ENERGY MANAGEMENT AND CONFIGURATION FOR URBAN RAIL TRANSIT TRACTION NETWORKS WITH HYBRID ENERGY STORAGE SYSTEMS BASED ON REINFORCEMENT LEARNING

LI GUANNAN

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Electrical and Electronic Engineering

Energy Management and Configuration for Urban Rail Transit Traction Networks with Hybrid Energy Storage Systems Based on Reinforcement Learning

LI Guannan

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

January 2025

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

______________________(Signed)

_____LI Guannan______(Name of Student)

Abstract

The integration of urban rail transit (URT) traction networks (TNs) with hybrid energy storage systems (HESSs) has become technologically and socioeconomically crucial to enabling highly efficient and convenient mass public transportation within urban areas while promoting the carbon-neutral transformation of URTs. The spatial-temporal uncertainties and complexities arising from passenger demand, urban traffic congestion, widespread distribution, operational disturbances, etc. have imposed significant challenges and limitations on the stable, efficient, sustainable, and intelligent operations of the HESS-integrated URT TNs, especially for those involving distributed HESSs (DHESSs).

This thesis reports using reinforcement learning (RL) as a machine-learning base technique to develop three different levels of energy management and configuration strategies for HESS-integrated URT TNs. These include: (1) a supervised RL-based energy-efficient train trajectory optimization (SRL–EETTO) approach for automatic train operation at the 1st (train) level, (2) a multi-task RL-based sizing and control optimization (MTRL–SCO) approach for HESS-integrated traction substation operation at the 2nd (substation) level, (3) a multi-task multi-agent RL-based multi-time scale energy management (MTMARL–MTSEM) approach for DHESS-integrated TN operation at the 3rd (network) level, and (4) a multi-task multi-agent RL-based data-driven multi-objective configuration optimization (MTMARL–DDMOCO) approach for furthering the DHESS-integrated TN operation at the 3rd (network) level. The research background, problem formulation, approach establishment, and case study verifications are also described for each energy management and configuration strategy.

At the 1st (train) level, the proposed SRL–EETTO approach is aimed to expand the capability of automatic train operation systems in addressing the real-time responsiveness and dynamic online challenges to state-of-the-art TTO approaches and their associated safety, punctuality, and ride comfort issues. A real-time train control model under uncertain disturbances is formulated as a Markov decision process, and a supervised twin-delayed deep deterministic policy gradient algorithm with improved effectiveness is developed to solve the real-time train control model. Satisfactory performances on reduced traction energy use and multiple evaluation indices are verified for the proposed SRL–EETTO approach, and the optimal configuration of the train trajectory set is investigated.

At the 2nd (substation) level, the proposed MTRL–SCO approach is intended to enhance the coordinated operations of the supercapacitor–battery HESSs and their integrated traction substations under dynamic spatial-temporal URT traffic. A dynamic traffic model is devised to characterize the multi-train traction load uncertainty induced by passenger flow fluctuations, real-time traffic regulations, and train parameters. An MTRL algorithm based on a dueling double deep Q network with knowledge transfer is presented to learn a generalized HESS control policy adapting to multiple train service patterns by leveraging a shareable cross-task experience. Simulations have validated the superior computational performance, sizing decisions, and control behaviors of the proposed MTRL–SCO approach.

At the 3rd (network) level, the proposed MTMARL–MTSEM approach strives for the economic and low-carbon operation of TNs integrating with photovoltaic–regenerative braking (PV–RB) DHESSs. A two-stage stochastic scheduling is performed on a long-time scale to minimize daily operation and carbon trading costs at the upper level and correct day-ahead scheduling deviations against multi-source uncertainties at the middle level. A real-time energy management algorithm based on MTMARL is established to optimize PV–RB power flow and promote utilization through decentralized coordination of DHESSs at the lower level. Representative daily TN operation scenarios are selected to demonstrate the improved economic and low-carbon benefits and PV–RB energy utilization of the proposed MTMARL–MTSEM approach.

Furthering the 3rd (network) level, the proposed MTMARL–DDMOCO approach is focused on promoting an optimal synergy between the economic and energy efficiencies of the DHESS-integrated TN operation and the travel time of the passengers. A multi-objective configuration optimization model considering the electrothermal aging of batteries is formulated to optimize DHESS capacities and train operation parameters based on the developed MTMARL–MTSEM approach. The non-dominated sorting genetic algorithm is incorporated with ensemble learning-based load prediction models to solve the multi-objective configuration optimization model in a data-driven manner. The configuration decisions of the proposed MTMARL–DDMOCO approach are analyzed thoroughly.

List of Publications

Journal Papers

[1] **Guannan Li** and Siu Wing Or. "Multi-Agent Deep Reinforcement Learning-Based Multi-Time Scale Energy Management of Urban Rail Traction Networks with Distributed Photovoltaic–Regenerative Braking Hybrid Electric Storage Systems", *Journal of Cleaner Production*, Vol. 466, Article 142842, 2024. [Impact Factor 9.8, Rank 24/358, Percentile Rank 93.4%, JCR Q1* in Environmental Sciences]

[2] **Guannan Li** and Siu Wing Or. "A Multi-Task Reinforcement Learning Approach for Optimal Sizing and Energy Management of Hybrid Electric Storage Systems Under Spatio-Temporal Urban Rail Traffic", *IEEE Transactions on Industry Applications,* vol. 61, no. 2, pp. 1876-1886, 2025. [Impact Factor 4.2, Rank 28/181, Percentile Rank 87.0%, JCR Q1 in Engineering, Multidisciplinary]

[3] **Guannan Li**, Siu Wing Or, and Ka Wing Chan, "Intelligent Energy-Efficient Train Trajectory Optimization Based on Supervised Reinforcement Learning for Urban Rail Transits", *IEEE Access*, vol. 11, pp. 31508–31521, 2023. [Impact Factor 3.4, Rank 122/353, Percentile Rank 65.6%, JCR Q2 in Engineering, Electrical & Electronic]

[4] **Guannan Li** and Siu Wing Or. "A Multi-Objective Train Timetable and Distributed Hybrid Energy Storage System Configuration Optimization Approach for Urban Rail Transits Based on Multi-Task Multi-Agent Reinforcement Learning", in preparation for submission to *IEEE Transactions on Intelligent Transportation Systems*.

Conference Papers

[1] **Guannan Li** and Siu Wing Or. “DRL-Based Adaptive Energy Management for Hybrid Electric Storage Systems Under Dynamic Spatial-Temporal Traffic in Urban Rail Transits”, in *Proceedings 2023 IEEE International Conference on Energy Technologies for Future Grids (ETFG)*, Wollongong, Australia, 2023, pp. 1–6, Paper #135,.

[2] **Guannan Li**, Siu Wing Or, and Ka Wing Chan. “Supervised Reinforcement Learning-Based Dynamic Online Train Trajectory Optimization for Improved Operations of Urban Rail Transits”, in *Proceedings 2023 PolyU Research Student Conference (PRSC)*, Hong Kong, China, 2023, Paper #8366.

Acknowledgements

I would like to express my genuine gratitude to my supervisor, Prof. Siu Wing Or, for his invaluable academic guidance and insightful comments on my research works, as well as his encouragement and constant patience throughout my PhD study. I am also especially grateful to my co-supervisors, Dr. Ka Wing Chan and Prof. Ka Wai Cheng, for their continuous support and guidance. I appreciate all of you for helping me grow up in the academic community and complete this thesis.

I would like to express my appreciation to my group members, Dr. Zhenyu Xing, Dr. Lvping Fu, Dr. Musah Jamal-Deen, Mr. Man Long Hin, Mr. Andy Chan, and Mr. Patrick Yip, as well as members from Dr. Ka Wing Chan's and Prof. Ka Wai Cheng's group, for their camaraderie and selfless help. I also highly appreciate Dr. Yingping Cao for the insightful academic suggestions and discussions that enriched my experience. In addition, I would like to thank my fellow PhD candidates who have visited our group and shared their research expertise.

I am especially grateful to my intimate friend, high school classmate, and roommate, Dr. Jierui Li, who has taken good care of me during the pandemic and has been providing helpful advice and encouragement. Special thanks to Dr. Shaolin Wu, Mr. Zhentao Du, and my good friends at PolyU. I remembered all the fun we had in those years. I would also like to thank my colleagues during my masters study at Wuhan University for their help, especially my friend Mr. Min Xiong.

Finally, I would like to express my gratitude to my family for their endless love, blessing, and care. Also, I would like to sincerely appreciate my soul mate, fiancée, and university friend, Ms. Xiaoqian Qin, for her unwavering love, support, and inspiration.

How time flies! Thank you to everyone I have met during this journey.

LI Guannan

April 2025, at Hung Hom

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AIoT	Artificial Intelligence of Things
ATO	Automatic Train Operation
CDF	Cumulative Distribution Function
D3QN	Dueling Double Deep Q Network
DAS	Day-Ahead Scheduling
DAIS	Day-Ahead and Intraday Scheduling
Dec-POMDP	Decentralized Partially Observable MDP
DHESS	Distributed Hybrid Energy Storage System
DoD	Depth of Discharge
DTM	Dynamic Traffic Model
DTPA	Dynamic Threshold and Power Allocation
DQN	Deep Q Network
EETTO	Energy-Efficient TTO
EL	Ensemble Learning
ESS	Energy Storage System
FT	Fixed Threshold
FTPA	Fixed Threshold and Power Allocation
FPA	Fixed Power Allocation
GA	Genetic Algorithm
HDM	High-Demand Mode
HESS	Hybrid Energy Storage System
iATO	intelligent Automatic Train Operation
IoT	Internet of Things
KT	Knowledge Transfer
KT-D3QN	D3QN with Knowledge Transfer

LCC	Life Cycle Cost
LDM	Low-Demand Mode
LHS	Latin Hypercube Sampling
MD	Manual Driving
MDP	Markov Decision Process
MPO	Maximum PV-Battery Output
MTMARL	Multi-Task Multi-Agent Reinforcement Learning
MTMARL–DDMOCO	MTMARL-Based Data-Driven Multi-Objective Configuration Optimization
MTMARL–MTSEM	MTMARL-Based Multi-Time Scale Energy Management
MTMDP	Multi-Task MDP
MTMH-SAC	Multi-Task Multi-Head Soft-Actor-Critic
MTRL	Multi-Task Reinforcement Learning
MTRL–SCO	MTRL-Based Sizing and Control Optimization
NCS	No Control Strategy
NDS	Non-Dominated Sorting
NSGA-II	Non-Dominated Sorting Genetic Algorithm II
OCV	Open-Circuit Voltage
OD	Origin-Destination
PDF	Probability Density Function
PV	Photovoltaic
PWM	Pulse Width Modulation
RB	Regenerative Braking
RDG	Renewable Distributed Generation
RER	Recurrent Experience Replay
RL	Reinforcement Learning
RNN	Recurrent Neural Network
RPOC	Real-Time PV-Battery Output Control

RTEMA	Real-Time Energy Management Algorithm
RTTR	Real-Time Timetable Rescheduling
SAC	Supervisor-Actor-Critic
SL	Supervised Learning
SoE	State-of-Energy
SHAP	Shapley additive explanation
S-TD3	Supervised TD3
SRL	Supervised Reinforcement Learning
SRL–EETTO	SRL-Based Energy-Efficient TTO
TD3	Twin-Delayed Deep Deterministic Policy Gradient
TEPT	Traction Energy-Passenger-Time
TN	Traction Network
TTO	Train Trajectory Optimization
URT	Urban Rail Transit

Chapter 1: Introduction

1.1 Research Background

With the rapid urbanization worldwide, urban rail transits (URTs), including metros, light rails, etc. [1], play an increasingly essential role in mass public transportation within densely-populated urban regions, leveraging their capability of providing highly efficient and convenient transit services (Table 1.2). By the end of 2023, the total mileage of global URT lines has exceeded 43400 km, and by 2025, the global URT passenger flow will reach a historical high of 954 billion person/km [2]. In mainland China [3], 53 cities have operated URT lines with a total mileage of 9018 km, and the number of carried passengers is 23750 billion. In Hong Kong, the mass transit railway system shares 48% of the franchised transport boarding [4].

With the expansion of URT systems, their energy consumption issues have been increasingly prominent. According to the latest survey [5], the total URT electricity consumption of China exceeds 24977 GW, and the year-on-year increase is 9.59%. Besides, the traction network (TN) energy consumption is 12934 GW, which is the most important component among all energy uses of URTs. Driven by the pressing need to mitigate energy shortages and global climate changes, many countries, including but not limited to the US, China, and the EU, have formulated action plans to enhance clean and diversified energy usage for the net-zero emission of URT and its TN operation [6]. In this regard, the utilization of renewable and train regenerative braking (RB) energy to reduce TN energy consumption has received widespread concerns [7–9].

1.1.1 Renewable Energy Utilization in Urban Rail Transits

Recent investigations [10–17] have revealed the large-scale renewable energy potential of distributed photovoltaics (PVs) and wind turbines using existing URT infrastructure such as trackside slopes, elevated station rooftops, and tunnels. So far, while distributed wind turbine applications in URTs are at an initial stage, several distributed PV demonstration projects have been applied in URTs, as illustrated in the Table 1.1 and Fig. 1.1. However, most of the existing projects and studies were utilized for non-traction energy supply such as HVAC and lightning. As TN energy consumption accounts for 40-60% of the total URT energy consumption [18], it is crucial to realize direct power supply from renewable resources to the TN in the near future [19–21].

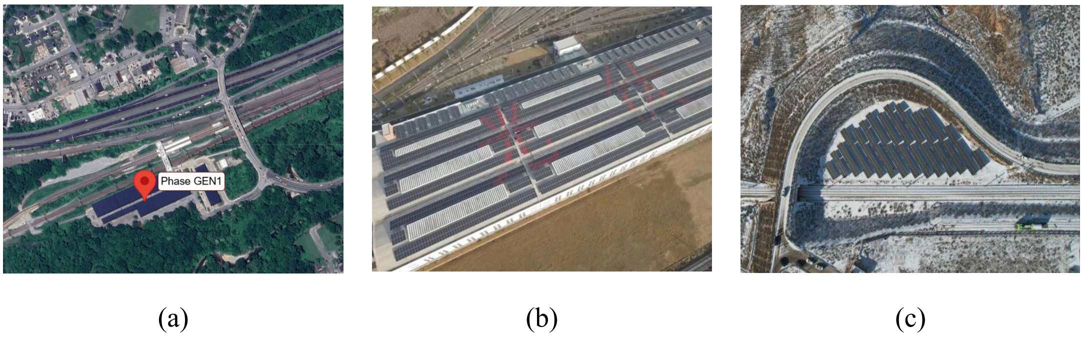


Fig. 1.1 PV projects in (a) Maryland, (b) Shanghai, and (c) Inner Mongolia [21, 22].

Table 1.1 Typical URT and railway PV projects [21, 22].

Year	Location	Parameter	Comment
2013	Beijing	60 kW	First solar-powered station in China
2018	Shanghai	10 MW	At vehicle base rooftops
2019	Singapore	1 MW	At metro depot rooftops
2024	Inner Mongolia	0.38 MW	First project for traction energy supply in China (high-speed rail)
2024	Maryland, USA	1.9 MW	Under construction

1.1.2 Regenerative Braking Energy Utilization in Urban Rail Transits

The RB energy is produced by converting train kinetic energy through traction inverters during train braking. It is reported that over a third of the total train traction energy can be converted to RB energy in some cases [23]. The main RB energy utilization schemes include reversible substation, timetable optimization, and energy storage system (ESS), as shown in Fig. 1.2.

The reversible substation enables the feedback of RB energy to the external AC power grid. Currently, several available reversible substation systems have been developed by Alstom [24], Siemens [25], and Ingeber [26], achieving 7–13% energy saving. However, the impact of RB energy feedback and converter harmonics on the power quality of the external grid needs to be fully considered and optimized.

Timetable optimization aims to extend the overlap time of one accelerating train and another braking train in adjacent power supply sections so that more RB energy can be directly absorbed by the accelerating train [27]. The reported energy saving can up to 14% [7]. This scheme does not require additional equipment investment but must be strictly subject to train operation requirements, such as punctuality constraints [28]. However, during off-peak hours or weekends, the utilization through timetable optimization is lower because the overlap occurs less frequently.

Different from the above schemes which use RB energy immediately, the ESS serves as a buffer hub to temporarily store and release RB energy when demanded. Their reported energy savings can be up to 30% [7]. The ESSs can be divided into onboard and wayside ESSs, where the onboard ESS is installed on the train, and the wayside ESS is installed at the substation or trackside. Considering the installation space and equipment weight, the wayside ESS has broader application perspectives [29].

In addition, the wayside ESS can also perform multiple functions, such as voltage regulation, substation peak load shaving, and power supply for non-railway energy use (e.g., electric vehicle charging [30]). Compared with ESS with a single storage medium, the state-of-the-art hybrid ESS (HESS) can leverage the advantages of multiple energy storage mediums to achieve more efficient and flexible configuration. Considering the improvements in renewable–RB energy utilization and the above operation benefits, establishing and operating novel URT TNs with advanced energy storage (HESS) and enabling technologies (e.g., artificial intelligence (AI) and smart sensors) to promote clean energy utilization has become technologically and socioeconomically crucial for achieving carbon-neutral transformation of URTs.

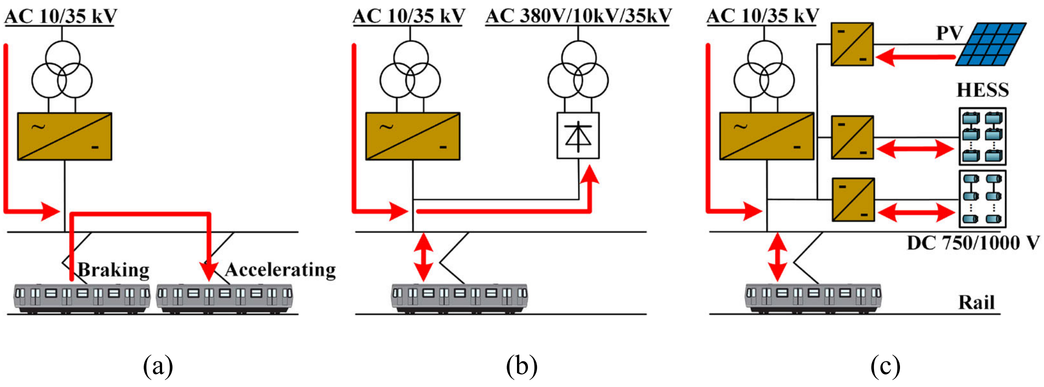


Fig. 1.2 Renewable and RB energy utilization via (a) timetable optimization, (b) reversible substation, and (c) energy storage (wayside).

1.1.3 Operation Challenges of Traction Networks with Hybrid Energy Storage Systems

As illustrated in Table 1.2, compared with the high-speed rails, on the one hand, URTs possess high-density train services and short distances between stations. On the other hand, The real-time URT train operation is more vulnerable to unexpected disturbances [31], including passenger flow uncertainty, pedestrian-vehicle conflict under mixed traffic (for light rails), urban traffic congestion, extreme weather, etc. For instance, field tests on a Beijing subway line with an average train running time of 90–195 s between stations indicated that the range of running time error could reach 15 s [32]. In Dutch, the number of severe light rail accidents per kilometer traveled with vulnerable road users are 12 times higher than those of cars [33]. These operation characteristics have imposed unique challenges on URT TN with HESSs (Fig. 1.3).

Specifically, **at the train level**, frequent train acceleration-deceleration can lead to remarkable train traction energy consumption and power fluctuations (exceeds 10 MW in a few seconds [34]) in the URT TN. Moreover, while the related safety, delay, and passenger satisfaction issues need to be addressed, the disturbances have also resulted in substantial train load uncertainty. **At the substation level**, the stochastic volatility of the train power and renewable generation has limited the efficient utilization of renewable and RB energy via HESSs. Since most traction substations only allow unidirectional energy flows from the external power grid to the TN, excessive renewable and RB energy not utilized by nearby accelerating trains can cause the rise of TN voltage, resulting in stability and thermal management issues [7]. In addition, the possible temporal mismatch between tidal passenger demands and peak renewable generations can undermine the cost-effectiveness brought by HESS installation. **At the network level**, the geographically and temporally dispersed passenger flows, trains, traction substations, HESSs, and renewable distributed generations (RDGs) have resulted in highly dynamic and complex energy flows in the URT TN, exacerbating the challenges at the train and substation levels. Moreover, with the application of multiple sets of HESSs in URTs in the near future, the operation of distributed HESSs (DHESSs) requires a thorough investigation to improve their overall application performances.

