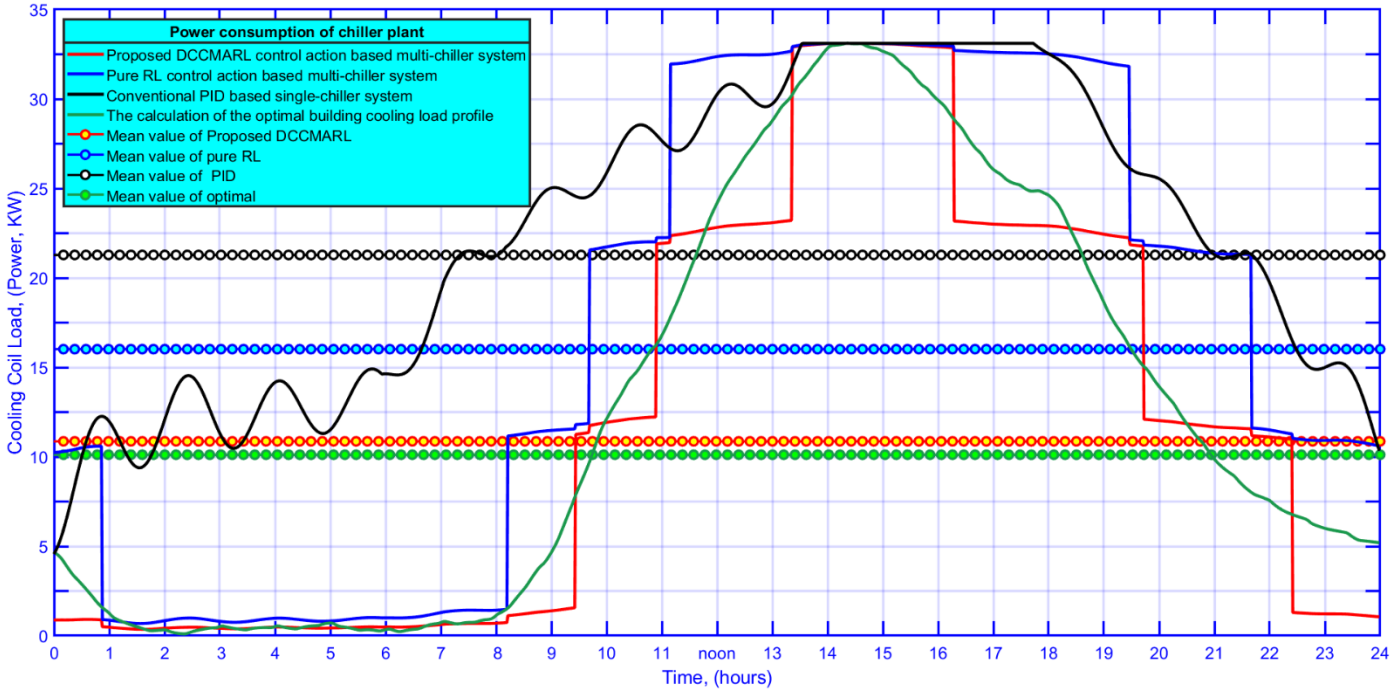

Fig. 12. The cooling coil loads are handled over the energy usage of the HVAC systems by three different algorithms.

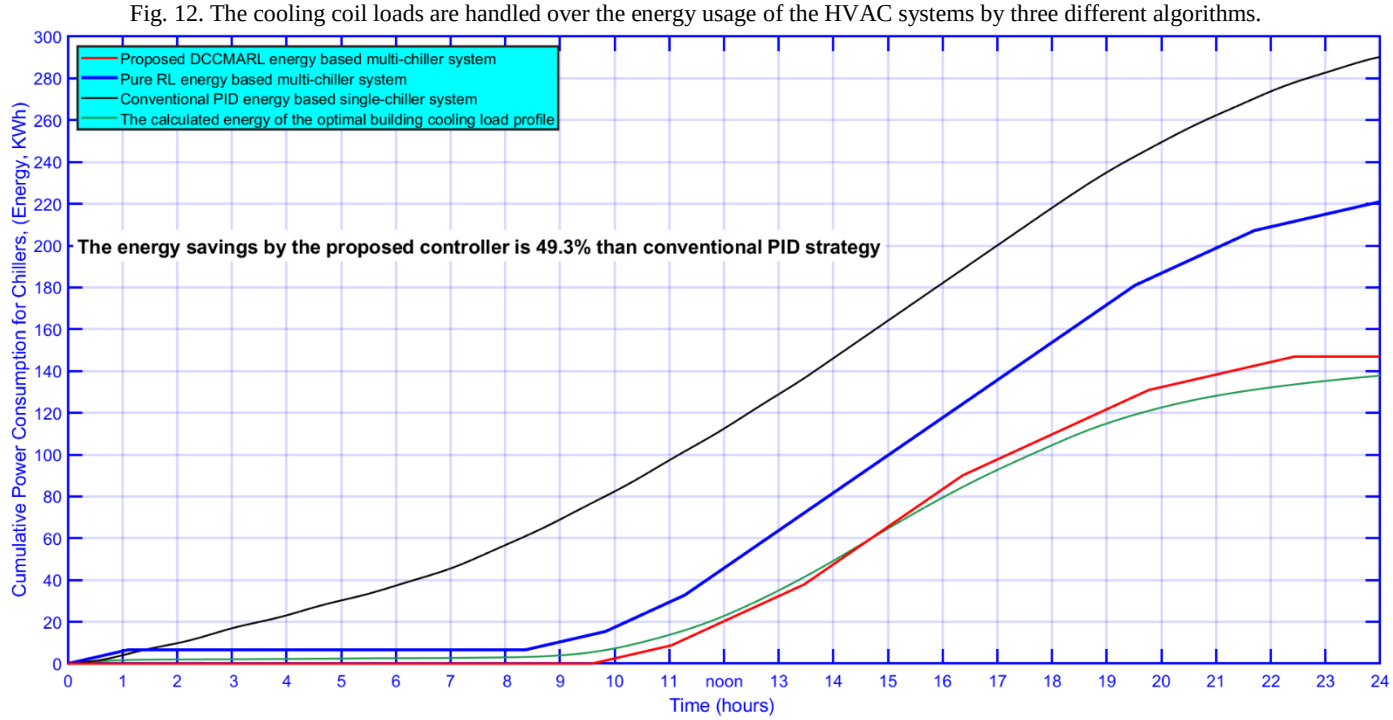

Fig. 13. The cumulative energy consumption of the HVAC systems by three different algorithms.

6.2. The analysis and discussion of whole building

This subsection deals with performance aspects of cooperative multi-agent systems (CMAS) action for DCCMARL duties under varied cooling loads for the whole building.

6.2.1. Performance analysis of DCCMARL model

The learning of cooperative multi-agent RL (MARL) is based on the multi-state framework of the environment to tackle the multi-objective task. The optimal value function is the maximum value obtained when solving an MDP iteratively, which can quickly yield the closest fit to an optimal policy's target value. The RL learning iteratively renders the existence of uncertainty at every moment in the value function structure, which leads to rectifying action policy to handle the defectives of agents' decisions. As shown in Fig. 14a, the predicted value function structure is developed using unadulterated RL learning based on the original setup. A value function developed over a dataset by unsupervised clustering of environmental statistics and refining its values using the NLSR approach to calibrate the structural defect relies heavily on feedback iteration learning to achieve its energy-saving development aims. The value function surface significantly improves smoothness due to DCCMARL rendering the pure RL

results, as depicted in Fig. 14b, which leads to realizing the optimal policy of the agent. Although there is a remarkable similarity in the shape of the two value functions, and both have almost the same global minimum and maximum values, they use the same reward function to model value states. But it is important to note that surface smoothness is a criterion-referenced tool that evaluates control robustness and chattering. According to the smoothness criterion, one finds that the performance of the proposed controller is quite good, as can be seen from the two Figs. 14a and 14b.