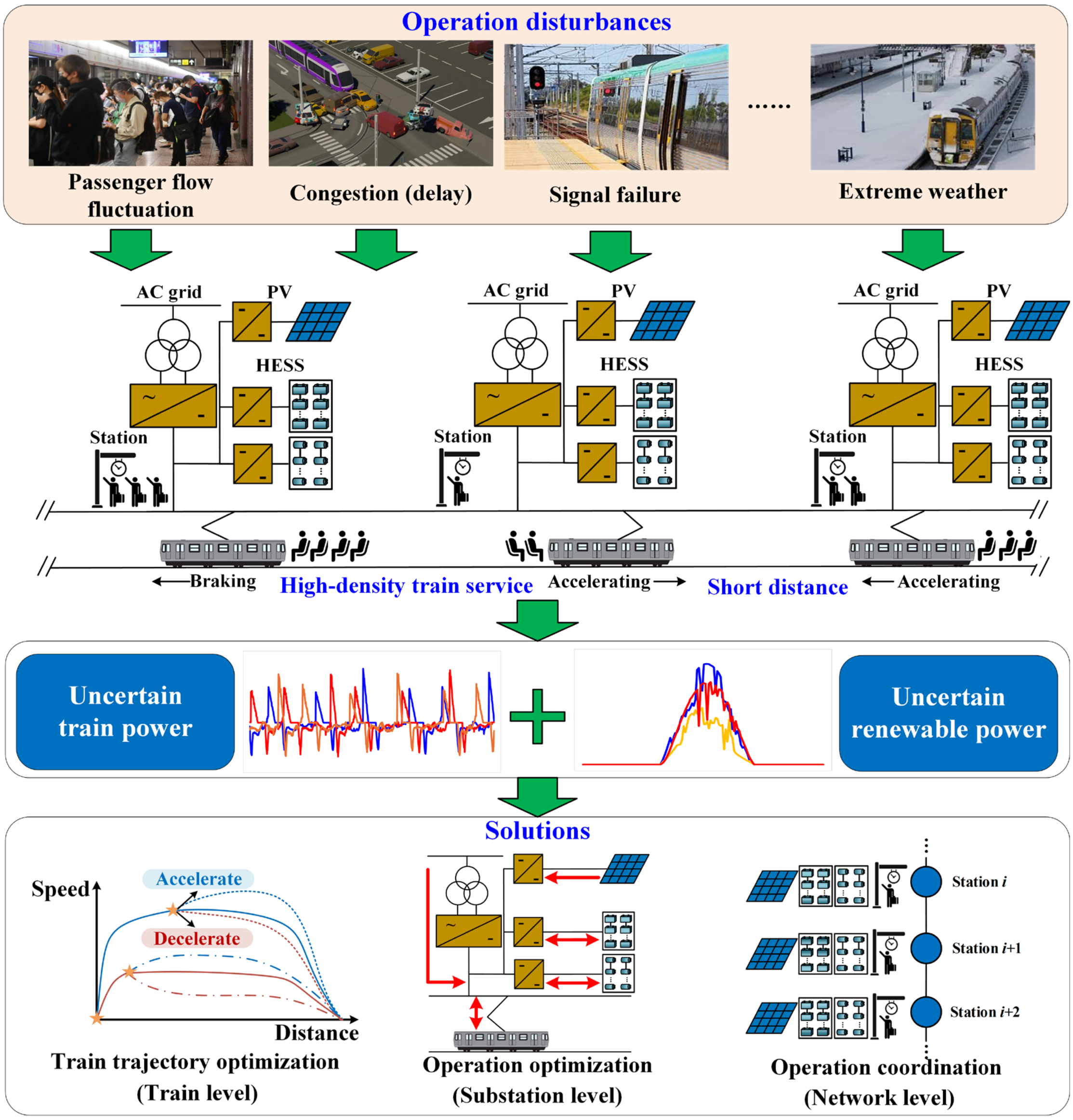


Fig. 1.3 Typical URT TN with HESSs.

In summary, the aforementioned spatial-temporal operation uncertainties and complexities have put forward considerable demands on the stable, efficient, sustainable, and intelligent operations of HESS-integrated TNs, especially for those involving DHESSs. Based on the facts above, this thesis reports using reinforcement learning (RL) [35] as a machine learning base technique to develop energy management and configuration strategies for the optimal operation of HESS-integrated URT TNs from train, substation, and network levels. For the remaining sections, section 1.2 introduces the existing HESS structure and applications in URTs. Section 1.3 reviews

RL principles and algorithms, followed by TN energy management and configuration strategies, where RL applications are especially mentioned. Sections 1.4–1.6 illustrate the technological challenges, research aim and objectives, and thesis outline, respectively.

Table 1.2 Comparison between URTs and high-speed rails.

Item	URT (Subway)	URT (Light rail)	High-speed rail
Right-of-way	 Independent track	 **Mixed traffic**	 Independent track
Range	**Urban areas**	**Urban areas**	Inter-cities
Speed	≤80 km/h	≤30 km/h	**200–350 km/h**
Station distance	**≤3 km**	**≤3 km**	30–60 km
Service frequency	**2–5 min (Peak hour)**	**2–5 min (Peak hour)**	Very low

1.2 Hybrid Energy Storage Systems Applications in Urban Rail Transits

ESSs have been applied to URT TNs for more than twenty years (Table 1.3), where supercapacitors, batteries, and flywheels are common energy storage mediums. Considering the satisfaction to the huge energy and power demand of the URT TNs and the expensive investment and safety issues of flywheels, recent URTs has adopted supercapacitor-battery-based HESSs to improve the application performance of ESSs. The Enviline HESS of the ABB company has operated in the SEPTA subway system of

Philadelphia, USA. In addition to the RB energy recovery function, it also provides balancing services to the local electricity grid. In 2021, Enviline HESS was improved to support 1500 V DC and has been applied to the Melbourne Metro [36]. In China, the 750 V DC HESS has been tested in the Beijing Subway in 2020 [37].

Table 1.3 Typical ESS applications in URTs [7, 36–38].

Year	Type	Location	Parameters	Comment
2003	Supercapacitor	Mannheim	1 kWh	30% energy saving
2003	Supercapacitor	Madrid	2.3 kWh	Voltage stabilization
2010	Li-ion	Kobe	640 V	Energy saving
2010	Supercapacitor	Daejeon	10.4 kWh	Energy saving and voltage stabilization
2012	HESS	Philadelphia	2.2 MW	Energy saving and balancing service
2015	Flywheel	Los Angeles	2 MW	10–18% Energy saving
2020	HESS	Beijing	1 MW	10–20% Energy saving
2021	HESS	Melbourne	12.2 kWh	Energy saving

As shown in Fig. 1.4, for a typical HESS in URTs, the battery and supercapacitor modules are connected in parallel to the DC bus through bidirectional DC-DC converters. Under the charge mode, switch S1 (e.g., IGBT) of both converters is off, and the energy flows from the DC bus to the HESS. Under the discharge mode, switch S2 of both converters is off, and the energy flows from the HESS to the DC bus. The control structure of the HESS can be generally divided into four components: the voltage threshold adjustment strategy, voltage control loop, power allocation strategy, and current control loop (Fig. 1.5). The voltage adjustment strategy determines the charge/discharge voltage threshold U^{CH} and U^{DIS} according to the current system status.

Then, the voltage control loop compares the referential voltage to U^{CH} and U^{DIS}, and generates the referential power by proportional-integral control. Next, the power allocation strategy determines the powers of supercapacitor and battery modules and generates their referential currents. Finally, the current control loop controls the duty ratios of switches (namely, on/off) based on the referential currents. Thus, voltage threshold adjustment and power allocation strategies are the key components to realize intelligent HESS control.

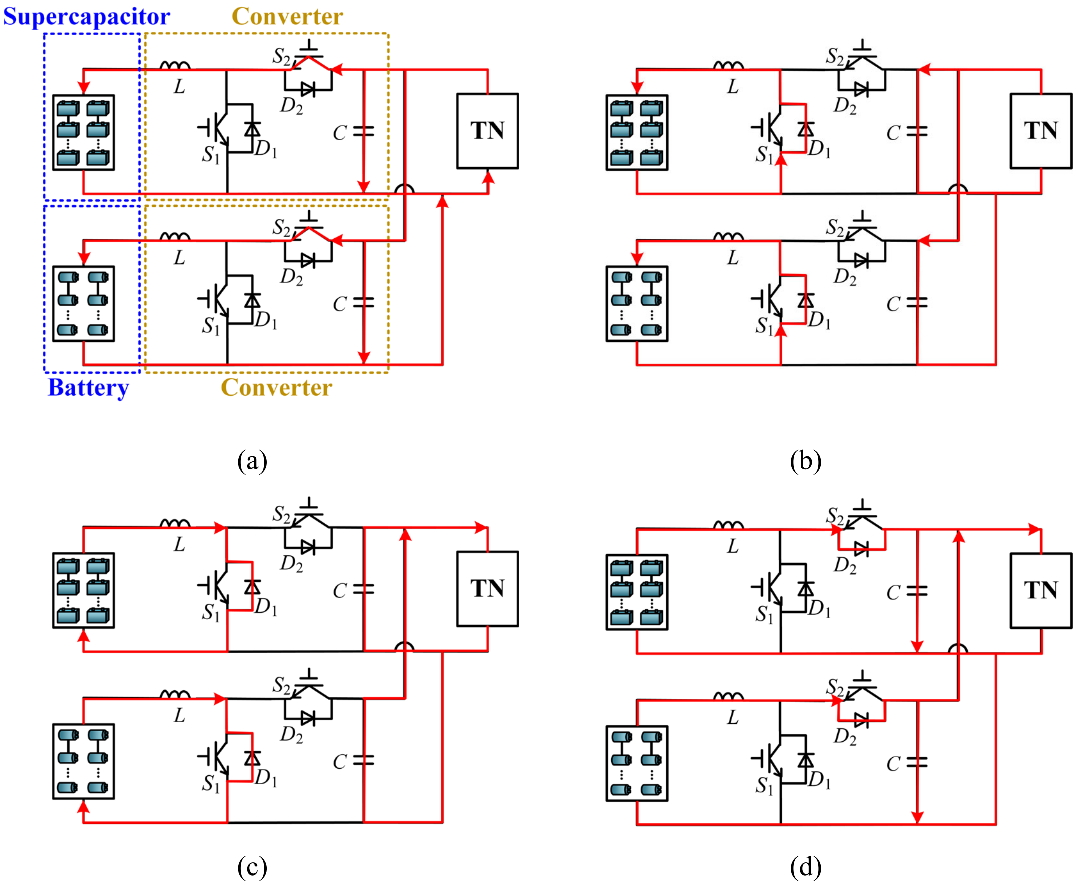


Fig. 1.4 HESS working principles of (a) S2 on, (b) S2 off, (c) S1 on, and (d) S1 off.

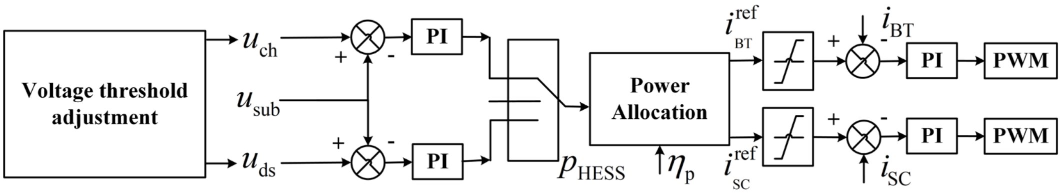


Fig. 1.5 Typical HESS control structure in URTs [37].

1.3 Literature Review

1.3.1 Reinforcement Learning Principles and Algorithms

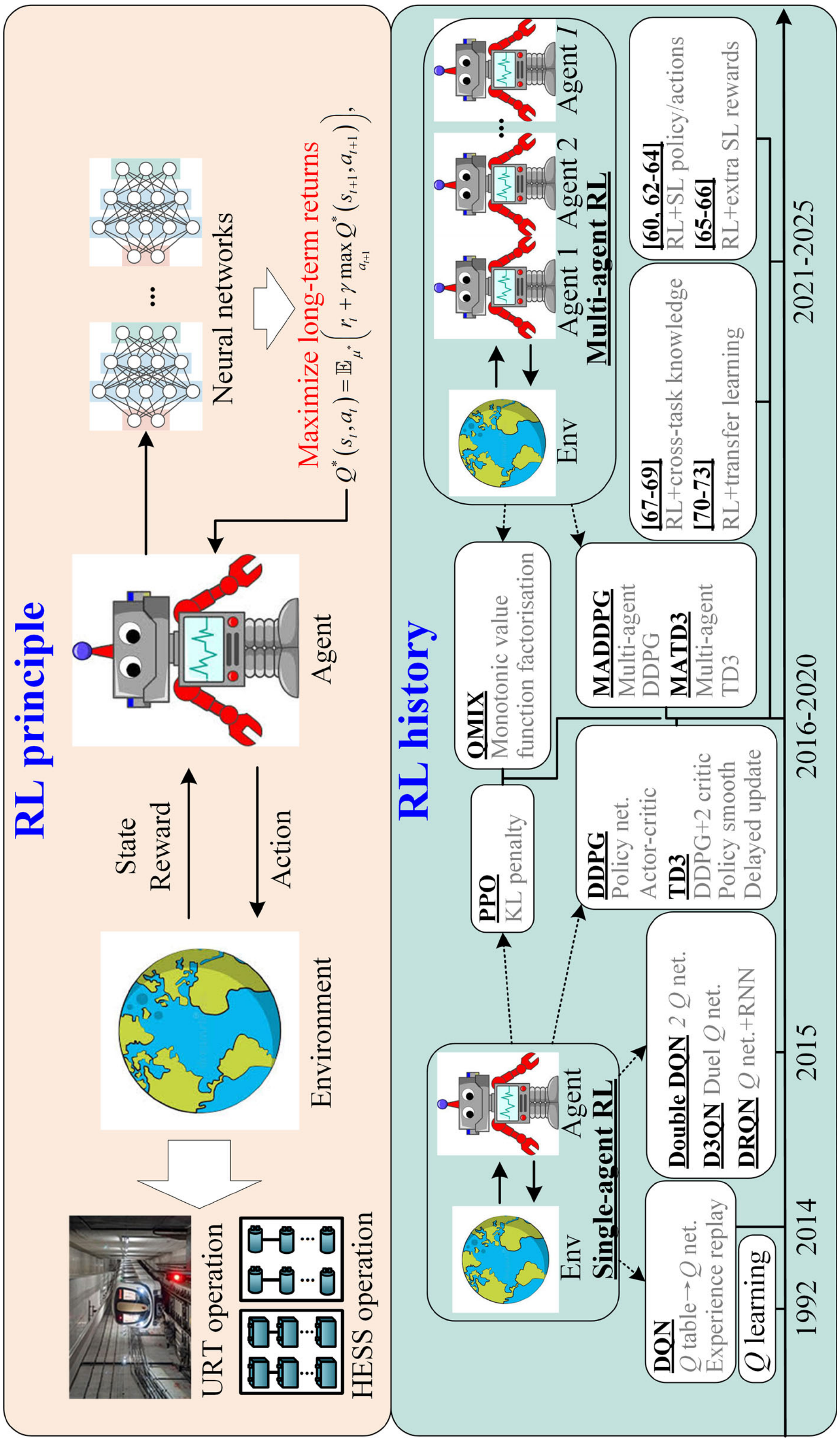


Fig. 1.6 RL principle and history.

1.3.1.1 Reinforcement Learning Principles

RL (Fig. 1.6) is the third basic technique in machine learning, in addition to supervised learning (SL) and unsupervised learning. The goal of RL is to learn a sequential optimal policy (e.g., a control strategy) by continuous interaction with the environment to maximize long-term returns. The RL can be model-free, which means it can estimate the optimal policy based on the experiences gained from the repeated interaction without prior knowledge of the environment model [35]. Compared with conventional model-based optimization methods, the advantages of RL lie in the following aspects: 1) RL avoids frequent execution of a complex optimization model for each environment state. 2) While model-based methods rely on accurate prediction data and exact modeling of uncertainty probability distributions, which are difficult to obtain in practice and the solving process is time-consuming, RL enables adaptive response to varying environment states without knowing these parameters.

The learner of RL is called an agent, and the agent-environment interaction process is defined as a Markov decision process (MDP), which contains five components $\langle \mathcal{S},\mathcal{A},\mathcal{R},\mathcal{P},\gamma \rangle$. $s \in \mathcal{S}$ is the state of the environment, $a \in \mathcal{A}$ is the available action of the agent, $r \in \mathcal{R}$ is the instant reward when state s_t transitions to s_{t+1} at time step t, $\mathcal{P}$ is the state transition function, γ is the discount factor that represents the relative importance between current and future rewards. Besides, μ is the policy, which represents the probability of executing action a_t at state s_t. The MDP assumes the Markov property of all states, which means that the next state only depends on the current state. Based on the MDP, the state-action value functions can be defined as

$$Q(s_t,a_t)=\mathbb{E}_{\mu}\left(\sum\nolimits_{i=t}^{T}\gamma^{(i-t)}r_i \mid s_t,a_t\right), \tag{1.1}$$

where $Q(s_t, a_t)$ is the state-action value function (also known as the Q value), namely, the expected return of executing action a_t at state s_t following the policy μ.

Based on the Bellman function [39], the expected return Q^* under optimal policy π^* (namely, the maximum cumulative discounted return) is

$$Q^*(s_t, a_t) = \mathbb{E}_{\mu^*}\left(r_t + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})\right), \tag{1.2}$$

where by definition, Q^* is the objective of the MDP. Thus, μ^* can be derived by

$$\mu^*(a_t \mid s_t) = \begin{cases} 1, & a = \arg\max Q^*(s_t, a_t), \\ 0, & \text{Otherwise.} \end{cases} \tag{1.3}$$

While the MDP well-defines the environmental interaction process of a single agent, it is not suitable for considering the involvement of multiple agents with cooperation and/or competition relationships. The multi-agent-environment interaction process is described by the Markov game [40], which contains 6 components $\langle I, \mathcal{S}, \mathcal{A}_i, \mathcal{R}_i, \mathcal{P}, \gamma \rangle$. I is the number of agents, $\mathcal{S}$ is the state set observed by all agents, $\mathcal{A}_i$ is the action set of agent *i*, $\mathcal{R}_i$ is the reward received by agent *i*. The relationship of agents can be cooperative, competitive, and mixed. In the cooperative setting, all agents can share a common reward, and they can be regarded as one agent to enable the single-agent RL algorithms. Another more general cooperative setting is to assign different rewards to agents and consider a team-average reward. In the competitive setting, it is essentially a zero-sum Markov game, where the reward of one agent is the loss of the other, and the sum of agents' rewards is zero. The Nash equilibrium will yield a robust policy that optimizes the worst-case long-term return. In the mixed setting, no clear constraint is imposed on agents.

1.3.1.2 Reinforcement Learning Algorithms

Initially, the classic Q learning algorithm was proposed for a fully observable environment with discrete action space [41]. Then, RL was enhanced by integrating deep neural networks for value and/or policy representation, which makes it suitable for handling high-dimensional complex decision-making problems.

Single-agent RL algorithms can be divided into value-based and policy-based algorithms. The value-based RL aims to learn the value of states and Q functions, where the deep Q network (DQN) [42] was the first attempt. In DQN, an experience reply mechanism was used to store the agents environmental interaction data . At each training step, the DQN updates the parameters by minimizing the loss

$$\mathcal{L}=\mathbb{E}_{r_w\sim\mathcal{R},s_w\sim\mathcal{S}_\mu}\left[r_w+\gamma\max_{a_{w+1}}Q_\theta\left(s_{w+1},a_{w+1}\right)-Q_\theta\left(s_w,a_w\right)\right]^2, \tag{1.4}$$

where data $\left(s_w,a_w,s_{w+1},r_w\right)$ are randomly extracted from the experience replay buffer. θ is the deep neural network parameter. Based on DQN, various extensions were developed to improve its performance [43–46].

Policy-based algorithms directly learn the optimal policy μ^*. For instance, in [47], two networks (the actor and critic) were utilized to update the value function and policy function parameter, respectively. In [48], parallel actor-environment interactions were allowed, and asynchronous training was executed. In [49], the Kullback–Leibler divergence constraint was applied to policy updates for training stability. In [50], a truncated alternative objective function is used for simplifying the above method. In [51], a deep deterministic policy gradient was proposed, which implemented a soft update on target networks for training stability and added stochastic noises in the actor for exploration efficiency. In [52], TD3 was proposed to improve the above method by suggesting delaying policy updates for mitigating value overestimation.

For multi-agent environments, the initial attempt is to take other agents as part of the environment and solve by single-agent algorithms [53]. Although satisfactory performance can be achieved in practice, convergence is not theoretically guaranteed.

According to the training paradigms, multi-agent RL algorithms can be divided into distributed and centralized training algorithms. Distributed training assumes information exchange between agents through communication networks to mitigate the non-stationary training issue. Generally, the agents share network weights or gradients (namely, parameter sharing) [54, 55]. For centralized training, it assumes a centralized controller that can collect the joint states, actions, and rewards of all agents. Considering the convenience of practical applications, extensive efforts have been made on centralized training and decentralized execution architecture. In this regard, the representative value-based multi-agent RL algorithms are VDN and QMIX [39, 56], which learn a centralized Q value function. The representative policy-based algorithms include COMA, MADDPG, and MATD3 [57–59], which utilize joint state and action to train the centralized critic network.

Despite the impressive progress, RL training was less efficient in complex environments with rich and dynamic data [60, 61]. To address this issue, one direction is to utilize a combination of RL and SL (namely, supervised RL), which has been explored in areas of recommendation systems [62], healthcare treatment [63–65], and automatic driving/navigation [60, 66]. In [62–64], SL provides a supervision signal to RL to learn a hybrid policy. Since RL may suffer from value overestimation, the introduced supervision can be a counterbalance to avoid focusing on the long-term return while sacrificing too much. In [60], supervised training data were utilized to provide candidate action for RL agents. Moreover, in [65, 66], self-supervised learning was implemented to generate extra rewards for RL calibration.

On the other hand, the complexity of the RL task can be reduced by decomposing it into multiple components or subtasks. The multi-task RL aims to train a single and universal policy that can be applied to a set of tasks, where each task has a unique MDP. In [67–69], the RL agent leverages shared knowledge between tasks. In [70], a well-learned RL policy is utilized as a start point to be transferred to new tasks via transfer learning. Besides, several multi-task learning techniques, such as gradient manipulation [71, 72] and loss function weight adaption [73], were introduced into RL training updates to address cross-task conflicts. Besides multi-task RL, another task-based approach is curriculum RL [74], which trains on simple tasks and gradually increases the task difficulty to improve learning efficiency. For instance, in [75, 76], curriculum RL was implemented to generate energy-saving car-following strategies. In [77], curriculum RL was utilized for energy management optimization of hybrid electric vehicles.

In summary, as a novel sequential decision-making method, RL has the advantages of being model-free and able to handle high-dimensional nonlinear objectives, which has inspired this thesis to introduce it into the energy management and configuration of HESS-integrated URT TNs.

1.3.2 Energy Management and Configuration Strategies

1.3.2.1 Energy Management Strategies

1) Train Level: Train-level energy management aims to realize energy-efficient train driving, where the train trajectory is regulated automatically to minimize traction energy consumption while ensuring other objectives, such as safety, punctuality, and passenger satisfaction of services [8, 78]. In this regard, the onboard automatic train operation (ATO) system is responsible for determining all train acceleration and braking commands (namely, train trajectories) through rigorous computation [79].