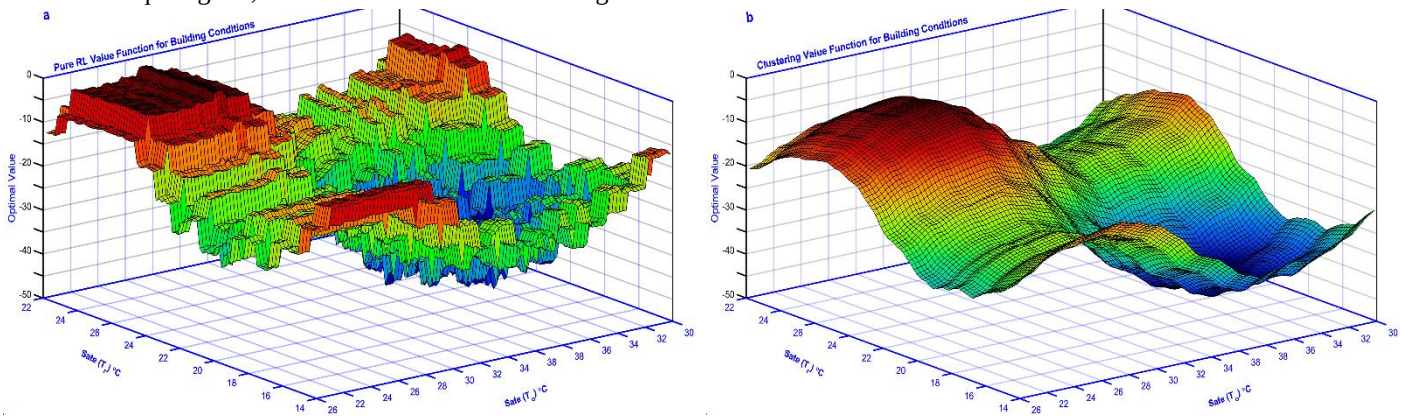

Fig. 14. The proposed DCCMARL function value is a smooth surface for given variables of states; on the contrary, the pure RL surface is so rough, as illustrated in (a) and (b).

Fig. 14 provided a comprehensive overview of the impact on the agent performance decision. In other words, if the value function is defined adequately, it will produce the best results when the inputs are inside the state of the environment domain. And the value function will be updated every time the input is modified, according to Eq. (6). As can be observed from both Figs. 14a and 14b, the response of the value function is more sensitive to changes in the variance of indoor temperature (T_r) than the outdoor temperature (T_o) variation. Thus, the potential thermal mass of the building caused delays in indoor comfort due to the outdoor temperature (T_o) variation.

The RL agents learn optimal policies according to Eq. (9) to explore the environment. The agent continuously refined its control policy by learning from the environment and taking the reward as feedback for their action. The policy function of the pure or standalone (i.e., without the DCCMARL) RL has used three states as independent variables, including indoor temperature (T_r), outdoor temperature (T_o), and relative humidity (RH), but the most influential variables are the T_r and T_o. The flow rate of chilled water related to valve position is adopted as an agent action sample to show the performance of the best policy in accordance with the levels of two independent variables (states), as demonstrated in Fig. 15a. The step beyond that is to involve the identification of concepts of the DCCMARL based on pure RL algorithm to focus on optimal policy to develop an agent acting for the chilled water valve position, which varied its positions as indicated in Fig. 15b. The best policy of the proposed agent achieves an improvement of the dynamic indoor conditions involves especially the parallel actions of the building agents with other multi-agent for the chiller plant. The improvement is achieved not only by enhancing the control structure or using the feedback executed by the clustering technique but also by using the dynamic NLSR algorithm that can be refined its control policy to the output of the agent's decision. Such achievement can be evidenced by different surface smoothness observed in pure RL and proposed policy, as distinctly presented in two Figs. 15a and b.

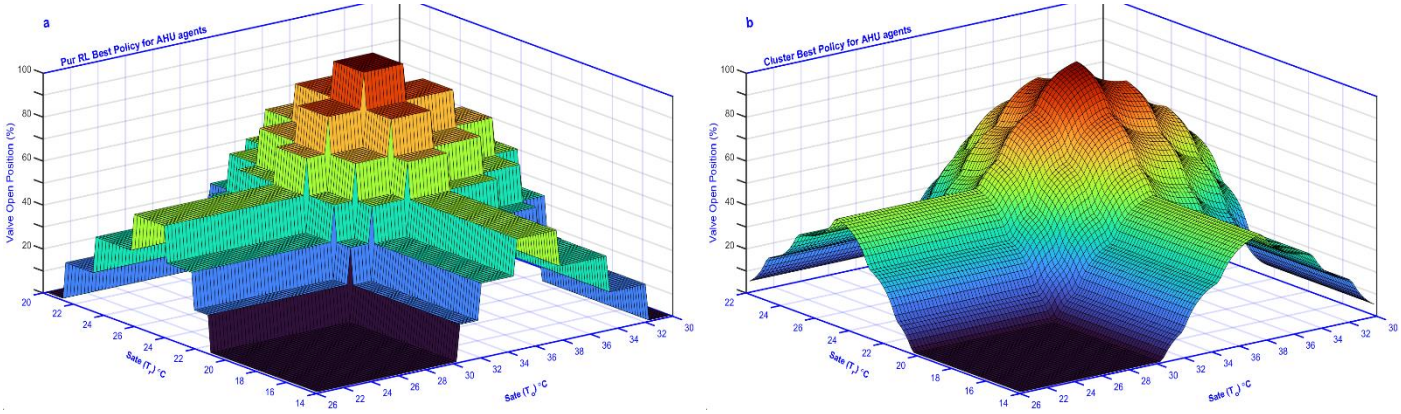

Fig. 15. The difference in the AHU best policy action of two cases: the pure RL and proposed DCCMARL are illustrated in (a) and (b) respectively.

As shown in Fig. 16, the behaviors of the five cooperating agents are resilient throughout the day and through significant temperature fluctuations. It is simple to distinguish strong communications between two agents with different positions of motors powered by both damper's new and returning air, which function oppositely due to their connection. The agents open the action of windows when the value of outdoor temperature is less than or equal to the indoor setpoint temperature. The action of the chilled water valve position follows the building cooling load profile in a day due to the AHU valve position being a function of the flow rate of chilled water. The action of the fifth agent is related to indoor illumination, which is learned in this case of control policy by an optional schedule (flexible). The adopted schedule mainly comes from two aspects of energy saving: the first one is harvesting natural daylight by fully exploiting the potential of windows and skylights, and the second one, in case of inactivity at night, the agent dims the level of lighting within the specified time frame, as depicted in Fig. 16.

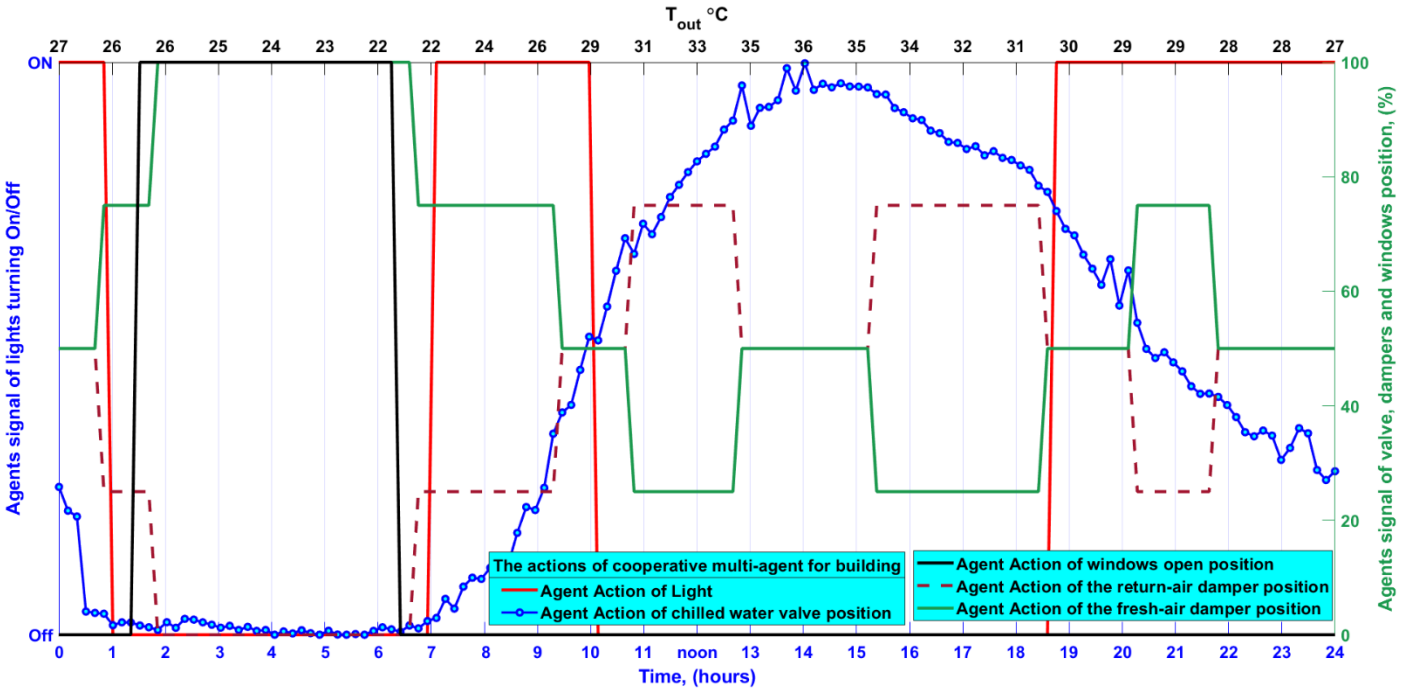

Fig. 16 The responses of five cooperative agents within 24 hours for the proposed DCCMARL.

6.2.2. The evaluation and comparison of variation in indoor conditions

The state transitions (controlled states) are induced by an action set of agents that is governed by rewards such as indoor temperature and RH. Where the agent selects numerous actions every timestep based on the indoor states (temperature and RH) to measure its performance indicators policy. The MARL learning in complex environments with constraints in chilled water flow rate and room thermal comfort is difficult. Since the DCCMARL is capable of multivariate building controls, the agents manipulated AHU and windows to maintain indoor air temperatures by the occupants' comfort bands. The agent of the chilled

water valve is highly effective at controlling indoor temperature and therefore learned through interaction with the occupants' comfort band, which ranges from 20 °C to 24 °C, every 24-h period. The daytime is divided between outdoor temperature values within the occupants' comfort band and out of the band. In the first period (outdoor temperatures within the occupants' comfort band), the agents cooperate in their actions to fully exploit the passive ventilation technologies of the building and switch off all chillers. In the second period (outdoor air temperatures out of the occupants' comfort band), the chillers are switched on by a sequencing agent action according to the environment to provide active cooling effects. In the first period, the target of all cooperative agents is trying to move the indoor state (T_r) as much as possible to be convergent at lower band temperatures (20 °C) and utilize night cooling as passive cooling to the thermal mass of the building. In the second period, the target of all cooperative agents is trying to move the indoor state (T_r) as much as possible to be convergent at higher band temperatures (24 °C) and utilize the thermal mass of the building. That's why the suggested DCCMARL leans toward a reward function suggestion over the two time periods; moreover, as shown in Fig. 17, its interior temperature acts in the preferred fashion (robust stability, steady-state tracking of the desired output and faster response). It is outperforming the two benchmarks across all time. And when the evaluation focuses on the behavior of the two benchmarks in their indoor temperature management, they are not effective enough to be following desired manner. It's clear that Fig. 17 indicates both temperatures of the two benchmark controllers (pure RL and PID) come very close to the same drawback of performance, where they suffer from the chattering phenomenon, and in addition, the PID causes a large amount of chattering phenomenon over time, also it violates the recommended indoor conditions. The side effect of the chattering phenomenon is the wear of the valve stem and the valve guide of the chilled water valve. The high relative humidity (RH) levels inside a building pose an inconvenient feeling for the occupants and are well-known for their destructive effects on building components. To address this challenge, the pre-cooling coil needs to remove the moisture from the fresh air, thus related to controlling the flow rate of chilled water in the pre-cooling coil by the action signal of its valve. The performance of three action signals for the pre-cooling coil valve is demonstrated in Fig. 18, where the agent of the DCCMARL exhibits a largely positive trend for its control policy. Yet, they seem to lie on the same behavior of indoor temperature. When comparing the profile action of three controllers in two Figs. 17 and 18 can assess the similarity of each one, and this is evidence to support the hypothesis presented earlier. Finally, the predicted mean vote (PMV) was calculated using the findings in Figs to measure the performance of the three controllers' behavior. 19 and 20 as the PMV inputs, as shown in Fig. 19. PMV, which is clear from the input variables, represents the criteria of total indoor thermal to building occupant comfort level (temperature and RH). The PID controller also violates the ISO7730 recommended value of the PMV, because its inputs (temperature and RH) are also violated thermal comfort, as illustrated in Figs. 17, 18, and 19.

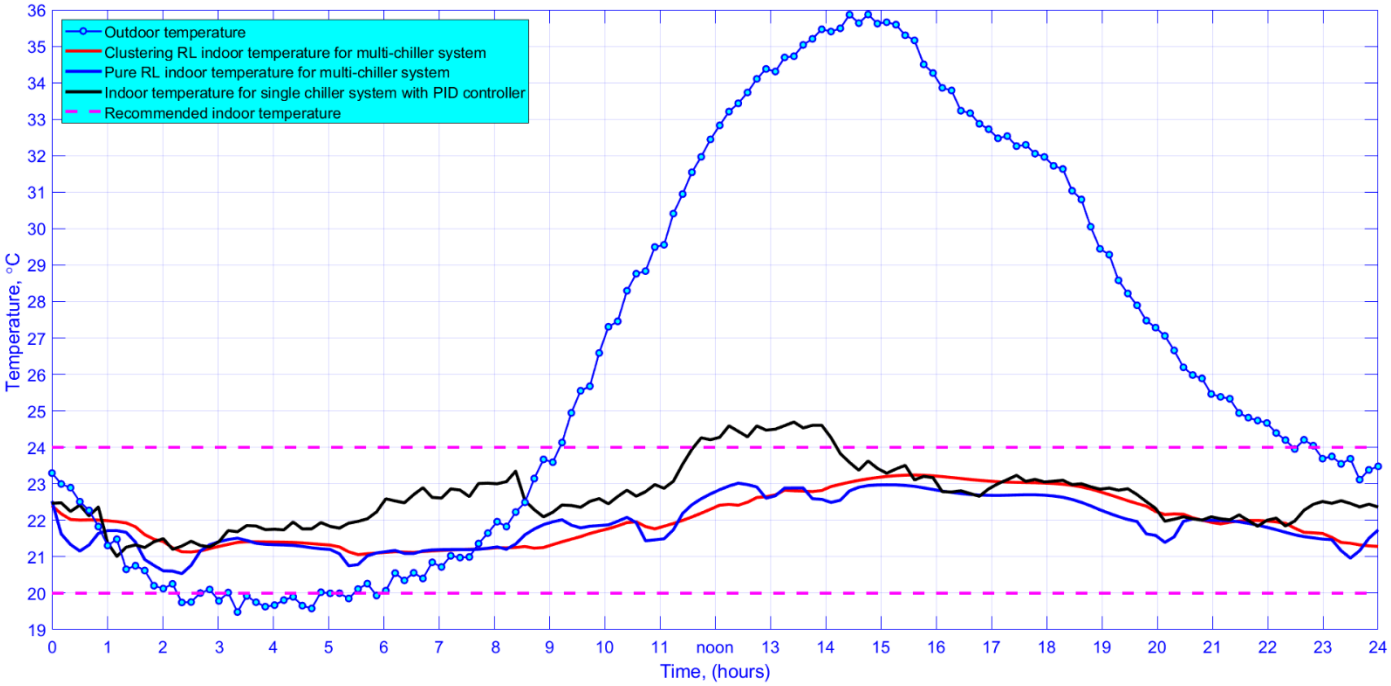

Fig. 17 A comparison between the performance of benchmark controllers and DCCMARL for indoor temperatures behavior.