Various train trajectory optimization (TTO) methods have been developed for the ATO operation, which can be categorized as analytical, numerical, heuristic, and RL-based methods. The analytical methods [78, 80–84] are based on Pontryagin's maximum principle [85], where the optimal train trajectory was defined as a sequence of optimal control regimes and their switching points. Although the theoretical optimal solution can be guaranteed, finding such a solution can take substantial computational time since real-world URT lines include special sections such as curves, slopes, and tunnels. Numerical methods [86–91] (e.g., dynamic programming, pseudo-spectral method, etc.) were proposed to find near-optimal solutions within feasible computational time. Alternatively, heuristic methods [92–96] (e.g., genetic algorithm (GA)) can provide solutions, but their theoretical optimality is not always guaranteed.

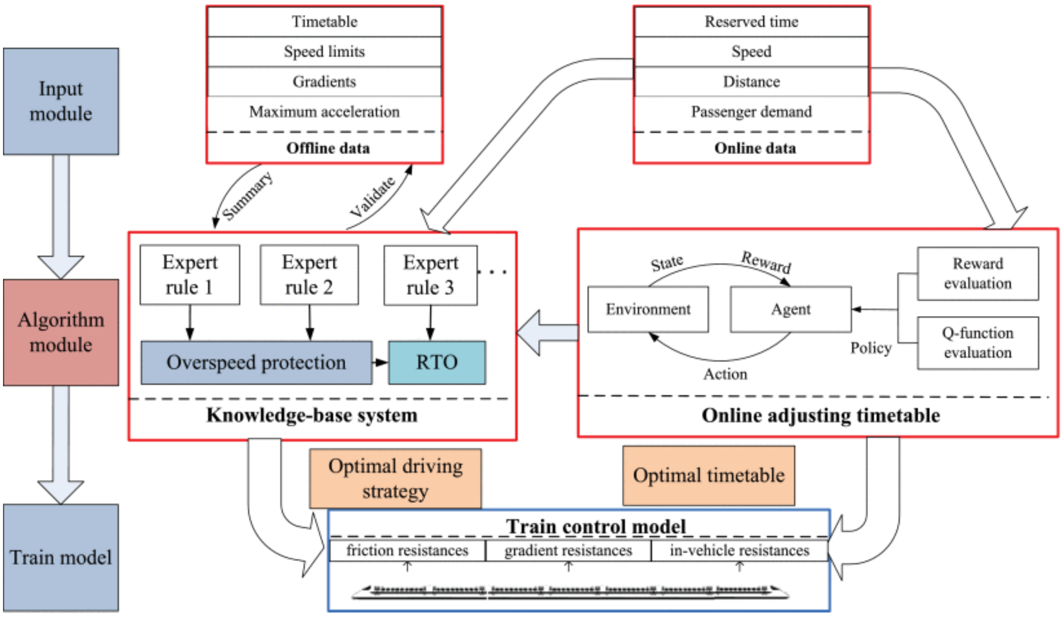


Fig. 1.7 Example of RL-based TTO model design [97].

However, most studies assume train trajectories to be optimized and embedded on the ATO in advance of real-time operation [98], which is insufficient to address real-time train operation disturbances. Moreover, these disturbances and uncertain train parameters [90] bring considerable challenges to the URT system modeling and forecasting. Compared with the above model-based TTO methods, the model-free RL algorithms can effectively handle these uncertainties. In [97], RL is first introduced into the TTO, where it was capable of adjusting trajectories dynamically between two stations. They [32] further utilized *Q* learning to address discrete train forces and uncertain delays, where the results show superior performance over MD and the existing ATO system. In [99], RL was combined with the long short-term memory to enhance its TTO performance. However, modern URT trains can output continuous traction force, and the dynamic trajectory adjustment capability of trains within the full running time range was not tested. In [100], RL was combined with expert knowledge rules to solve the TTO and can deal with the continuous train traction force. In [101], RL was combined with a reference system for proactive operation constraint handling. However, the on-road accidents and the dynamic trajectory adjustment capability were not considered in these works.

2) Substation Level: Based on non-real-time (historical and forecasting) load and RDG data, various energy management strategies for traction substation with an ESS have been proposed, which were mainly based on stochastic programming [30, 102], robust optimization [103], model predictive control [104, 105], and heuristic methods [106]. They are generally carried out on a time scale from minutes to days. Nevertheless, the real-time uncertainty and volatility of the traction load, RB, RDG output, passenger demands, etc., have not been addressed.

The real-time energy management strategies, alternatively, can be divided into rule-based, optimal-control-based, and RL-based strategies. In terms of rule-based strategies, the optimal operation rules can be either determined by tracking ESS state-of-energy (SoE) [107, 108] or comparing the TN voltage with its voltage thresholds [20, 109–112]. Some recent studies have investigated dynamic threshold control to better adapt to the URT TN operation characteristics [37, 113]. However, only [20] has taken RDGs into account. Besides, these rule-based strategies are easy to implement but heavily depend on intuition and experience. Considering the dynamic operation environment of TNs, the optimal rules can be difficult to set. Regarding optimal-control-based strategies, in [114] and [115], the optimal strategy was obtained by the Euler-Lagrange equation and Lagrange multipliers, respectively. These methods can handle a specific operation condition, while their adaptability to various train operation conditions can be inadequate. In [116], dynamic programming was applied to minimize the braking resistor loss. However, dynamic programming-based methods can suffer from the curse of dimensionality. In [117], a hierarchical control strategy was proposed, where a state machine was introduced in the energy management layer, and a multi-objective optimization algorithm was proposed in the converter layer. In [118], a comprehensive model integrating train control, substation output, and HESS was developed, and a model predictive control framework was proposed to minimize energy consumption. In [119], a bi-level multi-objective optimization was performed considering substation operation stability based on particle swarm optimization with compression factor. Nevertheless, their performance is affected by prediction accuracy.

So far, only a few studies have been presented on RL-based strategies. In [120], DQN was utilized to adjust the voltage thresholds of a supercapacitor-based ESS for energy saving and voltage stabilization. In [121], TD3 was adopted to allocate HESS power with similar objectives. In [122], a parallel RL framework was established to improve energy utilization efficiency with a fast convergence speed.

3) Network Level: With the increasing integration of RDGs into URT TNs, a few recent studies have concerned the coordinated operation of multiple networked substations with DHESSs. Similar to the substation level, some studies [123–126] have formulated day-ahead and intraday scheduling plans based on non-real-time data. For instance, in [123], the optimal operation of a TN with PVs, wind turbines, supercapacitors, and batteries was formulated as a multi-period optimal power flow problem and solved by nonlinear programming. For real-time strategies, in [127], a novel optimization method based on GA was proposed to jointly consider the energy management, location, and size of supercapacitor-based distributed ESSs for obtaining optimal economic efficiency and voltage profile. In [128], a control strategy based on energy transfer was proposed for peak power shaving. According to the load characteristics of the URT TN, part of the RB energy absorbed by battery-based distributed ESSs was transferred from off-peak hours to peak hours for release. In [129], A dynamic priority-based power allocation strategy was developed to optimize the DHESS operation, where the on-board supercapacitors were utilized to accommodate the train traction energy demand, and the in-station batteries were responsible for voltage stabilization. In [130], a multi-time scale coordinated energy management strategy was proposed for supercapacitor-based distributed ESSs based on a genetic and fuzzy algorithm. The simulations reported superior performance compared with existing strategies using a single time scale. For RL-based strategies, in [131], a decentralized multi-agent cooperative control algorithm was proposed for the coordination of supercapacitor-based DHESSs. Nevertheless, these strategies were performed without considering RDGs. In [132], a real-time control strategy under a centralized control scheme was presented for a multi-source traction system integrating the conventional TN, PVs, wind turbines, and ESSs. The case studies reported an energy-saving rate of 36% and a peak power reduction rate of 46%. In [133, 134], a centralized-decentralized energy management framework was developed to address multiple operation objectives of a mainline railway and was evaluated by a field test.

Furthermore, several real-time energy management strategies have been developed for other electrified railways [135–138]. However, these approaches may not be suitable for HESS-integrated URT TNs since the operation characteristics of URTs (i.e., headway changes and passenger flow fluctuation) and their impacts on PV–RB energy utilization require further consideration.

1.3.2.2 Configuration Strategies

In this subsection, the current status of existing research on the optimal configuration strategies of HESS-integrated URT TNs in terms of the train trajectory and ESS is introduced. Conventionally, each running section is equipped with 3–5 available train trajectories for the ATO system selection, where each trajectory makes a trade-off between running time and energy consumption to different degrees [139–141]. Recently, considering the short distance between stations and the increasingly frequent train services to address passenger demands, in [142, 143], a multi-objective particle swarm optimization were performed to generate Pareto front of train trajectories for more flexible train scheduling. In [144], considering the uncertainties in train operation, a two-stage energy-efficient timetable design method was proposed to reduce train delays based on an optimal running time-energy consumption solution set. In [145], a preference dominance criterion was proposed to handle the train mass uncertainty, and a set of performance-robust driving schemes can be obtained. However, these Pareto solutions only consider energy saving and punctuality as the main objectives, while other factors, such as safety and riding comfort, need to be considered simultaneously.

So far, the optimal configuration strategies for ESSs primarily consider the impact of their energy management strategies, TN topologies, and/or train service patterns [121, 146–150]. To name a few, In [146], the capacities of ESSs were determined by predicting the maximum RB energy delivered to each substation. However, the power and energy limits of an actual ESS were ignored, which can lead to practically infeasible configuration solutions. In [149, 150], deterministic and stochastic programming were performed to investigate the optimal configurations and energy management strategies of the ESS jointly under different operation scenarios for a catenary-free tramline. In fact, the energy flows of the TN are the result of the combined impacts of trains, substations, and ESSs. Recently, some studies have investigated the joint optimization of ESS and other TN parameters. In [151], the energy-efficient train timetable and onboard ESS capacity allocation were jointly optimized, where each train was allowed to carry an onboard ESS with different capacities. Although the RB energy utilization shows a significant increase, the simplified physical model of the ESS (e.g., ignoring ESS size and weight limits) undermines the effectiveness of the result. In [152], a timetable optimization model considering ESS installation was proposed, where its performance under several ESS configuration scenarios was compared. In [153], another timetable optimization model with minimum time overlap was developed to match the deceleration and acceleration time of trains with the ESS working properties. The theoretical maximum RB energy was obtained, which can be a reference to set ESS capacities. Nevertheless, the energy conversion loss, transmission loss, and the ESS physical power limits were not considered in these studies. In [154], the ESS size, train timetable, RB control parameter, and no-load voltage were synthetically optimized based on a multi-train operation simulator. However, the passenger flow fluctuations were ignored.

Currently, the research on ESS configuration is primarily carried out without considering the integration of RDGs. Besides, the optimal capacity allocation among different energy sources and locations of DHESSs requires further investigation. More importantly, the optimal synergy of the configuration strategy, energy management strategy, and train operation parameters (e.g., RB control parameter and timetable) has been crucial for the optimal planning and operation of HESS-integrated URT TNs.

1.4 Technological Challenges and Limitations

Compared with various model-based energy management and configuration strategies, such as rule-based and optimal-control-based (including heuristic-based) strategies, the RL-based strategies do not rely on accurate system modeling and uncertainty predictions [79, 155] while generating end-to-end sequential decision-making solutions for online applications instead of frequent execution of a complex optimization model for each environment state. Leveraging these advantages, some achievements of RL-based strategies have been made regarding the energy management and configuration issues within the HESS-integrated URT TNs. However, the spatial-temporal uncertainties and complexities of the URT TN operation environment arising from passenger demand, urban traffic congestion, widespread distribution, operational disturbances, etc., plus the pressing need for carbon-neutral transformation have imposed significant challenges and limitations to existing RL-based strategies in terms of the stable, efficient, sustainable, and intelligent operations of the HESS-integrated URT TNs, especially for those involving DHESSs. Specifically:

1) **At the train level,** due to the short running times between stations and frequent disturbances in URTs, even short delays that last for several seconds can lead to knock-on delays, resulting in extra traction energy use and decreased ride comfort to guarantee punctuality. The insufficient capability of the ATO-based technologies in terms of calculation and adjustment of energy-efficient trajectories online in response to uncertain disturbances and adapting to rescheduled trip times significantly limit their application prospects. Thus, it is urgent to address the energy-efficient TTO (EETTO) and its associated safety, punctuality, and ride-issues under real-time uncertain disturbances to enhance the ATO performance. However, existing RL-based methods [32, 97, 99–101] only addressed one or several objectives in this regard, and therefore, a more thorough consideration is needed. Besides, viable trajectory configuration suggestions based on the EETTO are required for practical use.
2) **At the substation level,** regarding HESS-integrated URT TN operation, few studies [119, 154] focused on both sizing and real-time control, while none of the sizing strategies was incorporated with an RL-based energy management strategy. Although several RL-based energy management strategies have been employed [120–122], they focused on learning individual strategies for each specific train headway and/or train mass task. Considering the same TN topology these tasks shared, it is beneficial to leverage shareable cross-task experience to improve RL performance and data efficiency. In addition, the joint optimization of voltage threshold adjustments and power allocations to fully explore the flexibility in the HESS power regulation has not been involved, substantially undermining its economic operation.

3) **At the network level,** for DHESS-integrated URT TNs operation, besides the above substation-level challenges, on a short time scale (within seconds), the coordination of DHESSs for the optimal complementation of solar, RB, and electricity energy and real-time uncertainties lacks a thorough investigation [131, 138]. More importantly, it is necessary to adopt an adaptive and decentralized DHESS control scheme to handle single-point failure, communication burden, and scalability issues. On the other hand, on a long time scale (sub-hourly or hourly), the daily train service pattern changes and temporal mismatches between peak PV output and peak passenger demand require optimal dispatches of DHESS outputs. Nevertheless, except for the heuristic-based strategy in [130], the synergetic consideration of multiple time scales has not been addressed.
4) In addition, existing works have not fully characterized the multi-source operation uncertainties of HESS-integrated URT TNs, e.g., only a few studies [148, 156, 157] partially considered the impacts of passenger flows, delays, and/or temporary traffic regulations on the HESS and/or DHESS control. Moreover, the spatial-temporal uncertainties and correlations of dispersed PVs and passenger demands have strengthened the complexity of the energy management and configuration problem.

1.5 Research Aim and Objectives

The aim of this thesis is to develop RL-based energy management and configuration strategies for HESS-integrated URT TNs, targeting three different levels of automatic train operation, HESS-integrated traction substation operation, and DHESS-integrated TN operation. In detail, it includes the following objectives:

1) **At the 1st (train) level,** developing a supervised RL-based energy-efficient train trajectory optimization (SRL–EETTO) approach to expand the ATO system capability in addressing the real-time responsiveness and dynamic online challenges to energy-efficient TTO and its associated safety, punctuality, and ride comfort issues.
2) **At the 2nd (substation) level,** proposing a multi-task RL-based sizing and control optimization (MTRL–SCO) approach to enhance the coordinated operations of HESSs and their integrated traction substations under dynamic spatial-temporal traffic of URTs.
3) **At the 3rd (network) level,** presenting a multi-task multi-agent RL-based multi-time scale energy management (MTMARL–MTSEM) approach to promote the economic and low-carbon operation of DHESS-integrated TNs considering operation uncertainties of URTs and RDGs.
4) **Furthering the 3rd (network) level,** extending a multi-task multi-agent RL-based data-driven multi-objective configuration optimization (MTMARL–DDMOCO) approach to improve the synergy between the economic and energy efficiencies of DHESS-integrated TN operation and the travel time of the passengers.

1.6 Thesis Outline

This thesis presents the work on RL-based energy management and configuration for URT TNs with HESSs as follows.

Chapter 1 introduces the research background and significance, followed by literature reviews of RL algorithms, typical energy saving and emission reduction measures, energy management strategies, and configuration strategies. Then, the rest of this chapter analyzes the technological challenges of existing studies and summarizes the research aim and objectives of this thesis.

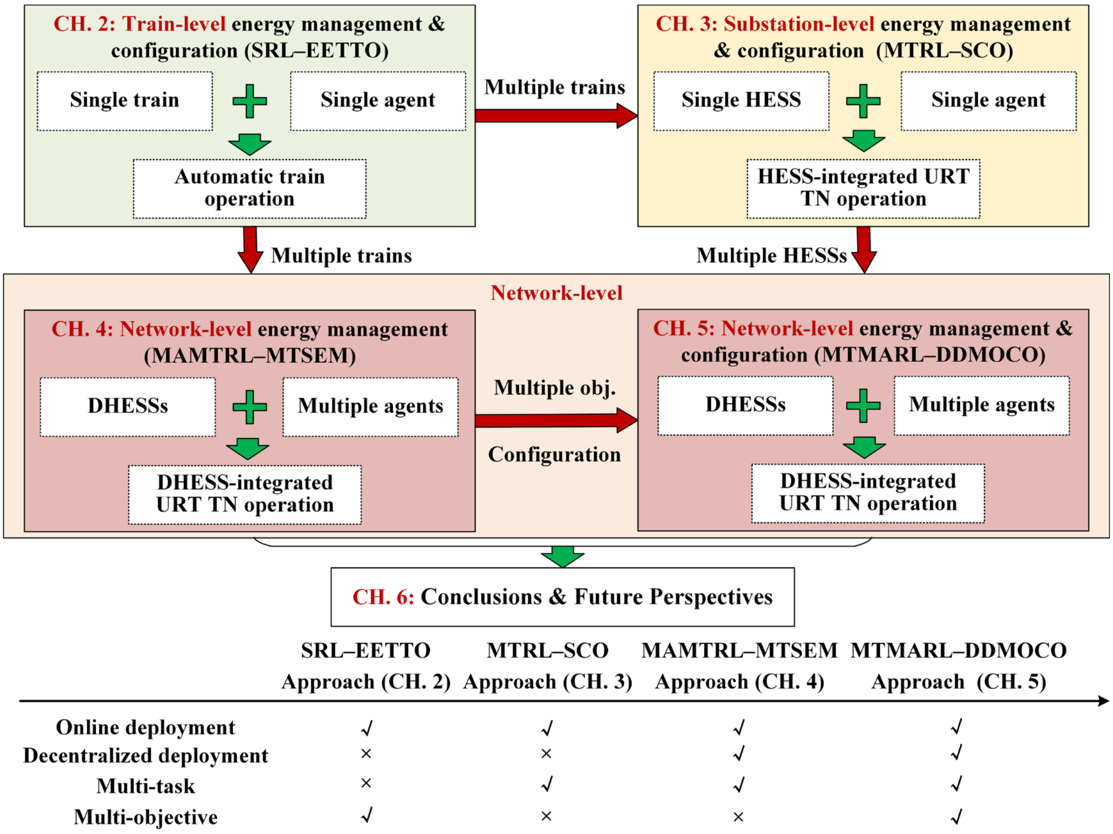
Chapter 2 begins with the formulation of the train control model considering real-time train operation disturbances. Then, the proposed SRL–EETTO approach is presented with MDP formulation for model design, followed by developing a supervised twin-delayed deep deterministic policy gradient algorithm for model training. Finally, case studies are investigated for model verification.

Chapter 3 illustrates the structure and modeling of HESS-integrated traction substations in TNs, followed by the formulation of the HESS sizing and control optimization model and the analysis of HESS control parameters and URT operation uncertainties on the operation cost and RB energy utilization. Then, the proposed MTRL–SCO approach is developed with the formulation of the dynamic traffic model, the multi-task MDP, and a novel KT-D3QN algorithm. Finally, the effectiveness of the proposed MTRL–SCO approach is validated.

Chapter 4 starts with the structure and modeling of DHESS-integrated TNs. Then, the tri-level framework of the proposed MTMARL–MTSEM approach is formulated, including a two-stage stochastic scheduling model at the upper and middle levels and a real-time energy management algorithm at the lower level. Representative daily TN operation scenarios are selected to demonstrate the performance of the proposed MTMARL–MTSEM approach.

Chapter 5 states the formulation of the multi-objective optimization model considering the electrothermal aging of batteries. Then, the proposed MTMARL–DDMOCO approach is presented, consisting of the ensemble-learning-based load prediction modeling and the data-driven implementation of the non-dominated sorting genetic algorithm based on the developed MTMARL–MTSEM approach. Finally, the configuration decisions of the proposed MTMARL–DDMOCO approach are analyzed thoroughly.

Chapter 6 concludes the major findings of this thesis and indicates the future directions of the following works.


CH. 2: Train-level energy management & configuration (SRL–EETTO)
Single train
Single agent
Automatic train operation
Multiple trains
CH. 3: Substation-level energy management & configuration (MTRL–SCO)
Single HESS
Single agent
HESS-integrated URT TN operation
Multiple trains
Multiple HESSs
Network-level
CH. 4: Network-level energy management (MAMTRL–MTSEM)
DHESSs
Multiple agents
DHESS-integrated URT TN operation
Multiple obj.
Configuration
CH. 5: Network-level energy management & configuration (MTMARL–DDMOCO)
DHESSs
Multiple agents
DHESS-integrated URT TN operation
CH. 6: Conclusions & Future Perspectives

	SRL–EETTO Approach (CH. 2)	MTRL–SCO Approach (CH. 3)	MAMTRL–MTSEM Approach (CH. 4)	MTMARL–DDMOCO Approach (CH. 5)
Online deployment	√	√	√	√
Decentralized deployment	×	×	√	√
Multi-task	×	√	√	√
Multi-objective	√	×	×	√

Fig. 1.8 Thesis outline.

Chapter 2: Energy-Efficient Train Trajectory Optimization for Automatic Train Operation Based on Supervised Reinforcement Learning

Nomenclature in this chapter

A. Supervised Reinforcement Learning Elements

a , $\tilde{a}$	Actions of agent and target agent
B	Replay buffer
DP	Delayed policy update frequency
J'_a , J_a , J_s , J_c	Weighted actor loss, actor loss, supervision loss, and critic loss
Q^* , Q_∞	Expected return and its value when $r_\mathrm{g} = r_\infty$
r , r_g , r_∞	Reward, goal-state reward, and reward per time t
r_T , r_E , r_C	Coefficients of r_g
s , s^s , s_g	Agent state, supervisor state, and goal state
α_s	Weight coefficient of supervision loss
μ^* , μ , μ_sl	Optimal agent policy, agent policy, and supervisor policy
γ	Discount factor
τ	Soft update rate
ξ_a , ξ_c	Learning rates of actor and critic networks
$\mathcal{A},\mathcal{R},\mathcal{S},\mathcal{P}$	Sets of actions, rewards, states, and transitions
ϕ , ϕ'	Parameters of actor and target actor networks
θ , θ'	Parameters of critic and target critic networks

B. Indices

$i \in \{1, 2, \cdots, I\}$ Index of stations or traction substations

$m \in \{1, 2, \cdots, M\}$ Index of supervisors

$w \in \{1, 2, \cdots, W\}$ Index of transitions sampled from replay buffer

C. Time Scales

t, T Current time step and time horizon on a short time scale for real-time train and HESS control (e.g., sub-minutely)

D. Variables

$a^{\mathrm{TR}}, v^{\mathrm{TR}}, x^{\mathrm{TR}}$ Train acceleration (m/s^3), speed (m/s), and position (m)

$\mathcal{D}^{\mathrm{bd}}, \mathcal{D}^{\mathrm{do}}$ Disturbances before departure and during operation (s)

F^{TB} Traction or braking force of the train (N)

M^{TOT} Total train mass (kg)

$R_{\mathrm{b}}, R_{\mathrm{l}}, R_{\mathrm{c}}, R_{\mathrm{t}}$ Basic, line, curve, and tunnel resistances (N)

$t_{\mathrm{d}}^{\mathrm{bd}} \sim \mathcal{T}^{\mathrm{bd}}, t_{\mathrm{d}}^{\mathrm{do}} \sim \mathcal{T}^{\mathrm{do}}$ Running time changes and their distributions caused by disturbances $\mathcal{D}^{\mathrm{bd}}$ and $\mathcal{D}^{\mathrm{do}}$ (s)

$T^{\mathrm{ACT}}, T^{\mathrm{PL}}, T^{\mathrm{RTTR}}$ Actual, planned, and rescheduled running times (s)

$x_{\mathrm{e}}^{\mathrm{ACT}}, x_{\mathrm{e}}$ Actual and planned end positions where the train stops (m)

x_{d} Position where the train receives rescheduling notifications (m)

x^{TU} Length of tunnel (m)

α^{SA} Slope angle (°)

α^{R} Curve radius (m)

E. Parameters

$a_{\mathrm{lim}}^{\mathrm{ACC}}, a_{\mathrm{lim}}^{\mathrm{DCC}}$ Train acceleration and deceleration limits (m/s^3)

a_{eb}^{DCC} Minimum deceleration for emergency train braking (m/s^3)

b	Train acceleration or deceleration constraints according to expert knowledge rules (m/s^3)
c_1, c_2, c_3	Coefficients of the Davis formula
I_v, I_p, I_t, I_e, I_c	Evaluation Indices for on-road speed limit violation (times), stopping accuracy (m), punctuality (s), energy saving (J/(km·kg)), and ride comfort (m/s^3)
$T_{\text{lim}}, E_{\text{lim}}, C_{\text{lim}}, p_{\text{lim}}$	Tolerances for punctuality (s), energy saving (J/(km·kg)), ride comfort (m/s^3), and stopping accuracy (m)
$v_{\text{lim}}^{\text{TR}}$	On-road speed limits (m/s)
ξ_r	Train rotating mass factor (kg·m^2)
Δx	Safe braking distance (m)

2.1 Background

Nowadays, scholars have investigated various energy-efficient train trajectory optimization (EETTO) methods for enhancing automatic train operation (ATO) system performances (literature reviews in section 1.3). Nevertheless, the EETTO and its associated ride-comfort, punctuality, and safety issues under uncertain disturbances and rescheduled trip times in modern urban rail transits (URTs) require comprehensive consideration (train-level challenges in section 1.4). Therefore, this chapter focuses on developing a three-step supervised reinforcement learning-based energy-efficient train trajectory optimization (SRL–EETTO) approach for intelligent automatic train operation (iATO) by hybrid-integrating reinforcement learning (RL) and supervised learning (SL) at the 1st (train) level. Specifically, the main contributions of this chapter are outlined as follows:

- The real-time train control model under uncertain disturbances is formulated as a Markov Decision Process (MDP). A binary function-based goal-directed reward design method is proposed to systematically integrate multiple real-time train operation objectives associated with energy saving, ride comfort, punctuality, and safety into the MDP. A fine-tuning process based on bilinear programming is used to correct RL reward parameters toward optimal states.
- A two-step supervisor-actor-critic (SAC) architecture based on a supervised twin-delayed deep deterministic policy gradient (S-TD3) is devised to generate optimal train trajectories online. While the first step develops multiple EETTO models to obtain optimal fixed-time train trajectories, the second step improves model generalization capability within the practical running time range by simultaneously optimizing traction energy efficiency and learning supervisory actions from pre-trained EETTO models.

Finally, simulations are implemented to validate the effectiveness of the SRL–EETTO. Section 2.2 states the problem formulation, including the illustration of the real-time train operation process and the train control model. Section 2.3 illustrates the SRL–EETTO approach, including an overview of the optimization process and EETTO model design, training, and verification. Section 2.4 reports case studies and their results. Section 2.5 gives the summary.

2.2 Problem Formulation

2.2.1 Automatic Train Operation Principle

The ATO system operates under an advanced onboard automatic train control (ATC) system, and the ATC system is responsible for regulating the train operation according to one or several referential train trajectories and fulfilling specific operational requirements such as safety and punctuality. Besides the ATO system, a typical ATC system [12] also consists of an automatic train protection (ATP) system and an automatic train supervision (ATS) system (Fig. 2.1).

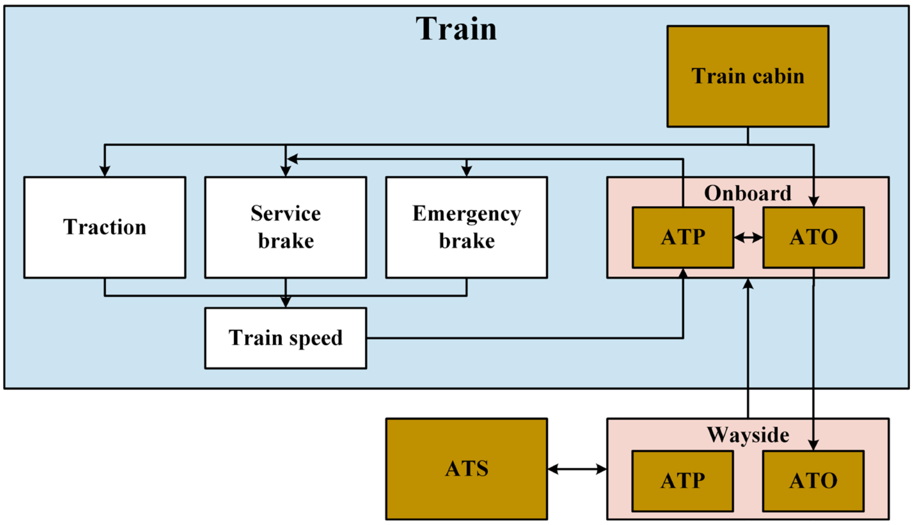


Fig. 2.1 ATC system realization.

The primary function of the ATP system is to prevent train accidents by automatically monitoring and controlling the trains speed and movement according to the safety speed profile and the safe operating distance between the trains. The ATS system is responsible for train schedule creation and updates, automatic routing optimization, operation data collection and analysis, and train status supervision. The ATO system calculates the optimal referential train trajectories according to the predetermined timetable and line data.

Fig. 2.2 illustrates the real-time train operation process. The planned running times between stations are denoted as $T_{1,2}^{\mathrm{PL}}$, …, $T_{i,i+1}^{\mathrm{PL}}$, …, $T_{I-1,I}^{\mathrm{PL}}$. When disturbances occur, the train timetable can be rescheduled, where the planned trip time $T_{i,i+1}^{\mathrm{PL}}$ between station i and i+1 can be changed to $T_{i,i+1}^{\text{RTTR}}$ in real-time. Based on $T_{i,i+1}^{\text{RTTR}}$, the train trajectory can be re-optimized by the ATO. Correspondingly, the actual running times can be denoted as $T_{1,2}^{\text{ACT}}$, …, $T_{i,i+1}^{\text{ACT}}$, …, $T_{I-1,I}^{\text{ACT}}$. From the fact above, finding the optimal train trajectories online is crucial to the ATO system operation.

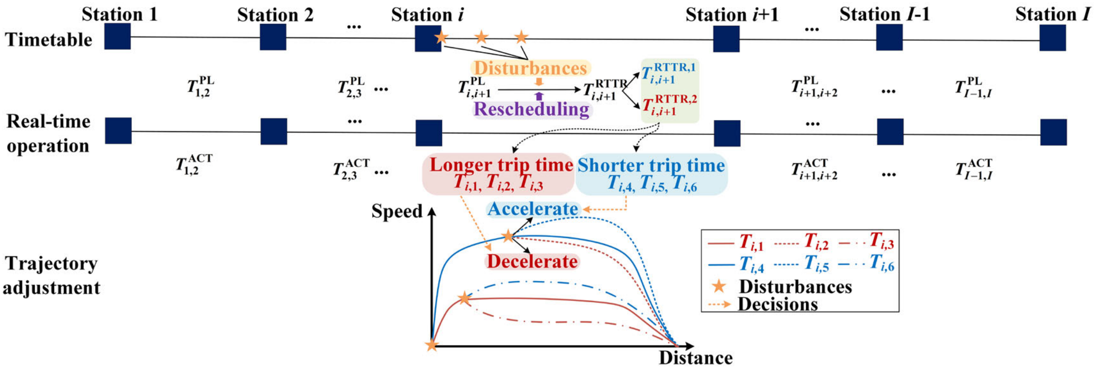


Fig. 2.2 Real-time train operation process.

2.2.2 Train Control Model

Generally, the train motion equation is written as

$$M^{\text{TOT}}\xi_r v^{\text{TR}}(x^{\text{TR}})\frac{\mathrm{d}v^{\text{TR}}(x^{\text{TR}})}{\mathrm{d}x}=F^{\text{TB}}(v^{\text{TR}})-R_{\text{b}}(v^{\text{TR}})-R_{\text{l}}(x^{\text{TR}}), \tag{2.1}$$

$$\frac{\mathrm{d}t(x^{\text{TR}})}{\mathrm{d}x^{\text{TR}}}=\frac{1}{v^{\text{TR}}(x^{\text{TR}})}, \tag{2.2}$$

where x^{TR} is the train position, M^{TOT} is the total mass of the train, $F^{\text{TB}}(v^{\text{TR}})$ is the traction or braking force, $R_{\text{l}}(x^{\text{TR}})$ is the line resistance at position x^{TR}, $R_{\text{b}}(v^{\text{TR}})$ is the basic resistance at speed v^{TR}, $v^{\text{TR}}(x^{\text{TR}})$ is the train speed at x^{TR}, $t(x^{\text{TR}})$ is the time at position x^{TR}, ξ_r is the rotating mass factor,.

$F^{\text{TB}}(v^{\text{TR}})$ is bounded by the maximum traction and braking forces. In reality, the maximum traction force is not a constant but a function of the speed v^{TR}. According to [98], nonlinear functions of a group of hyperbolic or parabolic formulas can be used to approximate the maximum traction force. For maximum braking force, due to safety concerns, it is only reserved for an emergency stop. Thus, the maximum braking force in this chapter is considered a constant that is much smaller than the actual maximum braking force.

The basic resistance $R_{\mathrm{b}}(v^{\mathrm{TR}})$ is a quadratic function of speed, where c_1, c_2, and c_3 are the coefficients of train characteristics. $R_{\mathrm{b}}(v^{\mathrm{TR}})$ can be described as [158]

$$R_{\mathrm{b}}(v^{\mathrm{TR}}) = c_1 + c_2 v^{\mathrm{TR}} + c_3\left(v^{\mathrm{TR}}\right)^2. \tag{2.3}$$

The line resistance $R_{\mathrm{l}}(x^{\mathrm{TR}})$ includes the resistance introduced by slopes, curves, and tunnels on the track, which is formulated as

$$R_{\mathrm{l}}(x^{\mathrm{TR}}) = M^{\mathrm{TOT}} g \sin(\alpha^{\mathrm{SA}}(x^{\mathrm{TR}})) + R_{\mathrm{c}}(\alpha^{\mathrm{R}}(x^{\mathrm{TR}})) + R_{\mathrm{t}}(x^{\mathrm{TU}}, v^{\mathrm{TR}}), \tag{2.4}$$

where g is the gravity acceleration, $\alpha^{\mathrm{SA}}(x^{\mathrm{TR}})$ is the slope angle at position x^{TR}, $R_{\mathrm{c}}(\alpha^{\mathrm{R}}(x^{\mathrm{TR}}))$ is the curve resistance when the radius of the curve at x^{TR} is α^{R}, $R_{\mathrm{t}}(x^{\mathrm{TU}}, v^{\mathrm{TR}})$ is the tunnel resistance, x^{TU} is the length of the tunnel.

The empirical curve resistance formula [158] is given as follows

$$R_{\mathrm{c}}(\alpha^{\mathrm{R}}(x^{\mathrm{TR}})) = \begin{cases} \dfrac{6.3M^{\mathrm{TOT}}}{\alpha^{\mathrm{R}}(x^{\mathrm{TR}}) - 55}, & \alpha^{\mathrm{R}}(x^{\mathrm{TR}}) \geq 300\mathrm{m}, \\ \dfrac{4.91M^{\mathrm{TOT}}}{\alpha^{\mathrm{R}}(x^{\mathrm{TR}}) - 30}, & \alpha^{\mathrm{R}}(x^{\mathrm{TR}}) < 300\mathrm{m}. \end{cases} \tag{2.5}$$

For tunnel resistance, if the tunnel exists a limiting gradient, namely, the maximum gradient that can be climbed without the help of a second power unit, $R_{\mathrm{t}}(x^{\mathrm{TU}}, v^{\mathrm{TR}})$ is calculated by [98, 158]

$$R_{\mathrm{t}}(x^{\mathrm{TU}}, v^{\mathrm{TR}}) = 1.296 \times 10^{-9} x^{\mathrm{TU}} M^{\mathrm{TOT}} g (v^{\mathrm{TR}})^2. \tag{2.6}$$

If the tunnel has no limiting gradient,

$$R_{\mathrm{t}}(x^{\mathrm{TU}}, v^{\mathrm{TR}}) = 1.3 \times 10^{-7} x^{\mathrm{TU}} M^{\mathrm{TOT}} g. \tag{2.7}$$

The train operation is subject to the following constraints,

$$a_{\mathrm{lim}}^{\mathrm{DCC}} \le a^{\mathrm{TR}}(x^{\mathrm{TR}}) \le a_{\mathrm{lim}}^{\mathrm{ACC}}, \forall x^{\mathrm{TR}} \in [0, x_{\mathrm{e}}], \tag{2.8}$$

$$v^{\mathrm{TR}}(x^{\mathrm{TR}}) \le v_{\mathrm{lim}}^{\mathrm{TR}}, \forall x^{\mathrm{TR}} \in [0, x_{\mathrm{e}}], \tag{2.9}$$

$$v^{\mathrm{TR}}(0) = 0, v^{\mathrm{TR}}(x_{\mathrm{e}}) = 0, t(0) = 0, | T^{\mathrm{PL}} - T^{\mathrm{ACT}} | \le T_{\mathrm{lim}}, \tag{2.10}$$

where $a^{\mathrm{TR}}(x^{\mathrm{TR}})$ is the acceleration at position x^{TR}, $a_{\mathrm{lim}}^{\mathrm{DCC}}$ and $a_{\mathrm{lim}}^{\mathrm{ACC}}$ are the deceleration and acceleration limits, respectively, $v_{\mathrm{lim}}^{\mathrm{TR}}$ is the on-road speed limit, x_{e} is the end position where the train stops, T_{lim} is the punctuality tolerance (maximum allowed trip time error), $| T^{\mathrm{PL}} - T^{\mathrm{ACT}} |$ is the absolute difference between the planned running time and the actual running time (trip time error), $T^{\mathrm{ACT}} = t(x_{\mathrm{e}})$.

We use $\mathcal{D}^{\mathrm{do}}$ and $\mathcal{D}^{\mathrm{bd}}$ to denote rescheduling commands or disturbances during operation and before departure, respectively. The uncertain trip time changes caused by $\mathcal{D}^{\mathrm{do}}$ and $\mathcal{D}^{\mathrm{bd}}$ are defined as $t_{\mathrm{d}}^{\mathrm{do}}$ and $t_{\mathrm{d}}^{\mathrm{bd}}$, respectively. Suppose the train receives notifications of the rescheduled command at position $x_{\mathrm{d}} \in [0, x_{\mathrm{e}})$,

$$T^{\mathrm{RTTR}} = \begin{cases} T^{\mathrm{PL}} - t_{\mathrm{d}}^{\mathrm{bd}}, & t_{\mathrm{d}}^{\mathrm{bd}} \sim \mathcal{T}^{\mathrm{bd}}, \quad \text{if } \mathcal{D}^{\mathrm{bd}}, \\ T^{\mathrm{PL}} - t_{\mathrm{d}}^{\mathrm{do}}, & t_{\mathrm{d}}^{\mathrm{do}} \sim \mathcal{T}^{\mathrm{do}}, \quad \text{if } \mathcal{D}^{\mathrm{do}}, \end{cases} \forall x \in [x_{\mathrm{d}}, x_{\mathrm{e}}], \tag{2.11}$$

$$| T^{\mathrm{RTTR}} - T^{\mathrm{ACT}} | = | T^{\mathrm{PL}} - (T^{\mathrm{ACT}} + t_{\mathrm{d}}^{\mathrm{bd}} \text{ or } T^{\mathrm{ACT}} + t_{\mathrm{d}}^{\mathrm{do}}) | \le T_{\mathrm{lim}}, \tag{2.12}$$

where $\mathcal{T}^{\mathrm{do}}$ and $\mathcal{T}^{\mathrm{bd}}$ are the time error distribution of $\mathcal{D}^{\mathrm{do}}$ and $\mathcal{D}^{\mathrm{bd}}$, respectively.

2.3 SRL–EETTO Approach

2.3.1 Approach Overview

To optimize the train trajectory online, an SRL–EETTO approach is proposed, which contains three steps: the model design, training, and verification steps (Fig. 2.3). At the model design step, the essential elements of the SRL environment are designed. First, the train operation features, including train operation states and constraints, are extracted, and multiple operation objectives (energy saving, ride comfort, punctuality, and safety) are formulated to establish evaluation indices for optimal train trajectory. Then, the SRL elements (state, action, and reward) are defined based on the train operation states and the evaluation indices.

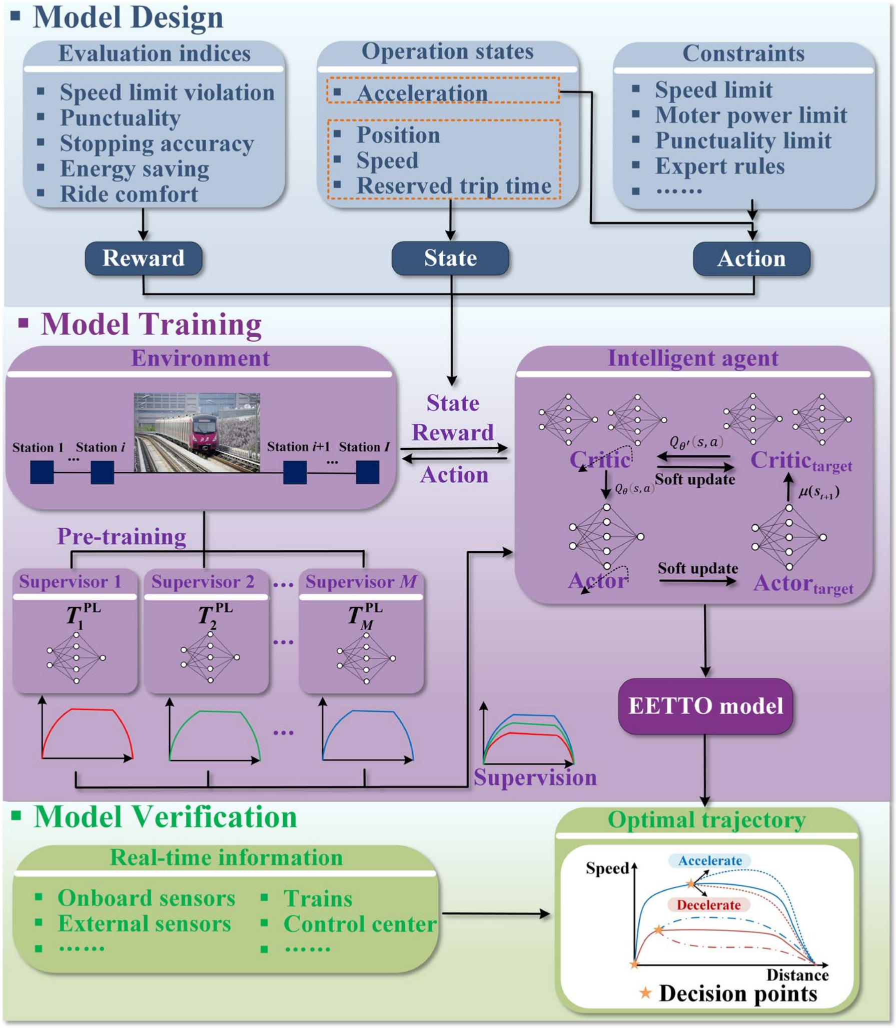


Fig. 2.3 Overview of SRL–EETTO.

At the model training step, the SAC architecture is adopted and a two-step training is implemented (Algorithm 2.1 and Algorithm 2.2). First, multiple EETTO models are pre-trained through the standard agent-environment interactions to serve as supervisors. These supervisors are pre-trained without using human driving data or ATO reference trajectories, which avoids the prior data collection process. Besides, each supervisor has a fixed but different planned running time T^{PL} within the practical running time range. This is to generate multiple optimal train trajectories on different planned trip times for agent learning to improve its generalization capability on disturbances or rescheduled trip times. Then, for the second step, an intelligent agent is trained under the supervision of supervisors.