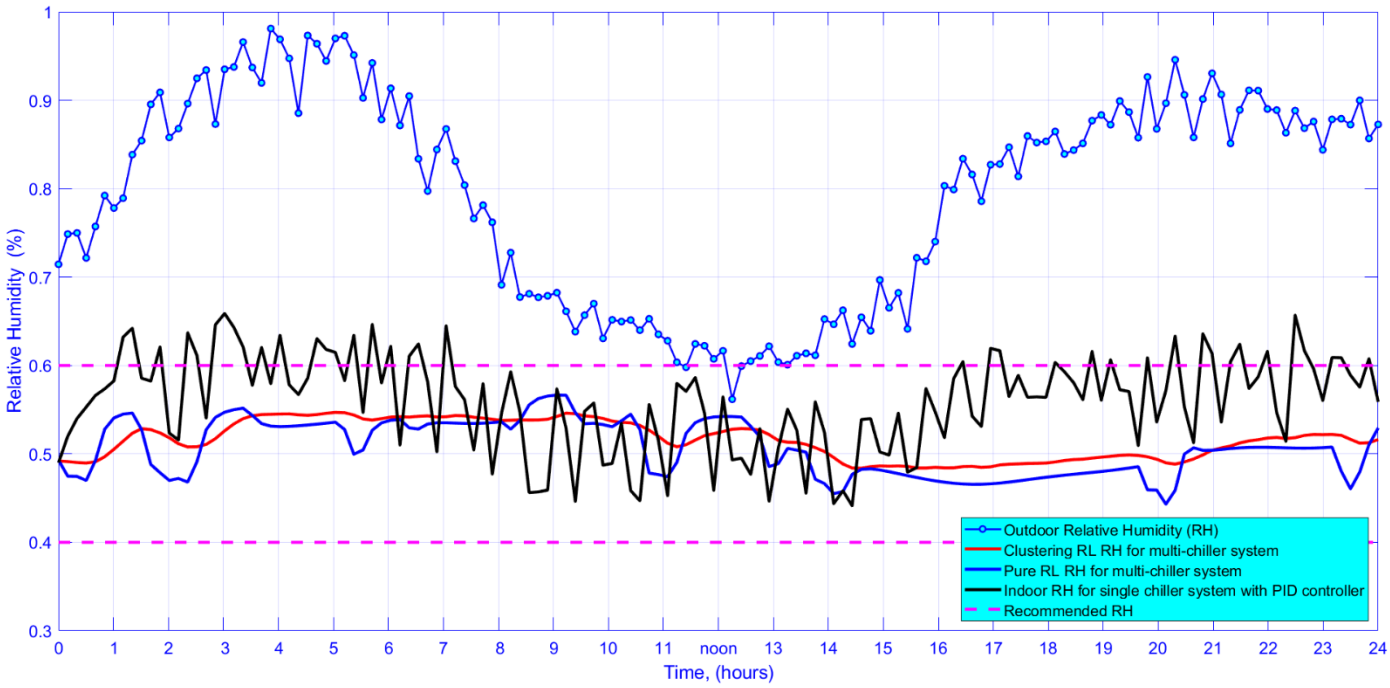

Fig. 18 A comparison between the performance of benchmark controllers and DCCMARL for indoor RH behavior.

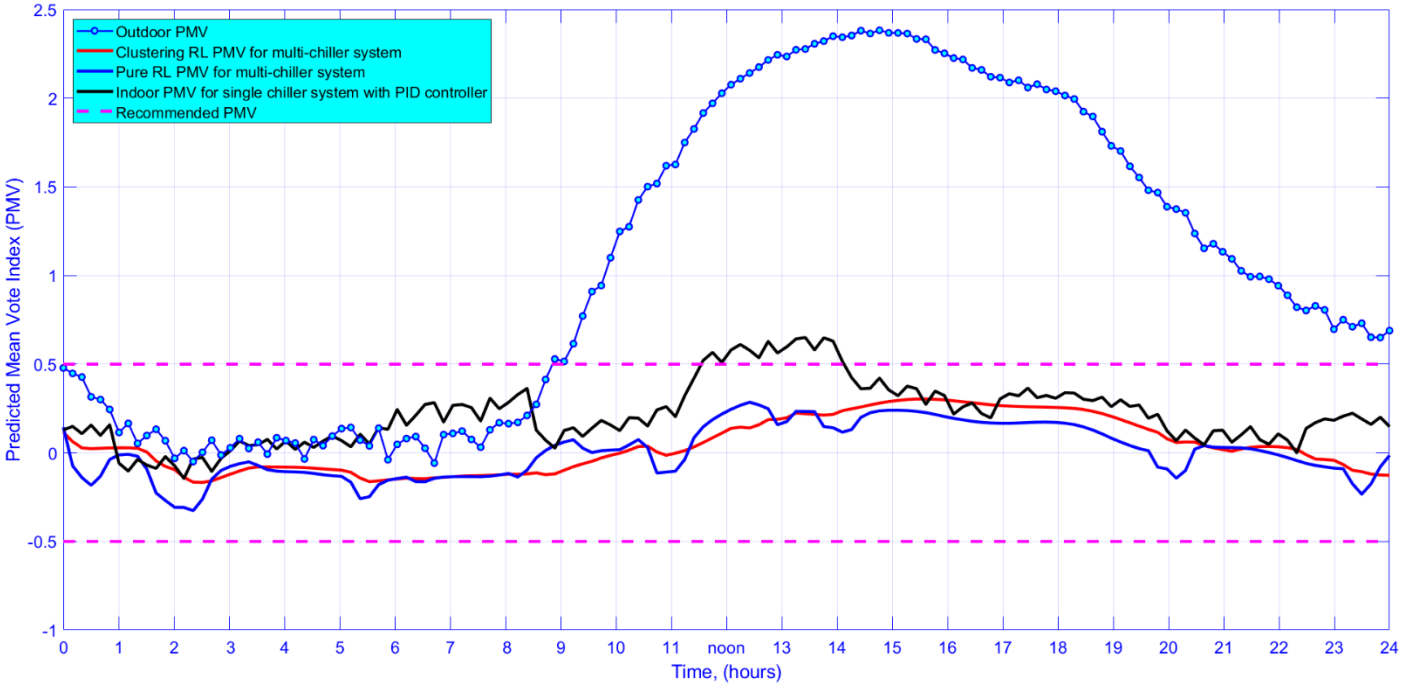

Fig. 19 A comparison between the performance of benchmark controllers and DCCMARL for indoor PMV behavior.