At the model verification step, the well-trained agent, namely the EETTO model, is tested by various cases that simulate real-world situations to verify its model performance and illustrate its practical usage.

2.3.2 Model Design

2.3.2.1 Operation States, Constraints, & Evaluation Indices

According to the train operation states, such as the train position, speed, reserved trip time [159], and acceleration, the real-time train acceleration (deceleration) constraints b can be calculated using (2.8)–(2.12). Besides, to ensure safety and reduce the complexity of the problem, the following rules derived from expert knowledge of experienced drivers and ATOs are added: 1) A safe braking distance $\Delta x = -v_{\mathrm{lim}}^2 / \left(2a_{eb}^{\mathrm{DCC}}\right)$ is defined. Once the distance between the current train position x^{TR} and the next station is less or equal to Δx, the train must decelerate in a constant a^{TR} for emergency brakes. a_{eb}^{DCC} =-1 m/s^2 [100]. 2) $a^{\mathrm{TR}} = 0$ whenever the speed of the train reaches 95% of the speed limit.

The optimal train trajectory generated by the proposed approach should be evaluated in various aspects, including safety (on-road speed limit restriction and stopping accuracy), punctuality, energy saving, and ride comfort. Correspondingly, indices I_v, I_p, I_t, I_e, and I_c are formulated [79, 100]

$$I_v = \begin{cases} 1, & \text{if } v^{\mathrm{TR}} \geq v_{\mathrm{lim}}, \\ 0, & \text{Otherwise,} \end{cases} \tag{2.13}$$

$$I_p = | x_{\mathrm{e}} - x_{\mathrm{e}}^{\mathrm{ACT}} |, \tag{2.14}$$

$$I_t = | T^{\mathrm{RTTR}} - T^{\mathrm{ACT}} |, \tag{2.15}$$

$$I_e = \frac{1}{M^{\mathrm{TOT}} x_{\mathrm{e}}^{\mathrm{ACT}}} \int_0^{x_{\mathrm{e}}} F^{\mathrm{TB}}(v^{\mathrm{TR}}) \mathrm{d}x, \tag{2.16}$$

$$I_c = \int_0^{T^{\mathrm{RTTR}}} \begin{cases} | \frac{\mathrm{d}a^{\mathrm{TR}}}{\mathrm{d}t} | \mathrm{d}t, & | \frac{\mathrm{d}a^{\mathrm{TR}}}{\mathrm{d}t} | > 0.3 \text{ m/s}^3, \\ 0, & \text{Otherwise,} \end{cases} \tag{2.17}$$

where $x_{\mathrm{e}}^{\mathrm{ACT}}$ is the actual distance between stations.

The key elements of the MDP are formulated as follows to implement SRL.

2.3.2.2 State & Action

For the agent, its state s contains train position, speed, and reserved trip time. For supervisors, their state s^{s} only contains train position and speed. This is because they do not need to observe T^{PL} since it is fixed. The initial state is defined as $s_0 = [0, 0, T^{\mathrm{PL}}]$ and $s_0^{\mathrm{s}} = [0, 0]$. Action a is the acceleration of the train, where the action space is bounded by the acceleration and deceleration limits.

$$s_t = [x_t^{\mathrm{TR}}, v_t^{\mathrm{TR}}, T^{\mathrm{PL}} - t], \tag{2.18}$$

$$s_t^{\mathrm{s}} = [x_t^{\mathrm{TR}}, v_t^{\mathrm{TR}}]. \tag{2.19}$$

$$a_t = a_t^{\mathrm{TR}}. \tag{2.20}$$

2.3.2.3 Reward

Following the goal-directed reward design method based on the binary function [160], the rewards can be classified into goal-state rewards r_{g} (namely, for this paper, final-state rewards) and rewards per step r_∞

$$r = \begin{cases} r_{\mathrm{g}}, & s_{t+1} = s_{\mathrm{g}}, \\ r_\infty, & \text{Otherwise,} \end{cases} \tag{2.21}$$

where s_{g} is the goal state (final state).

Thus, the expected return $Q^*(s,a)$ becomes a constant Q_∞ when $r_{\mathrm{g}} = r_\infty$,

$$Q_\infty = \sum_{t=1}^{T} \gamma^{t-1} r_\infty = r_\infty \frac{1-\gamma^t}{1-\gamma} \approx \frac{r_\infty}{1-\gamma}. \tag{2.22}$$

If $r_{\mathrm{g}} > Q_\infty$, the final-state rewards are more attractive than rewards in other states, leading the agent to the final state. Besides, r_{g} and r_∞ must be non-negative to encourage the agent to move from the current state to the next state. Thus, the relationship between r_{g} and r_∞ is established as

$$r_{\mathrm{g}} > \frac{r_\infty}{1-\gamma} \geq 0. \tag{2.23}$$

Thus, according to (2.23), we design different types of rewards following the binary reward function form to reflect various real-world objectives. Table 2.1 illustrates the designed rewards.

Table 2.1 Rewards.

Item	r_{g}	r_∞
-	+350	+2.5
Speed limit	-	-I_v
Punctuality	$\begin{cases} r_{\mathrm{T}} I_t, & I_t \geq T_{\mathrm{lim}}, \\ +50, & \text{Otherwise}, \end{cases}$	-0.5, if $I_t \geq T_{\mathrm{lim}}$
Stopping accuracy	-I_p, if $I_p \geq P_{\mathrm{lim}}$	-
Energy saving	$r_{\mathrm{E}} I_e$	-0.5, if $I_e \geq E_{\mathrm{lim}}$
Ride comfort	$r_{\mathrm{C}} I_c / I_{c,\max}$	-0.5, if $I_c \geq C_{\mathrm{lim}}$

For r_∞, the ride comfort, energy, on-road speed limits, and punctuality are considered. These rewards in r_∞ can give immediate feedback at every step to accelerate training. Specifically, 1) For speed limits, on the one hand, the agent should follow all speed limits; on the other hand, the agent should not stop before its arrival;

the agent gets a negative reward for every time step it breaks the above rules; since safety is the basic operation requirement and most important objective, the penalty weights are higher than other objectives. 2) For punctuality, the agent gets a negative reward for every time step its punctuality performance is worse than the tolerance T_{lim}; T_{lim} is set to be 3 s [100]. 3) For energy saving, the agent gets a negative reward for every time step its energy-saving performance is worse than the tolerance E_{lim}; E_{lim} is set to be equal to the practical energy consumption of the same line since we expect better energy saving in agent performance than in practice. 4) For ride comfort, the agent gets a negative reward for every time step its ride comfort performance is worse than the tolerance C_{lim}; C_{lim} is set to be 0.3 g/s [161]. 5) A bias term is added to ensure the non-negative nature of r_∞.

For r_{g}, we design rewards for ride comfort, energy, punctuality, and stopping accuracy. The stopping accuracy tolerance p_{lim} is set to be 0.3 m [162]. The upper bound is zero for all r_{g} coefficients since we aim to minimize these reward terms. Derived from (2.23), the lower bound is derived from the following equation that all coefficients must satisfy

$$r_{\text{g,max}} + r_{\text{T}} I_t + r_{\text{E}} I_e + \frac{r_{\text{C}} I_c}{I_{c,\text{max}}} - I_p = r_{\text{g,max}} > r_\infty / (1-\gamma). \tag{2.24}$$

I_p can be ignored if the maximum time step T is sufficiently large that the train position differences between each step is small. Therefore, (2.24) can be rewritten as

$$\begin{aligned} \max\quad & r_{\text{T}}, r_{\text{E}}, r_{\text{C}} \\ \text{s.t.}\quad & (2.24), \\ & 0 \le I_c \le I_{c,\text{max}}, T_{\text{lim}} \le I_t \le T^{\text{PL}}, 0 \le I_e \le E_{\text{lim}}, \\ & r_{\text{T}} \le 0, r_{\text{E}} \le 0, r_{\text{C}} \le 0, \end{aligned} \tag{2.25}$$

where $I_{c,\text{max}}$ is a sufficiently large value to consider the worst ride comfort case ($|\frac{\text{d}a^{\text{TR}}}{\text{d}t}| > 0.3\ \text{m/s}^3$). Since these coefficients of various r_{g} terms have a significant impact on model performance, a fine-tuning process is carried out to optimize their values. The results are illustrated in subsection 2.4.2.

2.3.3 Model Training and Verification

2.3.3.1 Training process

The model training architecture is shown in Fig. 2.4, where a two-step training process is implemented. At the pre-training step, each supervisor is trained with a fixed but different T^{PL} within the practical running time range. This range is determined by calculating the minimum and maximum planned trip time $T^{\mathrm{PL,min}}$ and $T^{\mathrm{PL,max}}$ of the trip. The calculation of $T^{\mathrm{PL,min}}$ and $T^{\mathrm{PL,max}}$ can be referred to [163], where $T^{\mathrm{PL,max}}$ is based on the assumption that the lowest average running speed of 40 km/h offered to passengers. T_1^{PL} , T_2^{PL} , …, T_m^{PL} for supervisor 1 , 2 , …, M are uniformly sampled from $T^{\mathrm{PL,min}}$ to $T^{\mathrm{PL,max}}$. The TD3 algorithm with prioritized experience replay [44] is used to train supervisors.

At the agent training step, an improved TD3 algorithm, S-TD3, which is suitable for training the agent under the SAC architecture, is proposed. Similar to TD3, the actor network outputs action a based on its policy μ, which is updated according to the Q value, and two critic networks are adopted to estimate the Q value. However, in S-TD3, multiple supervisor networks are added to calculate the supervision loss, namely, the differences between the supervisors policy $\mu_{\mathrm{s1}}(s)$ and the agents policy $\mu(s)$, to guide the actors action. Besides, different from TD3, the SRL training environment randomly assigns T^{PL} value for each episode, and the training data are stored separately in several independent buffers according to T^{PL} . Each supervisor samples data from its own buffer while the actor receives data from all buffers. In this manner, the supervisor avoids providing inappropriate supervisory information, and the actor can receive all supervisory information. The detailed training procedures are shown in Algorithm 2.1 and Algorithm 2.2.

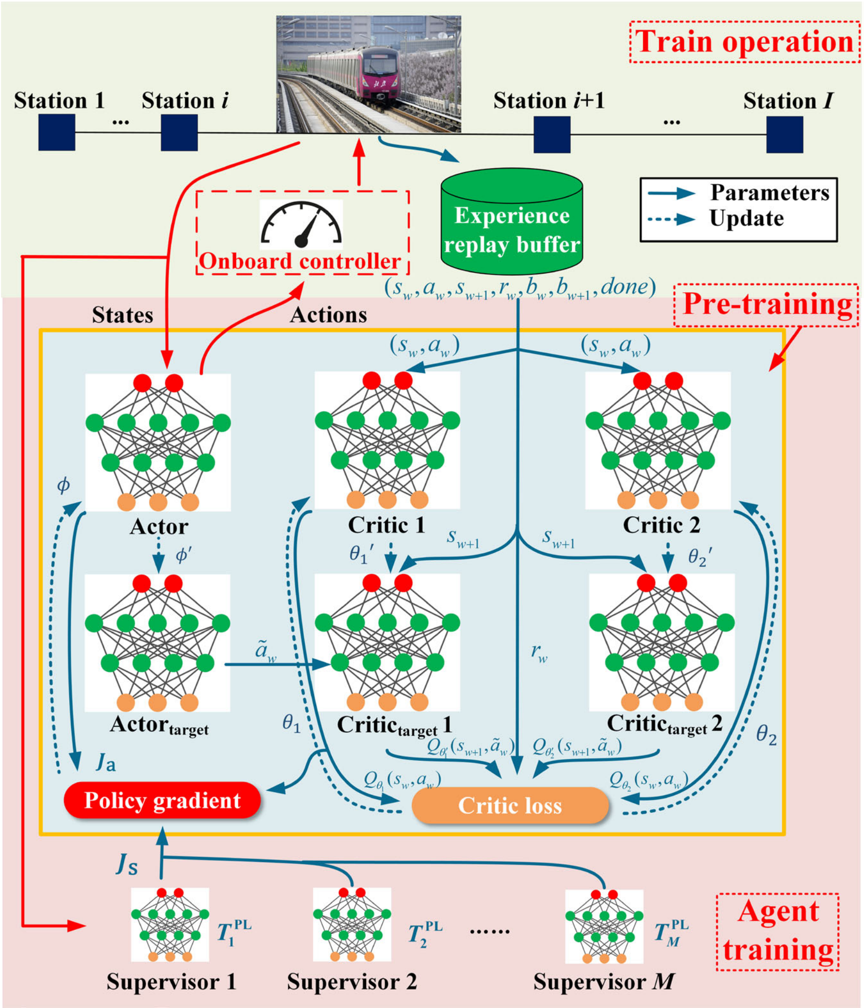


Fig. 2.4 SAC training architecture.

2.3.3.2 Loss Update of Actor Without Supervisor

To find the optimal agent's policy μ^*, the loss function of the actor can be updated by taking the gradient of the expected return

$$\nabla_\phi J_{\mathrm{a}}(\phi) = \mathbb{E}_{s_w \sim \mathcal{S}_\mu}\left[\nabla_a Q_\theta(s_w, a_w)|_{a_w=\mu_\phi(s)} \nabla_\phi \mu_\phi(s_w)\right], \tag{2.26}$$

$$\phi \leftarrow \phi + \xi_a \nabla_\phi J_a(\phi), \tag{2.27}$$

where J_a is the loss of the actor. θ is the parameter of critic, ξ_a is the learning rate of the actor. It is worth noting that the policy update of the actor is delayed by a rate *DP* to let the critic have a better estimation of Q.

2.3.3.3 Loss Update of Actor with Supervisor

With the supervisor, the loss update of the actor is modified to include the supervision loss by

$$\nabla_\phi J'_a(\phi) = (1-\alpha_\mathrm{s})\nabla_\phi J_\mathrm{a}(\phi) + \alpha_\mathrm{s}\nabla_\phi J_\mathrm{s}(\phi), \tag{2.28}$$

$$J_\mathrm{s}(\phi) = \sum_m \mathbb{E}_{s_w \sim \mathcal{S}_\mu}\left[(\mu_{\mathrm{sl},m}(s_w) - \mu_{\phi,m}(s_w))^2\right], \tag{2.29}$$

$$\phi \leftarrow \phi + \xi_a \nabla_\phi J'_a(\phi), \tag{2.30}$$

where J'_a is the revised loss of the actor, J_s is the supervision loss, $\alpha_\mathrm{s} \in [0,1]$ represents the trade-off between RL and SL contribution.

2.3.3.4 Loss Update of Critic

The update of the critic is by minimizing the critic loss (Q loss)

$$J_\mathrm{c}(\theta) = \mathbb{E}_{r_w \sim \mathcal{R}, s_w \sim \mathcal{S}_\mu}[(Q_\theta(s_w, a_w) - y)^2], \tag{2.31}$$

$$\theta \leftarrow \theta + \xi_c \nabla_\theta J_c(\theta), \tag{2.32}$$

where J_c is the critic loss, y is calculated by the target critic. ξ_c is the learning rate of the critic.

2.3.3.5 Target Network

In S-TD3, there are two target critic networks and one target actor network. For the target critics, by take the minimum Q value of both networks, the Q value overestimation issue can be mitigated.

$$y = r(s_w, a_w) + \gamma \min Q_{\theta'_{1,2}}(s_{w+1}, \tilde{a}_w)\,|_{\tilde{a}_w = (\mu_{\phi'}(s_{w+1}) + \mathrm{clip}(\mathcal{N}))}, \tag{2.33}$$

where θ' is the parameter of the target critic, $\tilde{a}$ is the action taken by the target actor, $\mu_{\phi'}$ is the policy of the target actor. A target policy smoothing is implemented by adding a small stochastic noise to the target actor to mitigate overfitting.

Algorithm 2.1 Pre-Training

```
Initialize actor $\mu_\phi$ and critic $Q_{\theta_1}$, $Q_{\theta_2}$ with random weights $\phi$, $\theta_1$, and $\theta_2$. Initialize target networks $\phi'$, $\theta_1'$, and $\theta_2'$ with weights $\phi' \leftarrow \phi$, $\theta_1' \leftarrow \theta_1$, and $\theta_2' \leftarrow \theta_2$. Initialize replay buffer $B$.
For episode = 1, Max do
    Receive the initial observation $s_0^{\mathrm{s}}$
    For $t$ = 1, $T$ do
        Select $a_t \sim \mu_\phi(s_t) + \mathcal{N}$, $\mathrm{clip}(a_t, -b_t, b_t)$, execute $a_t$ and observe $r_t$, $s_{t+1}$, $b_{t+1}$
        Store transition $(s_t, a_t, s_{t+1}, r_t, b_t, b_{t+1}, done)$ to $B$
        Sample $W$ random transitions from $B$
        Select $\tilde{a}_w \sim \mu_{\phi'}(s_{w+1}) + \mathcal{N}$, $\mathrm{clip}(\tilde{a}_w, -b_{w+1}, b_{w+1})$
        $\theta_{1,2} \leftarrow \arg\min_{\theta_{1,2}} W^{-1} \sum_w (y - Q_{\theta_{1,2}}(s_w, a_w))^2$, update $\theta_{1,2}$ by (2.32)
        If $t$ mod $DP$ then
            Update $\phi$ by (2.26)-(2.27)
            $\theta_1' \leftarrow \tau\theta_1 + (1-\tau)\theta_1', \theta_2' \leftarrow \tau\theta_2 + (1-\tau)\theta_2'$, $\phi' \leftarrow \tau\phi + (1-\tau)\phi'$
```

Target networks are updated at regular intervals τ to enable more stable learning (soft update), namely,

$$\theta_1' \leftarrow \tau\theta_1 + (1-\tau)\theta_1', \theta_2' \leftarrow \tau\theta_2 + (1-\tau)\theta_2', \quad \phi' \leftarrow \tau\phi + (1-\tau)\phi' . \tag{2.34}$$

2.3.3.6 Verification process

For practical application purposes, the proposed EETTO model can be deployed on the onboard ATO system. Prior to the real-time operation, the train and line data were loaded into the ATO system. Then, the proposed model generated multiple referential train trajectories according to the running time ranges and train parameters. In real-time operation, with the received information from onboard and external sensors, other trains, and control centers (the process can be referred to Fig. 2.1), the proposed model can dynamically adjust the referential train trajectories online to address uncertain disturbances and rescheduled trip times.

Algorithm 2.2 S-TD3

```
Initialize actor $\mu_{\phi}$ and critic $Q_{\theta_1}$, $Q_{\theta_2}$, target networks $\phi'$, $\theta_1'$, and $\theta_2'$, and replay buffers $B_1$, $B_2$, …, $B_M$. Load $\mu_{\mathrm{sl},1}$, $\mu_{\mathrm{sl},2}$, …, $\mu_{\mathrm{sl},M}$
For episode = 1, Max do
    Receive the initial observation $s_0$ and ${s_0}^{\mathrm{s}}$
    For $t$ = 1, $T$ do
        Select $a_t \sim \mu_{\phi}(s_t) + \mathcal{N}$, $\mathrm{clip}(a_t, -b_t, b_t)$, execute $a_t$ and observe $r_t$, $s_{t+1}$, $b_{t+1}$
        Store transition $(s_t, a_t, s_{t+1}, r_t, b_t, b_{t+1}, done)$ to $B_1$, $B_2$, …, $B_M$
        Sample $W$ random transitions from $B_1$, $B_2$, …, $B_M$ equally
        Select $\tilde{a}_w \sim \mu_{\phi'}(s_{w+1}) + \mathcal{N}$, $\mathrm{clip}(\tilde{a}_w, -b_{w+1}, b_{w+1})$
        $\theta_{1,2} \leftarrow \arg\min_{\theta_{1,2}} W^{-1} \sum_{w} (y - Q_{\theta_{1,2}}(s_w, a_w))^2$, update $\theta_{1,2}$ by (2.32)
        If $t$ mod $DP$ then
            Update $\phi$ by (2.26) and (2.28)-(2.30)
            $\theta_1' \leftarrow \tau\theta_1 + (1-\tau)\theta_1'$, $\theta_2' \leftarrow \tau\theta_2 + (1-\tau)\theta_2'$, $\phi' \leftarrow \tau\phi + (1-\tau)\phi'$
```

\

2.4 Case Study

In this section, the numerical analysis of the aforementioned formulations and algorithms is conducted. First, the optimal selection of reward coefficients is demonstrated based on the proposed evaluation indices. In addition, the performance of SRL–EETTO is investigated by comparing it with state-of-the-art EETTO algorithms under various operation scenarios, including normal operations, uncertain disturbances, uncertain train masses, and uncertain resistances. Finally, the impact of supervision weights on the generalization capability is investigated.

Table 2.2 Speed limits and gradients of the training section.

Item	Segment (km)	Value	Segment (km)	Value
Speed limits (km/h)	[0, 0.31]	50	(0.64, 1.32]	65
	(0.31, 0.64]	80	(1.32, 2.63]	80
Gradients (‰)	[0, 0.02]	0	(1.15, 1.55]	-3
	(0.02, 0.34]	2	(1.55, 2.06]	8
	(0.34, 0.65]	3	(2.06, 2.63]	-3
	(0.65, 1.15]	-10.4	-	-

Table 2.3 Train timetable.

Station	Arrival time (s)	Dwell time (s)	Mileage (m)	Station	Arrival time (s)	Dwell time (s)	Mileage (m)
SJZ	0	30	0	RJ	1112	30	12065
XC	220	30	2631	RC	1246	30	13419
XHM	358	30	3906	TJN	1440	30	15757
JG	545	30	6272	JH	1620	30	18022
YZQ	710	30	8254	CQN	1790	35	20108
WHY	835	30	9274	CQ	1927	45	21394
WY	979	30	10785	YZ	2087	-	22728

2.4.1 Setup

In this subsection, the setup of the case study is illustrated. The simulation data of the infrastructure, train, and line conditions are from an in-service subway line in Beijing containing 13 sections (14 stations) and a total length of 22.73 km. However, the curve and tunnel data are not available. Considering that these data can vary greatly due to weather conditions, we ignore the curve and tunnel resistance terms in (2.4) but verify the model performance under resistance uncertainty in the following subsection instead. The model is trained on section SJZ–XC, where the section length is 2.63 km. Then, the evaluation is conducted on the entire subway line. The gradient profile and speed limits of the line can be found in [164] and Table 2.2. The train timetable is shown in Table 2.3. The train parameters are shown in Table 2.4. We set E_{lim}=162 J/(km · kg) based on practical and simulation data [139] of the same training section.

Table 2.4 Train parameters.

Parameter	Value
Maximum traction force (kN)	$\begin{cases} 310, & v^{\text{TR}} \le 36 \text{ km/h} \\ 310 - 20 \times (v^{\text{TR}} - 36), & 36 < v^{\text{TR}} \le 80 \text{ km/h} \end{cases}$
Basic resistance force (kN)	$3.48 + 0.144 v^{\text{TR}} + 0.0085\left(v^{\text{TR}}\right)^2$
Train mass (kg)	2×10^5

Table 2.5 SRL–EETTO parameters.

Parameter	Value	Parameter	Value
ξ_c	10^{-4}	M	5
ξ_a	10^{-5}	α_s	0.99
τ	5×10^{-4}	Optimizer	Adam
W	128	Buffer capacity	2^{20}
γ	0.99	Exploration policy	$\mathcal{N}(0,0.2)$, clip to $(-0.5,0.5)$

The parameters of the SRL–EETTO are listed in Table 2.5, which are suitable for both supervisor pre-training and agent training steps. For the supervisor, both actor and critic have 256, 256, 128, 64, and 64 units for hidden layers. For the agent, both the actor and critic have 400, 300, 200, 100, and 64 units for hidden layers. Each of the hidden layers is followed by Relu non-linearity. The output layer of the actor is followed by a Tanh non-linearity, while the output layer of the critic does not have any activation function. The target networks have the same structure as the corresponding actor or critic. The inputs are normalized for all networks. γ is set as 0.99 to fully account for the consideration of S-TD3 for future rewards. Besides, in order to make a trade-off between exploration and exploitation, the exploration noise is subject to a normal distribution $\mathcal{N}(0,0.2)$ and then clipped to the range (-0.5,0.5). The learning rate for the actor and the critic is 10^{-5} and 10^{-4}, respectively. The fine-tuning is carried out on Gurobi 9.5.2, and the SRL training is on Python 3.9.13 with PyTorch 1.12.1. The PC used for the computation has an Intel Core i7-12700KF processor at 3.61 GHz with 32 GB memory and an RTX3070.

Table 2.6 Model performance with different coefficient values[1].

r_{T}	r_{C}	r_{E}	I_t (s)	I_c (m/s^3)	I_e (J/(km·kg))	I_v
0.4	100	0.6	3.5±0.9	12.0±14.9	96.7±34.9	0.3±0.5
0.4	100	0.4	4.2±4.0	2.9±2.2	82.9±10.9	0±0
0.4	100	0.2	5.4±5.7	10.9±10.6	96.7±21.2	0.3±0.5
0.4	**0**	**0.6**	**2.1±0.6**	**1.7±0.4**	**82.5±10.2**	**0±0**
0.4	0	0.4	7.5±3.3	2.2±0.7	80.0±2.6	0±0
0.4	0	0.2	4.1±1.8	4.9±4.8	84.2±13.0	0±0
0.2	100	0.6	4.7±2.6	3.9±3.0	83.3±6.4	0±0
0.2	100	0.4	2.8±1.1	7.9±4.3	85.2±16.0	0±0
0.2	100	0.2	4.9±2.6	13.8±17.6	93.8±31.5	0.3±0.5
0.2	0	0.6	2.6±0.2	4.2±3.4	83.9±16.5	0±0
0.2	0	0.4	3.0±0.6	10.9±13.8	90.5±24.7	0±0
0.2	0	0.2	2.3±0.2	2.7±0.5	83.8±9.5	0±0

[1] ± denotes a single standard deviation, I_p is always zero.

2.4.2 Optimal Selection of Reward Coefficients

In this subsection, the optimal selection results of the reward coefficients of r_{g} is illustrated. By fixing r_{E} and r_{C} to their maximum, the lower bound of r_{T} can be calculated based on (2.25). Similarly, the lower bounds of r_{E} and r_{C} are obtained. $-0.62 \le r_{\mathrm{E}} \le 0$, $-0.43 \le r_{\mathrm{T}} \le 0$, and $-100 \le r_{\mathrm{C}} \le 0$. Table 2.6 shows the model performance under different coefficient values averaged by random 3 runs. The best parameters are $r_{\mathrm{E}} = 0.6$, $r_{\mathrm{T}} = 0.4$, $r_{\mathrm{C}} = 0$.

2.4.3 Analysis of Model Performance

2.4.3.1 Model Performance Under Normal Operations

In this subsection, the model performance of the proposed approach under normal operations is verified. The following approaches are compared to illustrate its overall model performance without disturbances or rescheduled trip times: *1) Manual driving (MD)*. *2) ATO-generated trajectories* with proportional-integral-derivative controller. 1)–2) are the practical driving data with no departure delays of the line in March 2012. Half of the data are MD, and the others are ATO since both types of driving were used at that time. *3) RTO algorithm* [32]: a comprehensive knowledge-based system with a collection of expert knowledge rules. The selection of experts requires prior data collection, surveying, expert selection, data mining, and summarizing. Noted that RTO is unable to handle disturbances. *4) STO algorithm* [100]: STO utilized advanced RL algorithms such as deep deterministic policy gradient and normalized advantage function to handle continuous action space for solving the TTO. *5) proposed approach*.

Fig. 2.5 shows the optimal trajectories generated by SRL–EETTO on the training set. Five pre-trained supervisors are used with planned running times of 185 s, 197 s, 209 s, 221 s, and 234 s, respectively ($T^{\mathrm{PL,min}}$=183 s and $T^{\mathrm{PL,max}}$=234 s). The trajectories are smooth and have no violations of on-road speed limits. Fig. 2.6 shows the overall model performance comparison under normal operation. From the figure, the upper part shows the optimal trajectories of SRL–EETTO, while the bottom part shows the comparative results of index I_e and I_c in each section. The bars represent the results of I_e, and the circles with the dotted line and the light-shaded area represent the results of I_c. Table 2.7 summarizes the average model performance across all sections, where the performance of SRL–EETTO is averaged across 5 runs, where P_{E} = (avg. I_e of MD – avg. I_e of the approach) / avg. I_e of MD × 100%; P_{C} = (min. I_c of MD – avg. I_c of the approach) / min. I_c of MD × 100%.

Table 2.7 Comparative model performances of approaches 1–5[1].

Item	I_c (m/s^3)	P_{C} (%)	I_e (J/(km·kg))	P_{E} (%)	I_t (s)
MD	7.5–14.0	-	147.0±31.0	-	2.5±2.4
ATO	-	-	154.7±31.0	-5.2	1.7±1.6
RTO	-	-	120.1±20.9	18.3	2.0±1.0
STO	4.0–5.8	-	-	-	-
SRL–EETTO	3.4±0.9	54.7	119.8±23.8	18.5	2.2±1.0

[1] ± denotes a single standard deviation, I_v and I_p are always zero.

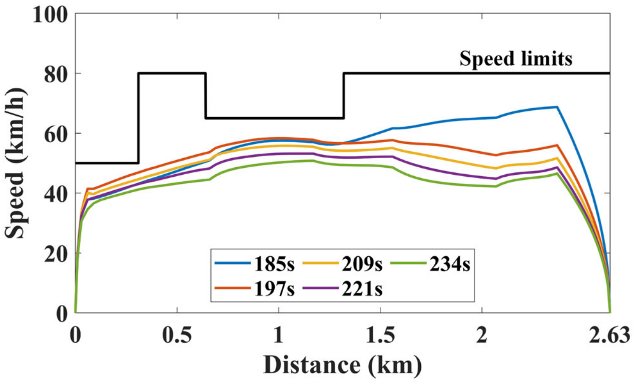


Fig. 2.5 Optimal trajectories generated on the training set.

It can be observed that SRL–EETTO can satisfy on-road speed limits and stopping accuracy in all sections. Besides, SRL–EETTO achieves the best performance on I_e among existing approaches and outperforms MD in an average energy saving of 18.5%. Although RTO achieves similar performance on I_e as SRL–EETTO, it is unable to handle disturbances. In addition, SRL–EETTO achieves the best performance on I_c compared with the practical solution and outperforms the practical solution in an average energy saving of 54.7%. Moreover, on the one hand, the variance of the trip time error of MD is huge, indicating that MD has unsatisfactory punctuality performance in some sections. On the other hand, although higher than ATO, the I_t of SRL–EETTO is still within 3 s.

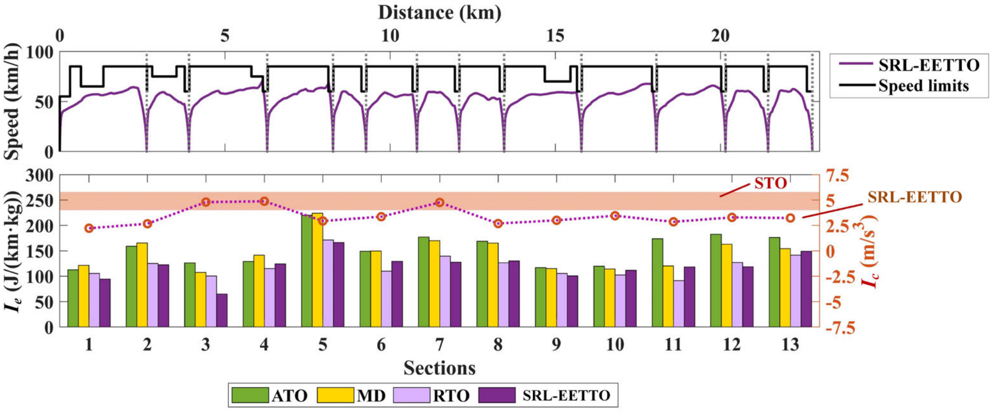
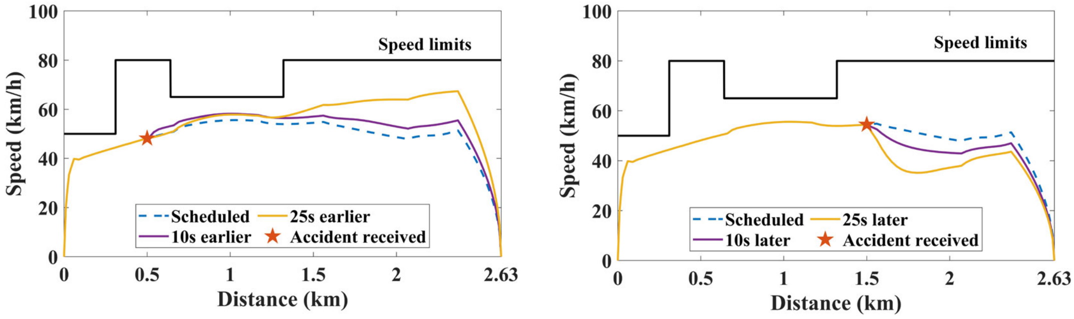


Fig. 2.6 Optimal trajectories and model performances across all sections.

2.4.3.2 Model Performance Under Uncertain Disturbances

In this subsection, the model performance of the proposed approach under uncertain disturbances, namely, the dynamic online train trajectory optimization capability under disturbances and rescheduled running times, is verified. First, we use Fig. 2.7 as an example to illustrate the model performance of the proposed model with disturbances and rescheduled trip times.



(a) (b)

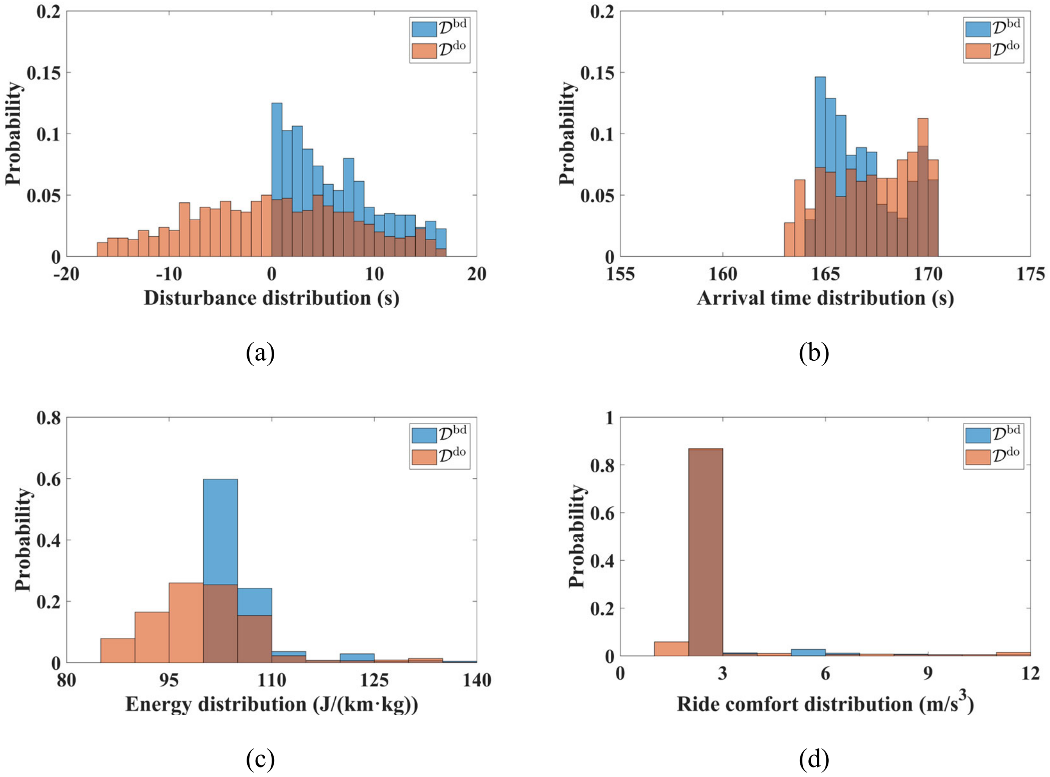
Fig. 2.7 Optimal trajectories with adjusted running times for scenarios (a) 1 and (b) 2.

Table 2.8 Model performances with adjusted running times[1].

Item	Adjustment	I_c (m/s^3)	I_e (J/(km·kg))	I_t (s)
Scheduled	-	2.6	73.3	1.1
Scenario 1	10 s earlier	3.0	80.6	1.1
	25 s earlier	2.5	102.0	2.3
Scenario 2	10 s later	2.4	71.0	1.0
	25 s later	3.0	77.8	1.0

[1] I_v and I_p are always zero.

Suppose the planned running time is scheduled as 210 s for the training section. Fig. 2.7(a) shows an accident occurs when the train runs 500 m. The train is informed at this moment to arrive at the station 10 s / 25 s earlier, respectively. This indicates that T^{PL} is changed to 200 s / 185 s, respectively. A red star marker represents the position where the accident happened. Fig. 2.7(b) are similar, except that the accident happens when the train runs 1500 m, and the train is required to arrive 10 s / 25 s later, respectively. It can be observed that when the train receives the accident information, the proposed model will change the driving strategy (action a) since the model input (state s) is changed due to the change of reserved trip time. Table 2.8 shows the detailed model performance. The trip time error is always within 3 s. I_e is larger than the example with scheduled T^{PL}, indicating extra energy consumption due to acceleration. I_c is larger than the example with scheduled T^{PL}, indicating slightly uncomfortable passengers may feel due to acceleration or deceleration.



(a)

(b)

(c)

(d)

Fig. 2.8 Probability distributions of (a) disturbances, (b) arrival times, (c) energies, and (d) ride comforts under Monte Carlo simulations.

To test the overall model performance under disturbances / rescheduled trip times, we then perform 2000 times of Monte Carlo simulations. Section TJN-JH is between two busy stations and is suitable for demonstrating the test results. $T^{\mathrm{PL,min}}$ and $T^{\mathrm{PL,max}}$ of section TJN–JH are 150 s and 185 s, respectively. The distributions of trip time changes are referred to [32, 165]. $t_{\mathrm{d}}^{\mathrm{bd}}$ is subject to a Weibull distribution where the shape parameter is 0.8, and the scale parameter is $(T^{\mathrm{PL,max}} - T^{\mathrm{PL,min}})/2$. $t_{\mathrm{d}}^{\mathrm{bd}}$ is non-negative since disturbances or rescheduling commands before departure usually cause delays. $t_{\mathrm{d}}^{\mathrm{do}}$ is subject to a Normal distribution where the mean value is 0, and the variance value is $(T^{\mathrm{PL,max}} - T^{\mathrm{PL,min}})/4$. For simulation purposes, x_{d} is set to occur within the first half of the trip. This is because when the train is close to the destination, it is difficult or even impossible to significantly change trip time by train control.

Table 2.9 Model performances across Monte Carlo results[1].

Disturbance type	I_c (m/s^3)	I_e (J/(km·kg))	I_t (s)
-	3.0	100.7	1.6
$\mathcal{D}^{\mathrm{bd}}$	3.2±3.1	109.2±13.6	1.6±0.9
$\mathcal{D}^{\mathrm{do}}$	2.7±1.9	101.5±11.7	1.9±1.0

[1] ± denotes a single standard deviation, I_v and I_p are always zero.

Table 2.10 Comparative model performances with RL-based algorithms.

Item	Maximum I_t (s)	Disturbance (s)
ITO	5.0	≤ 20
ITOR	4.3	10
SRL–EETTO	3.8 ($\mathcal{D}^{\mathrm{do}}$) 3.3 ($\mathcal{D}^{\mathrm{bd}}$)	≤ 35

Fig. 2.8 shows the Monte Carlo results in histograms. The blue and red colors of the histograms represent simulations that are subject to $\mathcal{D}^{\mathrm{bd}}$ and $\mathcal{D}^{\mathrm{do}}$, respectively. Fig. 2.8(a) shows the distribution of $t_{\mathrm{d}}^{\mathrm{bd}}$ and $t_{\mathrm{d}}^{\mathrm{do}}$, which denotes the distribution of disturbances. Fig. 2.8(b) shows the distribution of $(T^{\mathrm{ACT}} + t_{\mathrm{d}}^{\mathrm{bd}} \text{ or } T^{\mathrm{ACT}} + t_{\mathrm{d}}^{\mathrm{do}})$ and denotes the arrival time. Fig. 2.8(c) and Fig. 2.8(d) show the distribution of energy and ride comfort. The average model performance of the Monte Carlo simulations is reported (Table 2.9). It can be observed that disturbances vary on a broad time distribution (0–17.5 s for $\mathcal{D}^{\mathrm{bd}}$, -17.5–17.5 s for $\mathcal{D}^{\mathrm{do}}$), but the arrival time distribution is concentrated around the planned trip time (around 163–170 s). This indicates that the

probability of delay is very small across all simulations (0–3.3 s for $\mathcal{D}^{\mathrm{bd}}$, 0–3.8 s for $\mathcal{D}^{\mathrm{do}}$) and the average trip time errors are within 2 s under both $\mathcal{D}^{\mathrm{do}}$ and $\mathcal{D}^{\mathrm{bd}}$. The punctuality against $\mathcal{D}^{\mathrm{do}}$ is worse than against $\mathcal{D}^{\mathrm{bd}}$. This indicates the additional trip time error caused by trajectory changes during operation. The energy distribution varies due to the extra energy consumption for acceleration and deceleration to guarantee punctuality. Most of the resulting energy consumptions are concentrated within a small range (around 85–110 (J/(km$\cdot$kg)) for $\mathcal{D}^{\mathrm{bd}}$, and 100–110 (J/(km$\cdot$kg)) for $\mathcal{D}^{\mathrm{do}}$) with the average energy consumption close to results without disturbances. The ride comfort distribution is similar to energy distribution, except that it is more concentrated.

The Monte Carlo simulation shows that SRL–EETTO can efficiently overcome disturbances before departure and during operation and maintain model performance in terms of punctuality, energy saving, and ride comfort via online timetable adjustment. We then compare SRL–EETTO with state-of-the-art RL-based EETTO algorithms reported in the literature that consider disturbances (Table 2.10). For comparison purposes, we chose the ITO and ITOR algorithms based on Q learning. It can be observed that SRL–EETTO reduces maximum trip time error by at least 24.0% and 11.6% against a broader disturbance range compared with ITO and ITOR, respectively. Moreover, the computational time to re-generate the optimal train trajectory after disturbances is about 0.07 s. This fast response time indicates that the SRL–EETTO can generate or regenerate the optimal train trajectory online.

In addition, according to the above performance of SRL–EETTO, a trajectory configuration suggestion can be given. Generally, for convenience of scheduling, the equipped train trajectories of the ATO are of equal time separation. Considering the average trip time error in Table 2.9, this separation can be 3 s for SRL–EETTO. As the running time range of sections is around 33–51 s, the number of equipped trajectories can be increased from 5 (current implementation) to 12–18 (SRL–EETTO).

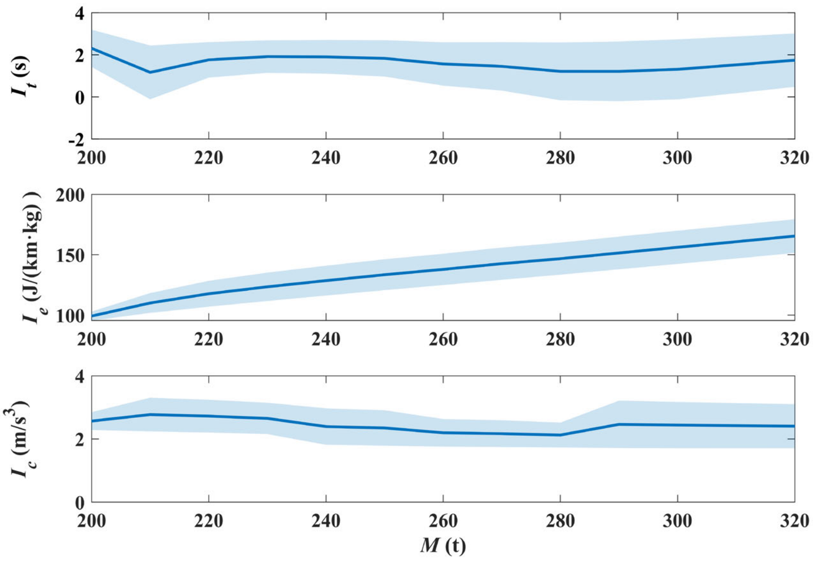


Fig. 2.9 Model performances with different train masses.