6.2.3. Using psychrometric charts to visualize indoor conditions

The psychrometric processes are beneficial to compute the outdoor/indoor states accurately and evaluate the energy change, also it is a vital tool to assess the validity and utility of controllers according to the ASHRAE comfort zone. The indoor moist air process lies on the curve that is correlated among the air properties, which are represented graphically. Such a graphical chart is used for animation points of the moist air to present the correct visualization of the process, one of them showing dynamic occupants' comfort zone of the building by plotting the air properties saturation of each point. The range of climatic variables that might be drawn on the psychrometric chart to indicate the thermal comfort zone as shown in Fig. 20 is ANSI/ASHRAE Standard 55. The data collected from the produced by a dynamic cooling load, as shown in Figs. 17 and 18, is applied to a psychrometric chart for 24 hours to demonstrate dynamic analysis of the reaction of three distinct types of controllers. It is evident from Fig. 20 that the DCCMARL agents manage the indoor thermal comfort to be turned toward concentration in a narrow space into a comfort zone standard when the extracted data is subjected to the psychrometric chart to show the behavior graph of process curves for indoor and outdoor conditions (cooling and dehumidifying process).

The other benchmarks' behaviors are dispersed throughout a vast region, which is not recommended in this form; also, the typical PID has violated the ASHRAE 55 standard comfort zone. The results obtained from the analysis of two Figs. 19 and 20 (PMV and ANSI/ASHRAE Standard 55) led to identical consequences of accepting or rejecting decisions for indoor conditions; thus, significant evidence indicates that the results are validated.

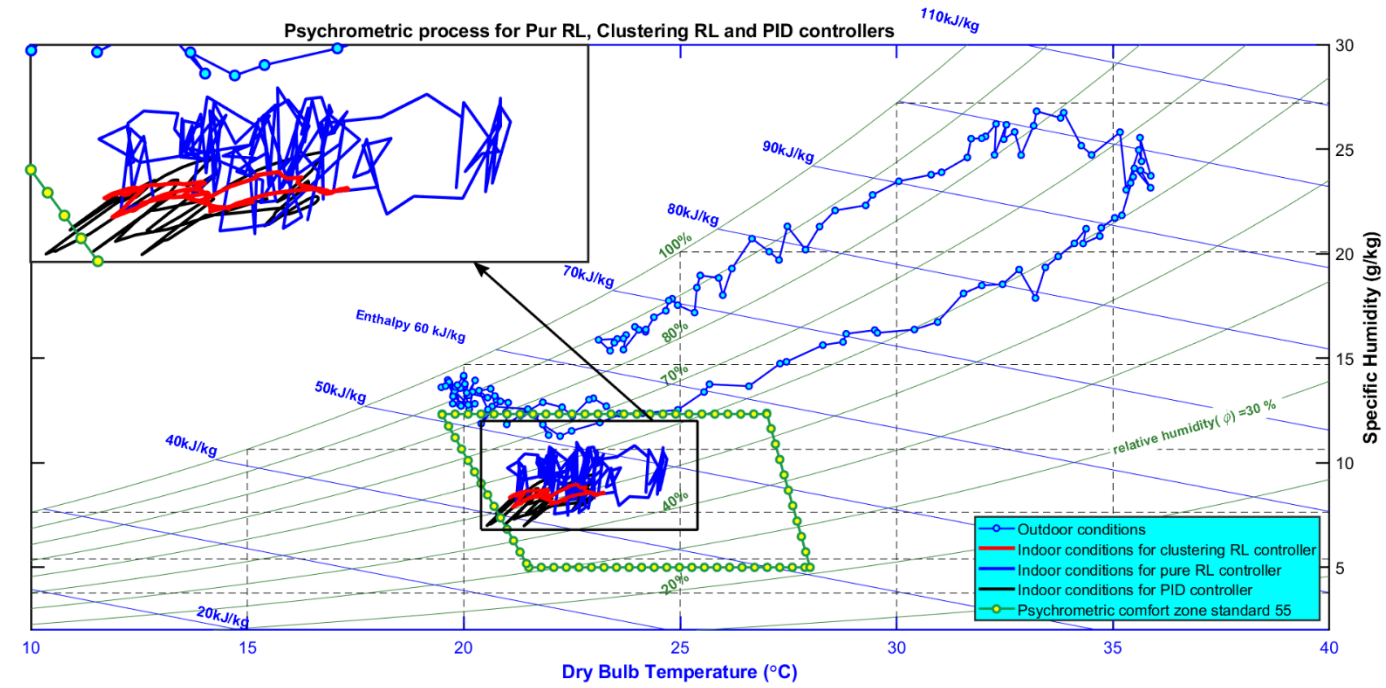

Fig. 20 The indoor psychrometric processes of the DCCMARL and the other conventional controllers.

6.2.4. The overall energy performance and energy saving

A multi-objective smart buildings interaction load management algorithm is presented to minimize electricity consumption and peak power consumption in smart buildings, employing the marginal of an occupants' comfort range to establish an energy saving that considers the load rate. The multi-objective performance of smart buildings has a significant potential to shave peak power demands through cooling load management. The previous subsections 6.1, 6.2.1, 6.2.2 and 6.2.3, show the verification results under different conditions of the proposed DCCMARL policy to achieve more stable, robust and reliable action, this will end up with an energy saving [57, 58]. Since the proposed DCCMARL adopted a multi-objective multi-agent approach, the best policy utilized the trade-offs method between energy saving and keeping thermal value within the occupants' comfort band. The building's overall thermal comfort is investigated in previous sections, whereas this section will focus on showing energy analysis. The HVAC systems and illumination are responsible for major buildings' energy consumption and greenhouse gas emissions; therefore, the energy of HVAC systems and illumination is included in the dynamic Bellman equation under an optimal policy. As a result, as shown in Fig. 16, the reward function employed the chilled water valve position and lighting schedule as independent variables, which are functions within a specific time interval. The chilled water flow rate policy is mainly tied to the exterior temperature due to utilizing the passive cooling approach, which minimizes the peak cooling demands of the structure. According to the reward function, the AHU agent considerably influences energy consumption since it efficiently regulates the flow rate of chilled water based on chilled water valve locations. As a result, the power consumption profile of an entire building, as shown in Fig. 21, is heavily influenced by the action profile of the chilled water valve position, which reflects the best behavior of DCCMARL (Fig. 16) when all agents cooperate. As can be seen in Fig. 21, the MATLAB iterative algorithm uses both the energy consumption of the chilled water plant systems (based on chilled water flow rate, fans, and pumps) and the energy consumption of the building itself (based on illumination, AHU fans, and other devices) to determine the total building energy consumption. This corroborates the findings above when each agent's performance action is investigated independently, demonstrating that the performances of the proposed multi-agent collaboration are more successful. From the results in Fig. 21, the proposed DCCMARL shows the power mean value was the lowest one with a return period of 24 h when compared to the other two benchmark controllers where the proposed = 15.675 kW, pure RL = 18.12 kW, and traditional PID = 28.35 kW.

Finally, another important criterion has to do with the energy of the whole building that is being consumed by the cooling load and other components. This criterion is about using cumulative power for a day to distinguish differences in performance across the different controllers being measured and check if worthy energy efficiency works effectively and reduces emissions. On the other hand, the total power consumption of these three separate controllers is strongly relies on the controller's activity. As a

result, the actions that respond to the large majority of building energy consumption, such as HVAC systems, lighting, and so on, must be specified. Along with HVAC system agents, the agents of lighting and other building systems have exploited the experience of RL for training due to expanding the highest amount of total building energy usage to improve agents' policy by clustering their data for the energy-saving framework. Fig. 22 clearly shows how to improve the policy of agents' choice skills. The trend lines of findings in Fig. 22 reflect the three controllers' management of the entire building's energy use. Such trend lines clearly illustrate that the performance of the multi-agent building employing the DCCMARL strategy exceeded the two benchmark controllers. The suggested agents are making good progress toward their eventual aim of substantial differences in overall energy savings potential from additional policy optimization reaching 44% per day and enhancing indoor conditions by roughly 20.5% of the typical controller (PID). Furthermore, the suggested controller's time profile interior air temperature gives the most apparent proof for a relationship between peak cooling load and indoor temperatures and, thus, energy saving.