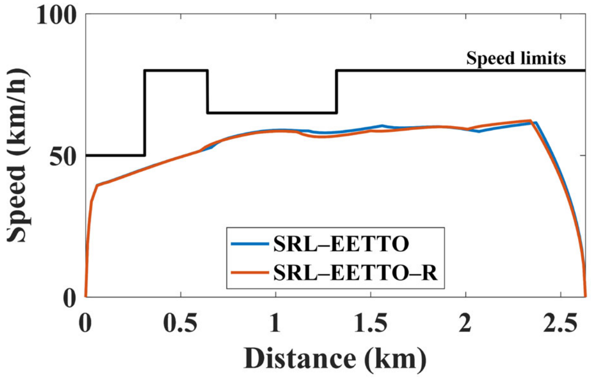


Fig. 2.10 Optimal trajectories under different resistances.

2.4.3.3 Model Performance Under Uncertain Train Masses

In this subsection, the model performance of the proposed approach under uncertain train masses is verified. Since the maximum train capacity during peak hours may reach 2000, assuming the average passenger weight is 60 kg, the maximum train mass can reach 320 t. Fig. 2.9 shows the trip time error, energy saving, and ride comfort of the SRL–EETTO under different train mass conditions within the possible train mass range. The shaded area denotes a single standard deviation from the average value across 5 runs. From the figure, it can be observed that punctuality is achieved across all train mass conditions, and the maximum trip time error (2.9 s) is at the maximum train mass point. As for energy saving, a larger train mass naturally leads to higher energy consumption. The index I_e increases almost linearly. The ride comfort, however, is improved with the increasing train mass. To summarize, the overall performance does not deteriorate when the train mass is changed.

Table 2.11 Changed gradients of section SJZ–XC.

Segment (km)	Value (‰)	Segment (km)	Value (‰)
[0, 0.57]	-3	(1.98, 2.29]	3
(0.57, 1.08]	8	(2.29, 2.61]	2
(1.08, 1.48]	-3	(2.61, 2.63]	0
(1.48, 1.98]	-10.4	-	-

Table 2.12 Comparative model performances with different resistances.

Item	I_c (m/s^3)	I_e (J/(km·kg))	I_t (s)
MD	7.5–14.0	121.4	8.8
ATO	-	112.6	0.6
SRL‒EETTO	2.6	99.2	2.3
SRL‒EETTO-R	2.7	105.2	2.9

2.4.3.4 Model Performance Under Uncertain Resistances

In this subsection, the model performance of the proposed approach under uncertain resistance conditions is analyzed. We simulate the resistance changes by reversing the on-road gradient conditions in Table 2.2, which is shown in Table 2.11. The trajectory and model performance on different resistances are reported in Fig. 2.10

and Table 2.12, respectively, where SRL–EETTO-R denotes the model performance under changed resistances. The results show increased energy saving and ride comfort of SRL–EETTO-R compared to MD and ATO. Although indices I_e and I_t of SRL–EETTO-R are higher than SRL–EETTO due to the resistance change, the trip time error is within 3 s, which satisfies the requirements of subway operations.

2.4.4 Impact of Supervision Weights on Generalization Capabilities

In this subsection, the impact of supervision weights on the model generalization capability is analyzed by investigating the effect of α_{s} and m. The reward curves under different parameters were depicted, where the reward values reflect the algorithm performance quantitatively during training. For discussion purposes, the following model performance is evaluated with T^{PL} sampled every 5 s from $[T^{\mathrm{PL,min}}, T^{\mathrm{PL,max}}]$, namely, 190 s, 195 s, …, 230 s.

2.4.4.1 Effect of α_{s}

Fig. 2.12 shows the sensitivity of α_{s} under fixed m. All curves are averaged across 5 runs, with the bold lines and the shaded area representing the average reward gained and a single standard deviation, respectively. From the figure, when $\alpha_{\mathrm{s}} = 0$, it is pure RL training. The rewards gained are significantly less than other curves within the maximum episode length, and the learning curve has a large variation even if trained for a long time. This indicates that pure RL training is time-consuming and more difficult to find the optimum due to the complexity of the problem compared with SRL. When α_{s} is larger, the SL supervision accelerates training and improves average performance on different trip times as more rewards are gained. Nevertheless, from the scale magnification of Fig. 2.12(a), if $\alpha_{\mathrm{s}} = 1$, the curve slowly decreases after gaining a high reward. This is due to the overfitting of the agent to the supervisors policy

$\mu_{\mathrm{sl}}(s)$. Note that $\alpha_{\mathrm{s}}=1$ does not represent pure SL training since RL is in effect for the critic.

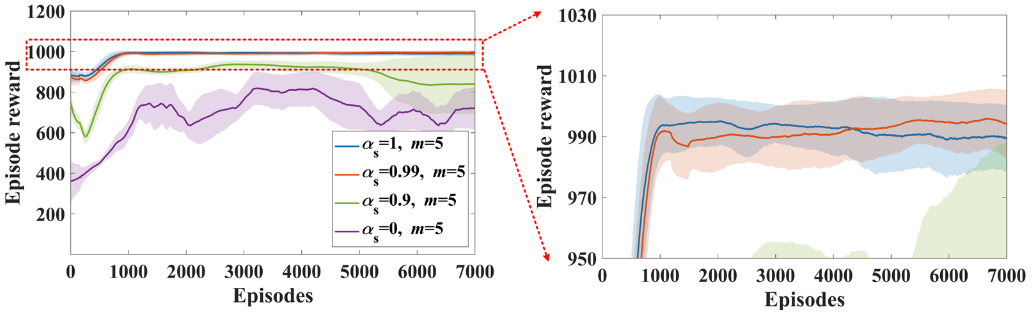
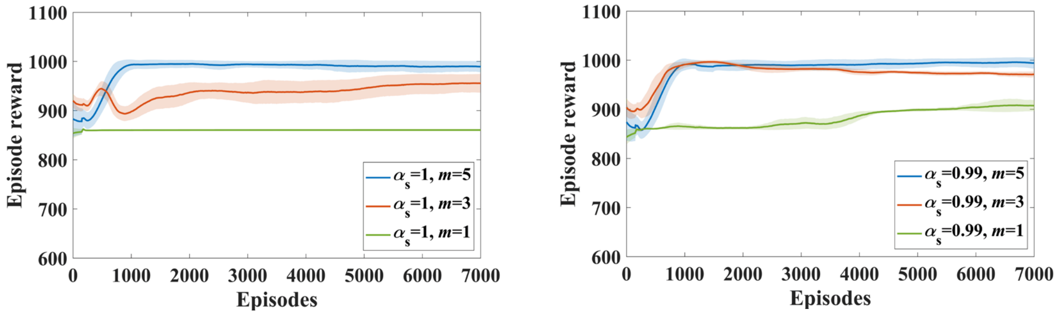
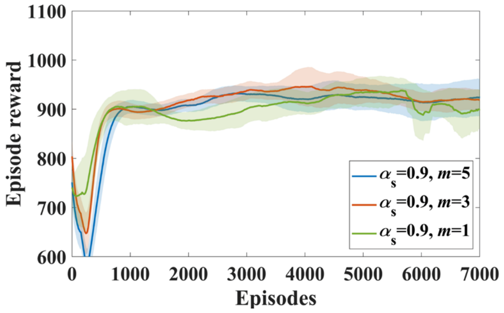


Fig. 2.11 Reward curves under different α_{s}.



(a)

(b)



(c)

Fig. 2.12 Reward curves under different *m* with (a) $\alpha_{\mathrm{s}}=1$, (b) $\alpha_{\mathrm{s}}=0.99$, and (c) $\alpha_{\mathrm{s}}=0.9$.

2.4.4.2 Effect of m

Fig. 2.12 shows the sensitivity of *m* under fixed α_{s}. The learning curves for different m with constant α_{s} are shown in Fig. 2.12. With larger m, the rewards gained within the maximum episode length increase. The highest reward is obtained by $m = 5$. When $\alpha_s = 0.9$, since the total SL contribution is small, the rewards gained under different m are similar. $\alpha_s = 0.99$ and $m = 5$ is the best and default parameter for SRL–EETTO. Compared with pure RL, the designed SRL training architecture improves model generalization capability while accelerating training.

2.5 Summary

In this chapter, an SRL–EETTO approach is proposed for enabling iATO of modern URTs. The research mainly includes the following aspects.

First, the real-time train operation under uncertain disturbances is formulated as an MDP with a goal-directed reward design method to systematically optimize multiple operation objectives of energy saving, ride comfort, punctuality, and safety. Then, a two-step SAC architecture based on the S-TD3 algorithm is developed to solve the MDP and generate optimal train trajectories online. Finally, simulations are implemented to validate the effectiveness of the SRL–EETTO using in-service subway line data.

The key findings of the designated case study are summarized as follows: The proposed approach shows superior average energy saving of 18.5% and ride comfort improvement of 54.7% compared to the practical driving data while providing satisfactory performance on punctuality and safety. 2) The adaptability of the proposed approach to online running time adjustments, uncertain train masses, and uncertain resistance conditions has been verified. 3) a train trajectory configuration suggestion based on the proposed approach have been given, where the increased number of trajectories can improve the scheduling flexibility.

Chapter 3: Sizing and Control Optimization for Hybrid Energy Storage System-Integrated Traction Substation Operation Based on Multi-Task Reinforcement Learning

Nomenclature in this chapter

A. Multi-Task Reinforcement Learning Elements

AE , VE	Action advantage estimation and value estimation
a , s , r , z	Action, state, reward, and task
B	Replay buffer
l	Layer of the Q network in soft module
P^l	Connection probability between layer l and $l+1$
Q^* , Q^s	Expected return and Q value of single-task agent
$\mathcal{A},\mathcal{R},\mathcal{S},\mathcal{P}$	Sets of actions, rewards, states, and transitions
$\mathcal{Z}$, $\mathcal{Z}_H$, $\mathcal{Z}_P$	Sets of tasks, headway tasks, and combination tasks of trajectories
$\mathcal{L}$, $\mathcal{L}_\varphi$, $\mathcal{L}_{KL}$	Weighted Q loss, Q loss, and knowledge transfer loss
μ	Agent policy
γ	Discount factor
τ	Soft update rate
λ	Weight of knowledge transfer
ξ_q	Learning rate of neural networks
φ , φ' , φ_r	Parameters of Q, target Q, and routing networks

B. Indices

$d \in \{1,2,\cdots,D\}$	Index of average onboard passengers
$h \in \{1,2,\cdots,H\}$	Index of train headways

$i \in \{1, 2, \cdots, I\}$	Index of stations or traction substations
$j \in \{1, 2, \cdots, J\}$	Index of stations or traction substations except for station or substation i
$k \in \{1, 2, \cdots, K\}$	Index of trains
$p \in \{1, 2, \cdots, P\}$	Index of train trajectories
$w \in \{1, 2, \cdots, W\}$	Index of transitions sampled from replay buffer

C. Time Scales

Δn, n, N	Increment, current time step, and time horizon on a long time scale for economic dispatch and prediction (e.g., sub-hourly or hourly)
Δt, t, T	Increment, current time step, and time horizon on a short time scale for real-time train and HESS control (e.g., sub-minutely)

D. Variables

a^{TR}, x^{TR}, v^{TR}, d^{TR}	Train acceleration (m/s^3), position (m), speed (m/s), and direction (up/down)
F^{TR}	Train total resistance (N)
H_k	Headway of train k (s)
I^{SC}, I^{BT}, I^{SUB}, I^{TR}	Currents of supercapacitor, battery, substation, and train (A)
J^{SUB}, J^{OM}	Costs of electricity trading and HESS operation ($)
J^{INV}, J^{REP}, J^{FIX}	Costs of investment, replacement, and installation ($)
K_n	Number of trains running at interval n.
L_{BT}	Estimated battery life (year)
N_{BT}, N_{SC}, N_{DC}	Number of battery, supercapacitor, and converter modules
N^{REP}	Replacement frequency
N^{B}	Passengers who are onboard

$P^{\text{SUB}}, P^{\text{TR}}$	Powers of substation and train (W)
$P^{\text{SC,CH}}, P^{\text{SC,DIS}}$	Charging and discharging powers of supercapacitor (W)
$P^{\text{BT,CH}}, P^{\text{BT,DIS}}$	Charging and discharging powers of battery (W)
$R^{\text{V}}, R^{\text{BT}}$	Resistances of contact line and battery (Ω)
$\text{SoE}^{\text{SC}}, \text{SoE}^{\text{BT}}$	SoEs of supercapacitor and battery (%)
$T^{\text{D}}, T^{\text{PL}}, T^{\text{RTTR}}$	Delay, planned running, and rescheduled running times (s)
ΔT	Running time increment in train trajectory set (s)
$U^{\text{CH}}, U^{\text{DIS}}$	Charge and discharge voltage thresholds of HESS (V)
$U^{\text{SC}}, U^{\text{C}}$	Terminal and capacitance voltages of supercapacitor (V)
$U^{\text{BT}}, U^{\text{OCV}}$	Terminal and open-circuit voltages of battery (V)
$U^{\text{SUB}}, U^{\text{TR}}, U^{\text{L}}$	Voltages of substation, train and pantograph (V)
X^{L}	Distance of between train and substation (m)
δ	Train trajectory of a specific section
θ^{S}	Sensitivity of traction energy consumption (J)
$E, \Delta E$	Traction energy consumption of a specific train trajectory and its difference between train trajectories (J)
η^{BR}	Proportion of train braking power to traction network (%)
η^{PA}	Power allocation ratio of HESS (%)
ρ, f_{ρ}	Scenario and its probability
$\mathbf{Y}$	Admittance matrix of the TN

E. Parameters

c_{SUB}	Unit cost of electricity trading (\$/kWh)
$c_{\text{SC}}^{\text{OM}}, c_{\text{BT}}^{\text{OM}}$	Unit costs of supercapacitor and battery operation (\$/MWh)

$c_{\mathrm{SC}}^{\mathrm{INV}}, c_{\mathrm{BT}}^{\mathrm{INV}}, c_{\mathrm{DC}}^{\mathrm{INV}}$	Unit costs of supercapacitor, battery, and converter investment ($/module)
C^{SC}	Equivalent capacitance of supercapacitor (F)
I^{R}	Interest rate (%)
I^{CRT}	Critical substation load current (A)
L	System lifetime (year)
$M^{\mathrm{TR}}, M^{\mathrm{P}}$	Total vehicle mass (kg) and passenger mass per person (kg/person)
R^{L}	Unit resistance of contact line (Ω/km)
$R^{\mathrm{P}}, R^{\mathrm{SUB}}, R^{\mathrm{SC}}$	Resistances of pantograph, substation, and supercapacitor (Ω)
$U_0^{\mathrm{SUB}}, U_1^{\mathrm{BR}}, U_2^{\mathrm{BR}}$	No-load voltage and two braking resistor voltage thresholds (V)
$Q^{\mathrm{BT,norm}}$	Nominal battery capacity (Ah)
α, β	OD element and passenger arrival rate
$\eta^{\mathrm{SC}}, \eta^{\mathrm{BT}}, \eta^{\mathrm{TR}}$	Efficiencies of supercapacitor, battery, and train motor
η^{CR}	Capital recovery factor

3.1 Background

The increasing traction energy consumption induced by the rapid growth of passenger demand has underscored the necessity of developing effective hybrid energy storage system (HESS) sizing and control technologies for traction substations to improve energy and cost efficiencies. Nowadays, various methods have been successfully applied to HESS sizing and control (literature reviews in section 1.3). Nevertheless, the synergistic optimization of HESS sizing and control under the dynamic and uncertain urban rail transit (URT) traction network (TN) energy flows requires comprehensive consideration (substation-level challenges in section 1.4). Therefore, this chapter focuses on developing a multi-task reinforcement learning-based sizing and control optimization (MTRL–SCO) approach for enhancing the economic operations of the supercapacitor–battery HESSs and their integrated traction substations under dynamic spatial-temporal URT traffic at the 2nd (substation) level. Specifically, the main contributions of this chapter are outlined as follows:

- A synergistic sizing and control optimization framework is proposed for the coordinated operations of HESSs and traction substations. An iterative sizing optimization approach considering daily service patterns is devised to minimize the HESS life cycle cost (LCC). The sizing-specific HESS control problem under various spatial-temporal traction load distributions is modeled as a multi-task Markov decision process (MTMDP), where the voltage thresholds and power allocations are jointly optimized for minimizing the operation cost.
- A dynamic traffic model (DTM) considering passenger flow fluctuation and delay-induced traffic regulation is formulated to characterize multi-train traction load uncertainty for enhancing HESS control decisions. A Copula-based passenger flow scenario generation method is proposed to capture dependencies between multi-station origin-destination (OD) demands. A real-time timetable rescheduling (RTTR) algorithm incorporating the traction energy-passenger-time (TEPT) sensitivity matrix is developed to optimize the energy-efficient rescheduled timetable and train trajectories under uncertain short delays.
- An MTRL algorithm based on a dueling double deep Q network with knowledge transfer (KT-D3QN) is presented for solving the MTMDP effectively. A policy distillation annealing method is developed to learn a

generalized multi-task HESS control policy simultaneously and stably from task-specific agents and dynamic train operation environments. Soft modulation and gradient manipulation techniques are employed to handle inter-task parameter sharing and conflicts.

Finally, comparative studies based on a real-world subway have validated the effectiveness of the proposed approach for LCC reduction of HESS-integrated traction substation operation under URT traffic. The remaining parts of the chapter are organized as follows. Section 3.2 illustrates the problem formulation with structure and modeling of HESS-integrated traction substations and their traction networks (TNs), followed by the formulation of the HESS sizing and optimization model and the analysis of HESS control parameters and URT operation uncertainties on the operation cost and RB energy utilization. Section 3.3 presents the proposed MTRL–SCO approach, including the formulation of the DTM, MTMDP, and KT-D3QN algorithm. Section 3.4 reports case studies and their results. Section 3.5 gives the summary.

3.2 Problem Formulation

3.2.1 Structure of Traction Substations With Hybrid Energy Storage Systems

The typical structure of HESS-integrated traction substations is shown in Fig. 3.1. The substation contains a unidirectional 24-pulse wave diode rectifier. The HESS connects to the traction substation by DC-DC converters. Generally, the passenger flow prediction is conducted at a large time interval (e.g., 15 min), while the HESS control is carried out at a small time interval (e.g., 1 s). Thus, we use $n = 1,2,,\cdots,N$ to denote the “prediction interval”, and $t = 1,2,,\cdots,T$ to denote the “control interval”. The following assumptions are considered: 1) The OD matrix is deterministic, while the passenger arrival rate varies [156]. 2) The train stops at each station with a pre-determined dwell time and running time. While delays extend the planned dwell time, the total running and dwell times are unchanged [144].

In a scenario ρ, $P_{i,\rho,t}^{\mathrm{SUB}}$ is the power of ith traction substation, $1 \le i \le I$, $P_{\rho,t}^{\mathrm{SC,CH}}$ and $P_{\rho,t}^{\mathrm{SC,DIS}}$ are the discharging and charging power of the supercapacitor, respectively, $P_{\rho,t}^{\mathrm{BT,CH}}$ and $P_{\rho,t}^{\mathrm{BT,DIS}}$ are the discharging and charging power of the battery, respectively, $P_{k,\rho,t}^{\mathrm{TR}}$ is the power of kth train, $1 \le k \le K$. When a delay time $T_{i,\rho}^{\mathrm{D}}$ is known, the planned running time $T_{i,j,\rho}^{\mathrm{PL}}$ is changed to $T_{i,j,\rho}^{\mathrm{RTTR}}$ by RTTR, and an optimal trajectory is selected from a pre-programmed trajectory set.

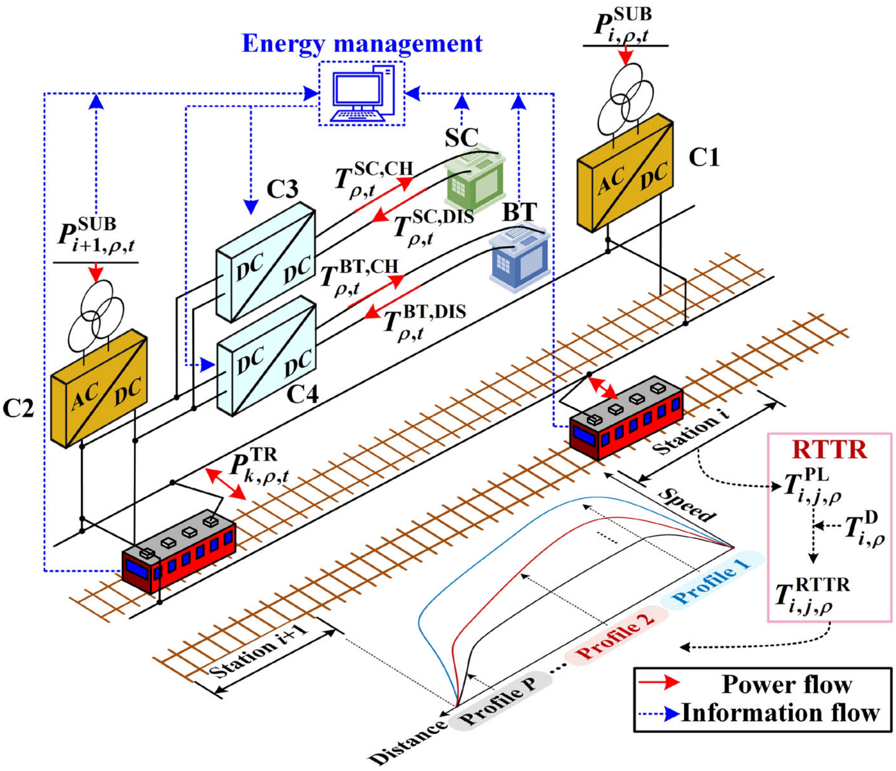

Energy management
$P_{i,\rho,t}^{\mathrm{SUB}}$
SC
C1
C3
$T_{\rho,t}^{\mathrm{SC,CH}}$
$T_{\rho,t}^{\mathrm{SC,DIS}}$
BT
$P_{i+1,\rho,t}^{\mathrm{SUB}}$
$T_{\rho,t}^{\mathrm{BT,CH}}$
$T_{\rho,t}^{\mathrm{BT,DIS}}$
AC DC
DC DC
DC DC
C4
C2
AC DC
$P_{k,\rho,t}^{\mathrm{TR}}$
Station i
RTTR
$T_{i,j,\rho}^{\mathrm{PL}}$
$T_{i,\rho}^{\mathrm{D}}$
$T_{i,j,\rho}^{\mathrm{RTTR}}$
Speed
Profile 1
Profile 2
Profile P
Station i+1
Distance
Power flow
Information flow

Fig. 3.1 The structure of HESS-integrated traction substations.