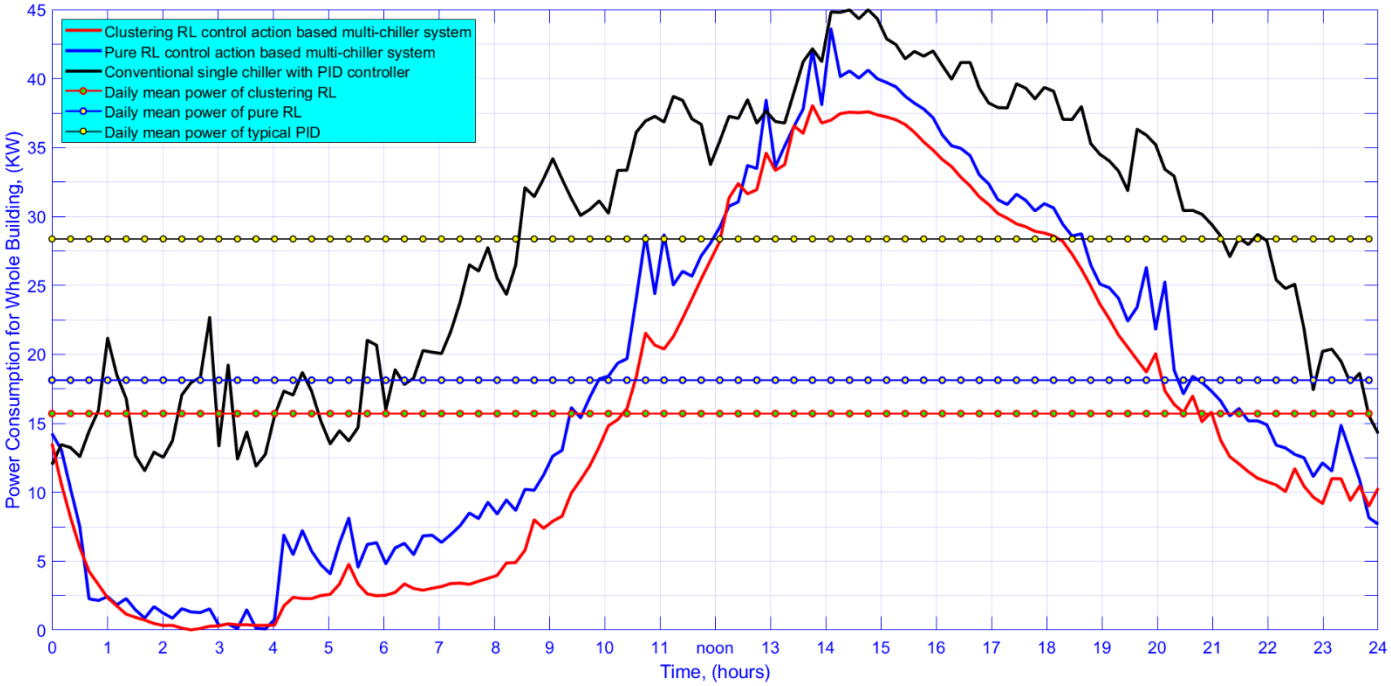

Fig. 21 The difference in the three controller profiles of overall power consumed by the whole building.

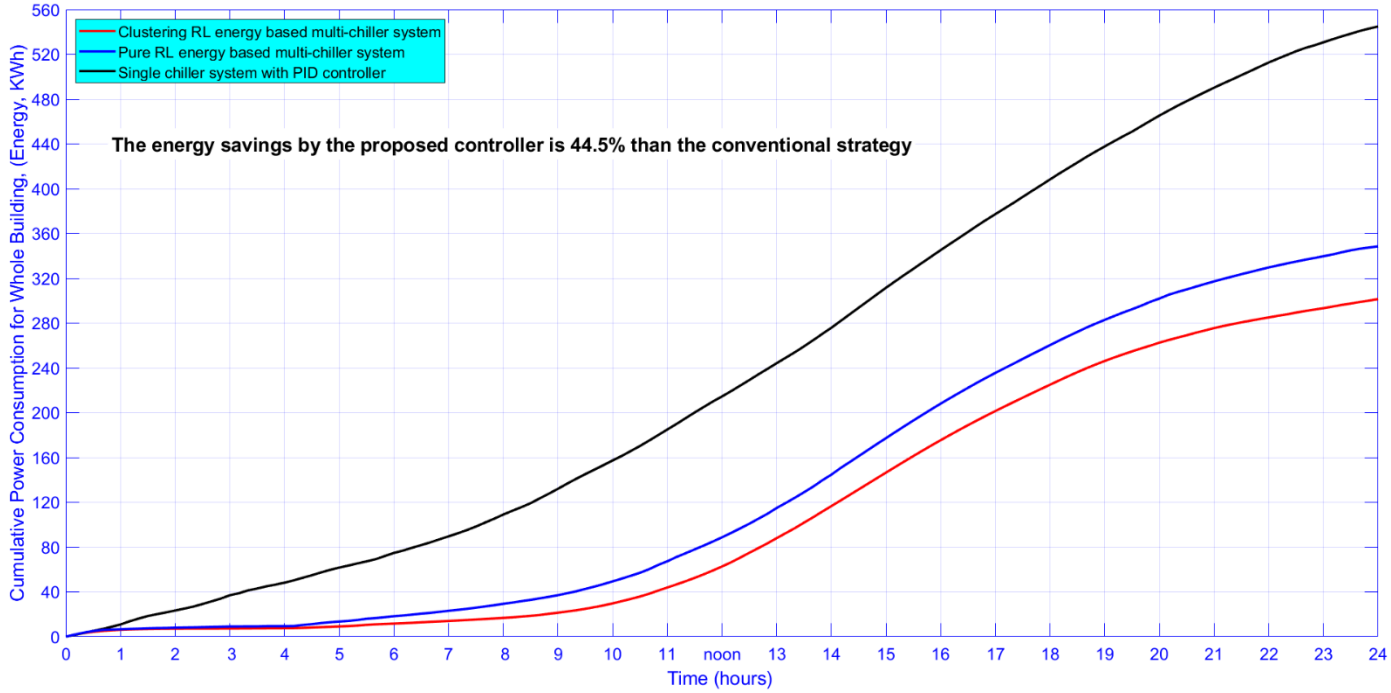

Fig. 22 Illustrated the cumulative power consumption of three type controllers of the whole building.

7. Conclusion

Since the HVAC and building consist of multiple subsystems, which leads to high-dimensional action spaces. The results and validation show that the DCCMARL can be used as MORL and can handle large data sets due to segmented data into multiple clusters. The improvement of the DCCMARL is achieved not only by enhancing the control structure or using the feedback executed by the clustering technique, but also by the dynamic NLSR algorithm that can be refined its control policy to the output of the agent's decision. The study concluded that most of the day-time, the full load is not required; instead of running all the chillers in part load, just run an optimized number of them and the multiple chiller plants can be managed by optimal chiller sequencing control (OCSC). Therefore, the DCCMARL adopted a new strategy by using the varied capacity of multiple-chiller to make energy-saving the main objective. The main objective tasks of the multi-objective multi-agent approach are successively managing multiple chiller plants to achieve the main aim of DCCMARL by realizing. It's the perfect action policy to save more than 49% of energy compared to PID. At the same time, the multi-agent of the whole building provides great tracking ability for required indoor conditions and excellent agent decision stability for more technical energy saving by up to 44.5% and improves the thermal comfort of the occupants by an average of 20.5% as compared to PID. However, we must keep in mind the study's limitations, which are based on the uncertainty of the two models (building and PMV) owing to their parameters regularly changing during their lifetime. As long as the environmental conditions varied dynamically over time, such constraints had little effect on assessing agent performance.

8. DCCMARL and its future applications

The concept of a smart building needs to monitor and control everything indoors and outdoors, which meets the MAS technique. In the potential of future DCCMARL applications could be developed to use its abilities to allow smart buildings to help with complex control tasks. The MAS technique of DCCMARL can service multiple distribution networks of demand response related to occupants, such as control of light, dampers, OCSC, windows, chilled water valve positions, access control, fire safety, and so on. Due to the smart buildings being described as MIMO systems, so the DCCMARL needs to increase the environment states to meet the MIMO system. Nevertheless, the DCCMARL shows its ability to deal with the MARL continuous actions space, as validated in the section on results and discussions. Furthermore, the clustering structure of DCCMARL can be modified according to how many of the agent's policies are, thus implemented by increasing its dimension to handle the expanded dataset of the multiple distribution networks.

9. References

[1] Karakurt, I. and Aydin, G., 2022. Development of regression models to forecast the CO2 emissions from fossil fuels in the BRICS and MINT countries. Energy, p.125650.

[2] Homod, R.Z., Togun, H., Abd, H.J. and Sahari, K.S., 2020. A novel hybrid modelling structure fabricated by using Takagi-Sugeno fuzzy to forecast HVAC systems energy demand in real-time for Basra city. Sustainable Cities and Society, 56, p.102091.

[3] Ahmed, M.S., Mohamed, A., Homod, R.Z. and Shareef, H., 2016. Hybrid LSA-ANN based home energy management scheduling controller for residential demand response strategy. Energies, 9(9), p.716.

[4] Farouk, N., Alotaibi, A.A., Alshahri, A.H. and Almitani, K.H., 2022. Using PCM in buildings to reduce HVAC energy usage taking into account Saudi Arabia climate region. Journal of Building Engineering, 50, p.104073.

[5] Homod, R.Z., Almusaed, A., Almssad, A., Jaafar, M.K., Goodarzi, M. and Sahari, K.S., 2021. Effect of different building envelope materials on thermal comfort and air-conditioning energy savings: A case study in Basra city, Iraq. Journal of Energy Storage, 34, p.101975.

[6] Ahmed, M.S., Mohamed, A., Homod, R.Z. and Shareef, H., 2017. A home energy management algorithm in demand response events for household peak load reduction. PrzeglAd Elektrotechniczny, 93(3), p.2017.

[7] Rao, R.V., 2016. Optimization of multiple chiller systems using TLBO algorithm. In Teaching Learning Based Optimization Algorithm (pp. 115-128). Springer, Cham.

[8] Catrini, P., Cellura, M., Guarino, F., Panno, D. and Piacentino, A., 2018. An integrated approach based on Life Cycle Assessment and Thermoeconomics: Application to a water-cooled chiller for an air conditioning plant. Energy, 160, pp.72-86.

[9] Yan, C., Cheng, Q. and Cai, H., 2019. Life-Cycle optimization of a chiller plant with quantified analysis of uncertainty and reliability in commercial buildings. Applied sciences, 9(8), p.1548.

[10] Scherle, M., Liedtke, J. and Nieken, U., 2022. Optimal sequencing and adsorbent design of multi-bed adsorption chillers. Applied Thermal Engineering, 200, p.117689.

[11] Bhattacharya, A., Vasisht, S., Adetola, V., Huang, S., Sharma, H. and Vrabie, D.L., 2021. Control co-design of commercial building chiller plant using Bayesian optimization. Energy and Buildings, 246, p.111077.

[12] Huang, S., Zuo, W. and Sohn, M.D., 2015, December. A new method for the optimal chiller sequencing control. In Proceedings of the 14th Conference of IBPSA, Hyderabad, India (pp. 316-23).

[13] May, Z., Nor, N.M. and Jusoff, K., 2011, February. Optimal operation of chiller system using fuzzy control. In Proceedings of the 10th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases (pp. 109-115).

[14] Li, X., Lin, S., Shan, K., Han, Z. and Wang, S., 2022. A self-organization method for logic control of distributed building automation system. Journal of Building Engineering, p.104688.
[15] Ting, C.C., Lai, C.W. and Huang, C.B., 2011. Developing the dual system of wind chiller integrated with wind generator. Applied energy, 88(3), pp.741-747.
[16] Sulaiman, M.H. and Mustaffa, Z., 2022. Optimal chiller loading solution for energy conservation using Barnacles Mating Optimizer algorithm. Results in Control and Optimization, 7, p.100109.
[17] Akram, S., Razia, A., Umair, M.Y., Abdulrazzaq, T. and Homod, R.Z., 2022. Double- diffusive convection on peristaltic flow of hyperbolic tangent nanofluid in non- uniform channel with induced magnetic field. Mathematical Methods in the Applied Sciences.
[18] Huang, S., Zuo, W. and Sohn, M.D., 2015, December. A new method for the optimal chiller sequencing control. In Proceedings of the 14th Conference of IBPSA, Hyderabad, India (pp. 316-23).
[19] Hussein, L.A., Ateeq, A.A. and Homod, R.Z., 2022, January. Energy Saving by Reinforcement Learning for Multi-Chillers of HVAC Systems. In IMDC-IST 2021: Proceedings of 2nd International Multi-Disciplinary Conference Theme: Integrated Sciences and Technologies, IMDC-IST 2021, 7-9 September 2021, Sakarya, Turkey (p. 118). European Alliance for Innovation.
[20] Behrooz, F., Mariun, N., Marhaban, M.H., Mohd Radzi, M.A. and Ramli, A.R., 2018. Review of control techniques for HVAC systems—Nonlinearity approaches based on Fuzzy cognitive maps. Energies, 11(3), p.495.
[21] Pacco, H.C., 2022. Simulation of temperature control and irrigation time in the production of tulips using Fuzzy logic. Procedia Computer Science, 200, pp.1-12.
[22] Homod, R.Z., Sahari, K.S.M., Almurib, H.A. and Nagi, F.H., 2012. Gradient auto-tuned Takagi–Sugeno Fuzzy Forward control of a HVAC system using predicted mean vote index. Energy and Buildings, 49, pp.254-267.
[23] Chen, D., Hu, X., Meng, D. and Leto, S., 2020. Optimal consumption modeling of multi–chiller system using a robust optimization algorithm with considering the measurement, control and threshold uncertainties. Journal of Building Engineering, 30, p.101263.
[24] Homod, R.Z., 2014. Assessment regarding energy saving and decoupling for different AHU (air handling unit) and control strategies in the hot-humid climatic region of Iraq. Energy, 74, pp.762-774.
[25] Ahmed, M.S., Mohamed, A., Homod, R.Z., Shareef, H. and Khalid, K., 2017. Awareness on energy management in residential buildings: A case study in Kajang and Putrajaya. Journal of Engineering Science and Technology, 12(5), pp.1280-1294.
[26] Mouneer, T.A., Aly, M.H. and Mina, E.M., 2021. Optimal design configuration and operating sequencing of hybrid chiller plant: A case study of hotel building in Cairo, Egypt. Journal of Building Engineering, 42, p.102796.
[27] Hou, Y.C., Mohamed Sahari, K.S., Weng, L.Y., Foo, H.K., Abd Rahman, N.A., Atikah, N.A. and Homod, R.Z., 2020. Development of collision avoidance system for multiple autonomous mobile robots. International Journal of Advanced Robotic Systems, 17(4), p.1729881420923967.
[28] Dawood, S.M., Hatami, A. and Homod, R.Z., 2022. Trade-off decisions in a novel deep reinforcement learning for energy savings in HVAC systems. Journal of Building Performance Simulation, 15(6), pp.809-831.
[29] Sun, J., Gong, M., Zhao, Y., Han, C., Jing, L. and Yang, P., 2022. A hybrid deep reinforcement learning ensemble optimization model for heat load energy-saving prediction. Journal of Building Engineering, 58, p.105031.
[30] Homod, R.Z., Togun, H., Hussein, A.K., Al-Mousawi, F.N., Yaseen, Z.M., Al-Kouz, W., Abd, H.J., Alawi, O.A., Goodarzi, M. and Hussein, O.A., 2022. Dynamics analysis of a novel hybrid deep clustering for unsupervised learning by reinforcement of multi-agent to energy saving in intelligent buildings. Applied Energy, 313, p.118863.
[31] Ahmadianfar, I., Noori, R.M., Togun, H., Falah, M.W., Homod, R.Z., Fu, M., Halder, B., Deo, R. and Yaseen, Z.M., 2022. Multi-strategy Slime Mould Algorithm for hydropower multi-reservoir systems optimization. Knowledge-Based Systems, p.109048.
[32] Sannad, M., Hussein, A.K., Abidi, A., Homod, R.Z., Biswal, U., Ali, B., Kolsi, L. and Younis, O., 2022. Numerical Study of MHD Natural Convection Inside a Cubical Cavity Loaded with Copper-Water Nanofluid by Using a Non-Homogeneous Dynamic Mathematical Model. Mathematics, 10(12), p.2072.
[33] Dawood, S.M., Hatami, A. and Homod, R.Z., 2022. HVAC system modeling and control methods: a review and case study. Journal of Energy Management and Technology, 6(4), pp.217-231.
[34] Tao, H., Alawi, O.A., Hussein, O.A., Ahmed, W., Abdelrazek, A.H., Homod, R.Z., Eltaweel, M., Falah, M.W., Al-Ansari, N. and Yaseen, Z.M., 2022. Thermohydraulic analysis of covalent and noncovalent functionalized graphene nanoplatelets in circular tube fitted with turbulators. Scientific Reports, 12(1), pp.1-24.
[35] Weigold, M., Ranzau, H., Schaumann, S., Kohne, T., Panten, N. and Abele, E., 2021. Method for the application of deep reinforcement learning for optimised control of industrial energy supply systems by the example of a central cooling system. CIRP Annals, 70(1), pp.17-20.
[36] Ahmed, M.S., Mohamed, A., Homod, R.Z., Shareef, H., Sabry, A.H. and Khalid, K.B., 2015, December. Smart plug prototype for monitoring electrical appliances in Home Energy Management System. In 2015 IEEE Student Conference on Research and Development (SCOReD) (pp. 32-36). IEEE.
[37] Fu, Q., Han, Z., Chen, J., Lu, Y., Wu, H. and Wang, Y., 2022. Applications of reinforcement learning for building energy efficiency control: A review. Journal of Building Engineering, 50, p.104165.
[38] Ahmed, M.S., Mohamed, A., Khatib, T., Shareef, H., Homod, R.Z. and Abd Ali, J., 2017. Real time optimal schedule controller for home energy management system using new binary backtracking search algorithm. Energy and Buildings, 138, pp.215-227.
[39] Jin, W., Fu, Q., Chen, J., Wang, Y., Liu, L., Lu, Y. and Wu, H., 2022. A novel building energy consumption prediction method using deep reinforcement learning with consideration of fluctuation points. Journal of Building Engineering, p.105458.
[40] Khalaf S Gaeid, Raad Z. Homod, Yousif Al Mashhadany, Takialddin Al Smadi, Mohammed Shweesh Ahmed, Aws Ezzulddin Abbas (2022), Describing Function Approach with PID Controller to Reduce Nonlinear Action. IJEER 10(4), 976-983. DOI: 10.37391/IJEER.100437.
[41] Pinto, G., Kathirgamanathan, A., Mangina, E., Finn, D.P. and Capozzoli, A., 2022. Enhancing energy management in grid-interactive buildings: A comparison among cooperative and coordinated architectures. Applied Energy, 310, p.118497.

[42] Homod, R.Z., Sahari, K.S.M., Almurib, H.A. and Nagi, F.H., 2011. Double cooling coil model for non-linear HVAC system using RLF method. Energy and buildings, 43(9), pp.2043-2054.
[43] Homod, R.Z., Sahari, K.S.M., Nagi, F. and Mohamed, H.A., 2010, December. Modeling of heat and moisture transfer in building using RLF method. In 2010 IEEE Student Conference on Research and Development (SCOReD) (pp. 287-292). IEEE.
[44] Ahmed, M.S., Mohamed, A., Homod, R.Z., Shareef, H. and Khalid, K., 2016. Modeling of electric water heater and air conditioner for residential demand response strategy. International Journal of Applied Engineering Research, 11(16), pp.9037-9046.
[45] Homod, R.Z., Sahari, K.S.M., Almurib, H.A. and Nagi, F.H., 2011. Erratum: Double cooling coil model for non-linear HVAC system using RLF method (Energy and Buildings (2011) 43 (2043-2054)). Energy and buildings, 43(9), pp.2043-2054.
[46] Homod, R.Z., 2013. Review on the HVAC system modeling types and the shortcomings of their application. Journal of Energy, 2013.
[47] Homod, R.Z., Sahari, K.S.M., Almurib, H.A. and Nagi, F.H., 2012. RLF and TS fuzzy model identification of indoor thermal comfort based on PMV/PPD. Building and Environment, 49, pp.141-153.
[48] Sahari, K.M., Jalal, M.A., Homod, R.Z. and Eng, Y.K., 2013, June. Dynamic indoor thermal comfort model identification based on neural computing PMV index. In IOP Conference Series: Earth and Environmental Science (Vol. 16, No. 1, p. 012113). IOP Publishing.
[49] Almusaed, A., Almssad, A., Homod, R.Z. and Yitmen, I., 2020. Environmental profile on building material passports for hot climates. Sustainability, 12(9), p.3720.
[50] Ahmed, M.S., Mohamed, A., Shareef, H., Homod, R.Z. and Abd Ali, J., 2016, November. Artificial neural network based controller for home energy management considering demand response events. In 2016 international conference on advances in electrical, electronic and systems engineering (ICAEES) (pp. 506-509). IEEE.
[51] Homod, R.Z., Togun, H., Ateeq, A.A., Al-Mousawi, F.N., Yaseen, Z.M., Al-Kouz, W., Hussein, A.K., Alawi, O.A., Goodarzi, M. and Ahmadi, G., 2022. An innovative clustering technique to generate hybrid modeling of cooling coils for energy analysis: A case study for control performance in HVAC systems. Renewable and Sustainable Energy Reviews, 166, p.112676.
[52] Homod, R.Z., Gaeid, K.S., Dawood, S.M., Hatami, A. and Sahari, K.S., 2020. Evaluation of energy-saving potential for optimal time response of HVAC control system in smart buildings. Applied Energy, 271, p.115255.
[53] Homod, R.Z., Abood, F.A., Shrama, S.M. and Alshara, A.K., 2019. Empirical correlations for mixed convection heat transfer through a fin array based on various orientations. International Journal of Thermal Sciences, 137, pp.627-639.
[54] Homod, R.Z., Sahari, K.S.M., Almurib, H.A. and Nagi, F.H., 2012. Gradient auto-tuned Takagi–Sugeno Fuzzy Forward control of a HVAC system using predicted mean vote index. Energy and Buildings, 49, pp.254-267.
[55] Homod, R.Z., Sahari, K.S.M., Almurib, H.A. and Nagi, F.H., 2012. Erratum :Gradient auto-tuned Takagi-Sugeno Fuzzy Forward control of a HVAC system using predicted mean vote index (Energy and Buildings (2012) 49 (254-267)). Energy and Buildings, 49, pp.254-267.
[56] Homod, R.Z., 2018. Analysis and optimization of HVAC control systems based on energy and performance considerations for smart buildings. Renewable Energy, 126, pp.49-64.
[57] Homod, R.Z., Sahari, K.S.M. and Almurib, H.A., 2014. Energy saving by integrated control of natural ventilation and HVAC systems using model guide for comparison. Renewable Energy, 71, pp.639-650.
[58] Homod, R.Z. and Sahari, K.S.M., 2013. Energy savings by smart utilization of mechanical and natural ventilation for hybrid residential building model in passive climate. Energy and Buildings, 60, pp.310-329.