3.2.2 Modeling of Hybrid Energy Storage System-Integrated Traction Substations and Their Traction Networks

3.2.2.1 Equivalent Circuit Model Overview

The equivalent circuit model of such TN (Fig. 3.2) includes traction substation (3.2), train (3.3)–(3.4), battery (3.7)–(3.8), and supercapacitor (3.9)–(3.10) models. In the figure, $X^{\mathrm{L}}_{k,i,\rho,t}$ and $X^{\mathrm{L}}_{k,i+1\rho,t}$ are the distance of the train k to station i and i+1, respectively. R^{P} is the pantograph resistance. R^{L} is the contact line resistance per km. Since the position of trains is changing during operation, the contact line resistance of each circuit branch becomes time varying. Therefore, the contact line resistance is modeled as a variable resistance. For instance, the resistance between train k and station i is $R^{\mathrm{V}}_{k,i} = R^{\mathrm{L}} X^{\mathrm{L}}_{k,i,\rho,t}$. Then, the admittance matrix $\mathbf{Y}$ of the circuit can be derived. For instance, for the TN in Fig. 3.2,

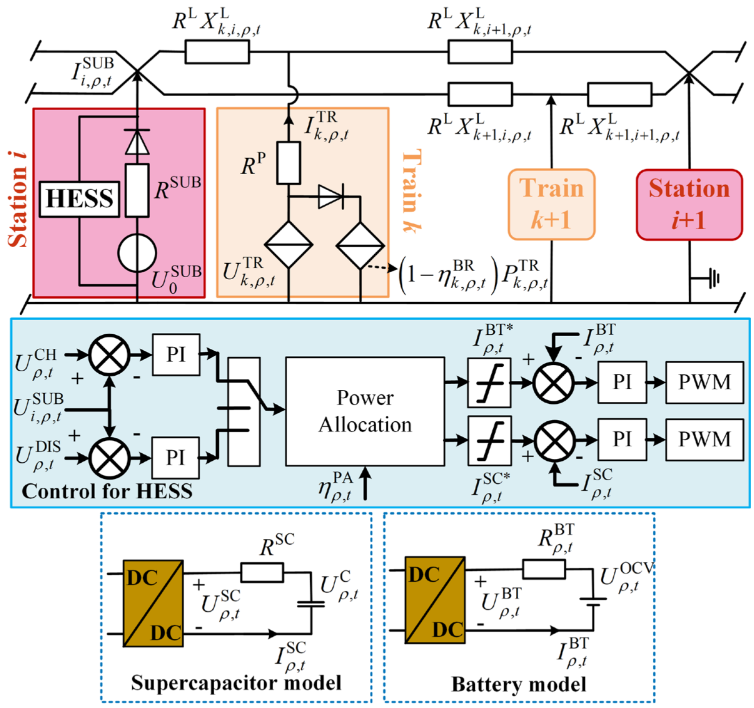

$R^{\mathrm{L}} X^{\mathrm{L}}_{k,i,\rho,t}$
$R^{\mathrm{L}} X^{\mathrm{L}}_{k,i+1,\rho,t}$
$I^{\mathrm{SUB}}_{i,\rho,t}$
$I^{\mathrm{TR}}_{k,\rho,t}$
$R^{\mathrm{L}} X^{\mathrm{L}}_{k+1,i,\rho,t}$
$R^{\mathrm{L}} X^{\mathrm{L}}_{k+1,i+1,\rho,t}$
Station i
HESS
R^{SUB}
U^{SUB}_0
R^{P}
$U^{\mathrm{TR}}_{k,\rho,t}$
Train k
Train k+1
Station i+1
$(1-\eta^{\mathrm{BR}}_{k,\rho,t})P^{\mathrm{TR}}_{k,\rho,t}$
$U^{\mathrm{CH}}_{\rho,t}$
$U^{\mathrm{SUB}}_{i,\rho,t}$
$U^{\mathrm{DIS}}_{\rho,t}$
+
-
PI
Power Allocation
$\eta^{\mathrm{PA}}_{\rho,t}$
$I^{\mathrm{BT*}}_{\rho,t}$
$I^{\mathrm{BT}}_{\rho,t}$
$I^{\mathrm{SC*}}_{\rho,t}$
$I^{\mathrm{SC}}_{\rho,t}$
PWM
Control for HESS
DC
DC
R^{SC}
$U^{\mathrm{SC}}_{\rho,t}$
$U^{\mathrm{C}}_{\rho,t}$
$I^{\mathrm{SC}}_{\rho,t}$
Supercapacitor model
$R^{\mathrm{BT}}_{\rho,t}$
$U^{\mathrm{BT}}_{\rho,t}$
$U^{\mathrm{OCV}}_{\rho,t}$
$I^{\mathrm{BT}}_{\rho,t}$
Battery model

Fig. 3.2 Equivalent circuit model of TNs with a HESS-integrated traction substation.

$$\mathbf{Y}=\begin{bmatrix}\frac{1}{R^{\mathrm{P}}}+\frac{1}{R_{k,i}^{\mathrm{V}}}+\frac{1}{R_{k,i+1}^{\mathrm{V}}} & 0 & -\frac{1}{R_{k,i}^{\mathrm{V}}} & -\frac{1}{R_{k,i+1}^{\mathrm{V}}} \\ 0 & \frac{1}{R^{\mathrm{P}}}+\frac{1}{R_{k+1,i}^{\mathrm{V}}}+\frac{1}{R_{k+1,i+1}^{\mathrm{V}}} & -\frac{1}{R_{k+1,i}^{\mathrm{V}}} & -\frac{1}{R_{k+1,i+1}^{\mathrm{V}}} \\ -\frac{1}{R_{k,i}^{\mathrm{V}}} & -\frac{1}{R_{k+1,i}^{\mathrm{V}}} & \frac{1}{R^{\mathrm{SUB}}}+\frac{1}{R_{k,i}^{\mathrm{V}}}+\frac{1}{R_{k,i+1}^{\mathrm{V}}} & 0 \\ -\frac{1}{R_{k,i+1}^{\mathrm{V}}} & -\frac{1}{R_{k+1,i+1}^{\mathrm{V}}} & 0 & \frac{1}{R^{\mathrm{SUB}}}+\frac{1}{R_{k+1,i}^{\mathrm{V}}}+\frac{1}{R_{k+1,i+1}^{\mathrm{V}}}\end{bmatrix}. \tag{3.1}$$

3.2.2.2 Traction Substation Equivalent Circuit Model

The 24-pulse rectifier in the traction substation contains two traction transformers and four sets of diode rectifier bridges, and the transformers operate in parallel. Due to the impedances of transformers, characteristics of rectifier bridges, topology, and operating conditions, the output characteristics of the 24-pulse rectifier is complex. On the one hand, the output voltage of the 24-pulse rectifier decreases with the increasing load current. On the other hand, the equivalent resistance of the rectifier varies with the load current. In practice, such output characteristics of the 24-pulse rectifier is generally simplified with piecewise linear functions [120].

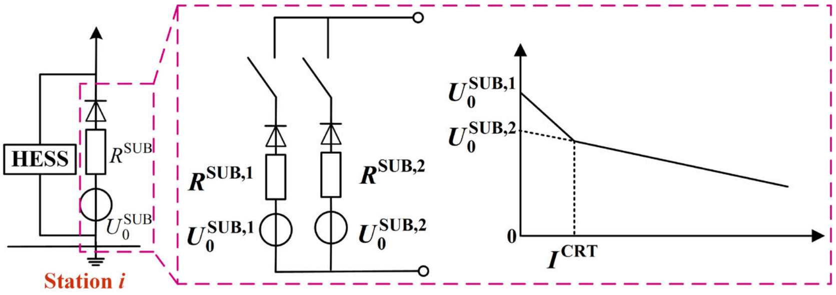


Fig. 3.3 Output characteristics and the equivalent circuit of the traction substation.

Specifically, the simplified output characteristics of the 24-pulse rectifier is shown in Fig. 3.3. The equivalent traction substation model can be established as a Thevenin circuit model, where the parameters of the ideal voltage source U_0^{SUB} and the internal resistance R^{SUB} is determined by the intercept and slope of the curve, respectively. A diode is in series with the Thevenin circuit to simulate the unidirectional energy flow of the rectifier.

$$U_{i,\rho,t}^{\mathrm{SUB}}=\begin{cases}U_0^{\mathrm{SUB,1}}-R^{\mathrm{SUB,1}}I_{i,\rho,t}^{\mathrm{SUB}}, & I^{\mathrm{CRT}}\geq I_{i,\rho,t}^{\mathrm{SUB}}\geq 0,\\ U_0^{\mathrm{SUB,2}}-R^{\mathrm{SUB,2}}I_{i,\rho,t}^{\mathrm{SUB}}, & I^{\mathrm{CRT}}<I_{i,\rho,t}^{\mathrm{SUB}},\end{cases} \tag{3.2}$$

where $U_{i,\rho,t}^{\mathrm{SUB}}$ and $I_{i,\rho,t}^{\mathrm{SUB}}$ are the traction substation voltage and current, respectively, U_0^{SUB} is the no-load traction substation voltage, R^{SUB} is the traction substation resistance. $U_0^{\mathrm{SUB,1}}$ and $U_0^{\mathrm{SUB,2}}$ are the ideal voltage source in current intervals 1 and 2, respectively, $R^{\mathrm{SUB,1}}$ and $R^{\mathrm{SUB,2}}$ are the internal resistance in current intervals 1 and 2, respectively, I^{CRT} is the critical load current.

3.2.2.3 Train Equivalent Circuit Model

The train is modeled as an equivalent controlled power source

$$U_{k,\rho,t}^{\mathrm{TR}}I_{k,\rho,t}^{\mathrm{TR}}=\eta_{k,\rho,t}^{\mathrm{BR}}P_{k,\rho,t}^{\mathrm{TR}}, \tag{3.3}$$

$$\eta_{k,\rho,t}^{\mathrm{BR}}=\begin{cases}100\%, & U_{k,\rho,t}^{\mathrm{L}}\leq U_1^{\mathrm{BR}},\\ (1-\dfrac{U_{k,\rho,t}^{\mathrm{L}}-U_1^{\mathrm{BR}}}{U_2^{\mathrm{BR}}-U_1^{\mathrm{BR}}})\times 100\%, & U_1^{\mathrm{BR}}<U_{k,\rho,t}^{\mathrm{L}}\leq U_2^{\mathrm{BR}},\\ 0\%, & U_2^{\mathrm{BR}}<U_{k,\rho,t}^{\mathrm{L}},\end{cases} \tag{3.4}$$

where $U_{k,\rho,t}^{\mathrm{TR}}$ and $I_{k,\rho,t}^{\mathrm{TR}}$ are the train voltage and current, respectively. $\eta_{k,\rho,t}^{\mathrm{BR}}$ simulates the braking resistor [18], which determines the proportion of braking power delivered to the network.

Generally, the pantograph voltage will rise if the regenerative braking (RB) energy is not absorbed by nearby accelerating trains or the HESS. Therefore, the braking resistor is adopted to avoid this situation. By controlling the duty cycle of the chopper, the redundant RB energy is consumed on the braking resistor. When the pantograph voltage $U_{k,\rho,t}^{\mathrm{L}}$ is lower than the start-up voltage threshold U_1^{BR}, the duty cycle is zero and $\eta_{k,\rho,t}^{\mathrm{BR}} = 100\%$. This indicates that all RB energy can be delivered to the traction network and the braking resistor is not working. When $U_{k,\rho,t}^{\mathrm{L}}$ is higher than U_1^{BR}, the duty cycle increases and $\eta_{k,\rho,t}^{\mathrm{BR}}$ drops linearly. When $U_{k,\rho,t}^{\mathrm{L}}$ reaches the maximum allowed voltage threshold U_2^{BR}, the duty cycle reaches 1 and $\eta_{k,\rho,t}^{\mathrm{BR}} = 0\%$. This indicates that the braking resistor consumes all RB energy.

3.2.2.4 HESS Equivalent Circuit Model

As shown in Fig. 3.2, the HESS in the traction substation generally adopts a voltage-current double-loop control [120]. The outer loop aims to stabilize the traction substation voltage at the charge or discharge voltage threshold of the HESS. When the traction substation voltage $U_{i,\rho,t}^{\mathrm{SUB}}$ is higher than the HESS charge voltage threshold $U_{\rho,t}^{\mathrm{CH}}$, the HESS charges. When the traction substation voltage $U_{i,\rho,t}^{\mathrm{SUB}}$ is lower than the HESS discharge voltage threshold $U_{\rho,t}^{\mathrm{DIS}}$, the HESS discharges. Then, the inner loop aims to control the HESS current at the referential value. A proportional-integral controller is utilized to generate the referential inner-loop current according to the traction substation voltage. Based on the referential inner-loop current and power allocation ratio $\eta_{\rho,t}^{\mathrm{PA}}$, the referential current of the supercapacitor and battery is determined. The referential current is limited by a current limiter, which compares the referential current with the maximum current of the supercapacitor and battery. Finally, another proportional-integral controller and a pulse width modulation (PWM) module are utilized to generate the charge/discharge command for the HESS.

Thus, the status of the supercapacitor at the network side (high-voltage side) can be described by constant-voltage and constant-current modes. When the referential current of the supercapacitor is lower than its maximum current, the current limiter is not in effect, and the supercapacitor is essentially a controlled voltage source. When the referential current of the supercapacitor reaches its maximum current, the current limiter is in effect, and the supercapacitor is essentially a current source that outputs its constant maximum current. The status of the battery at the network side is the same as that of the supercapacitor. Therefore, when both the supercapacitor and battery work in the constant-voltage mode, the traction substation voltage $U_{i,\rho,t}^{\mathrm{SUB}}$ is known, which equals the HESS charge/discharge voltage threshold $U_{\rho,t}^{\mathrm{CH}}$ (or $U_{\rho,t}^{\mathrm{DIS}}$). Nevertheless, when either or both the supercapacitor and battery work in the constant-current mode, their currents are known according to the power allocation ratio $\eta_{\rho,t}^{\mathrm{PA}}$, but the traction substation voltage $U_{i,\rho,t}^{\mathrm{SUB}}$ is higher than the HESS charge voltage threshold $U_{\rho,t}^{\mathrm{CH}}$ (or lower than the HESS discharge threshold $U_{\rho,t}^{\mathrm{DIS}}$).

On the low-voltage side, for the supercapacitor, the first-order RC equivalent circuit model is utilized. The parameters of such a model can be easily identified [120]. The model is essentially a resistor R^{SC} in series with a capacitor C^{SC}, where $U_{\rho,t}^{\mathrm{SC}}$ and $I_{\rho,t}^{\mathrm{SC}}$ are its output voltage and current, respectively, $U_{\rho,t}^{\mathrm{C}}$ is the capacitor voltage. For battery, it is modeled as an open-circuit voltage (OCV) source $U_{\rho,t}^{\mathrm{OCV}}$ in series with a resistor $R_{\rho,t}^{\mathrm{BT}}$. $U_{\rho,t}^{\mathrm{BT}}$ and $I_{\rho,t}^{\mathrm{BT}}$ are its output voltage and current, respectively. Both $U_{\rho,t}^{\mathrm{OCV}}$ and $R_{\rho,t}^{\mathrm{BT}}$ are nonlinear functions of the state-of-energy (SoE), which can be fit by polynomial functions [166]. The SoE-OCV and SoE-resistance relationship of the battery in this chapter can be fitted by [167]

$$U_{\rho,t}^{\mathrm{OCV}}=1.186\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^3-1.476\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^2+1.019\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}+1.758, \tag{3.5}$$

$$R_{\rho,t}^{\mathrm{BT}}=\begin{cases}1.055\times10^{-2}\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^3-1.706\times10^{-2}\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^2+ \\ 9.131\times10^{-3}\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}+9.757\times10^{-5}, & \text{Charge,} \\ 1.394\times10^{-2}\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^4-3.901\times10^{-2}\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^3+ \\ 3.929\times10^{-2}\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^2-1.670\times10^{-3}\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}+ & \text{Discharge.} \\ 3.880\times10^{-3}, \end{cases} \tag{3.6}$$

Moreover, the charging/discharging efficiency of the supercapacitor and battery is considered by η^{SC} and η^{BT}, respectively. Δt is the increment of the control interval. To sum up, the status of the supercapacitor and battery at the low-voltage side is modeled as

$$I_{\rho,t}^{\mathrm{BT}}=\begin{cases}P_{\rho,t}^{\mathrm{BT}}/\eta^{\mathrm{BT}}U_{\rho,t}^{\mathrm{BT}}, & P_{\rho,t}^{\mathrm{BT}}=P_{\rho,t}^{\mathrm{BT,DIS}}, \\ \eta^{\mathrm{BT}}P_{\rho,t}^{\mathrm{BT}}/U_{\rho,t}^{\mathrm{BT}}, & P_{\rho,t}^{\mathrm{BT}}=P_{\rho,t}^{\mathrm{BT,CH}}, \end{cases} \tag{3.7}$$

$$U_{\rho,t}^{\mathrm{BT}}=U_{\rho,t}^{\mathrm{OCV}}-I_{\rho,t}^{\mathrm{BT}}R_{\rho,t}^{\mathrm{BT}}, \tag{3.8}$$

$$U_{\rho,t}^{\mathrm{SC}}=U_{\rho,t}^{\mathrm{C}}-I_{\rho,t}^{\mathrm{SC}}R^{\mathrm{SC}},\quad U_{\rho,t}^{\mathrm{C}}=U_{\rho,t-1}^{\mathrm{C}}-I_{\rho,t}^{\mathrm{SC}}\Delta t/C^{\mathrm{SC}}, \tag{3.9}$$

$$I_{\rho,t}^{\mathrm{SC}}=\begin{cases}\dfrac{U_{\rho,t}^{\mathrm{C}}-\sqrt{\left(U_{\rho,t}^{\mathrm{C}}\right)^2-4R^{\mathrm{SC}}P_{\rho,t}^{\mathrm{SC}}/\eta^{\mathrm{SC}}}}{2R^{\mathrm{SC}}}, & P_{\rho,t}^{\mathrm{SC}}=P_{\rho,t}^{\mathrm{SC,DIS}}, \\ -\dfrac{U_{\rho,t}^{\mathrm{C}}-\sqrt{\left(U_{\rho,t}^{\mathrm{C}}\right)^2-4R^{\mathrm{SC}}\eta^{\mathrm{SC}}P_{\rho,t}^{\mathrm{SC}}}}{2R^{\mathrm{SC}}}, & P_{\rho,t}^{\mathrm{SC}}=P_{\rho,t}^{\mathrm{SC,CH}}. \end{cases} \tag{3.10}$$

3.2.2.5 HESS Degradation Model

For simplicity, the degradation of the battery and supercapacitor is estimated based on the rainflow counting method [166], which analyzes the cyclic loading history of a material or structure, while the effects of other aging parameters (e.g., temperature, current, etc.) are ignored. In chapter 5, we provide a more refined degradation model to consider the electrothermal coupling relationship of HESS. Besides, the detailed steps to implement both rainflow counting and electrothermal-coupled methods for degradation estimation are illustrated in Appendix A.

3.2.3 Formulation of Hybrid Energy Storage System Sizing and Control Optimization Problem

3.2.3.1 Control Optimization Model

The power flows are modeled by (3.11)–(3.14). The electricity trading and HESS operation costs are calculated by (3.15)–(3.16). $c_{\mathrm{SC}}^{\mathrm{OM}}$ and $c_{\mathrm{BT}}^{\mathrm{OM}}$ are the supercapacitor and battery operation cost per MWh, respectively, c_{SUB} is the trading cost per kWh, $J_{\rho,t}^{\mathrm{SUB}}$ is the electricity trading cost, $J_{\rho,t}^{\mathrm{OM}}$ is the HESS operation cost,

$$\mathbf{Y}\begin{bmatrix}\mathbf{U}^{l} & \mathbf{U}^{\mathbf{s}}\end{bmatrix}^{T}=\begin{bmatrix}\mathbf{I}^{\mathbf{t}} & \mathbf{I}^{\mathbf{s}}\end{bmatrix}^{T}, \tag{3.11}$$

$$\begin{aligned}&\mathbf{U}^{\mathbf{t}}=\begin{bmatrix}U_{1,\rho,t}^{\mathrm{L}} & \cdots & U_{K,\rho,t}^{\mathrm{L}}\end{bmatrix}^{T},\ \ \mathbf{I}^{\mathbf{t}}=\begin{bmatrix}I_{1,\rho,t}^{\mathrm{TR}} & \cdots & I_{K,\rho,t}^{\mathrm{TR}}\end{bmatrix}^{T},\\&\mathbf{U}^{\mathbf{s}}=\begin{bmatrix}U_{1,\rho,t}^{\mathrm{SUB}} & \cdots & U_{I,\rho,t}^{\mathrm{SUB}}\end{bmatrix}^{T},\ \ \mathbf{I}^{\mathbf{s}}=\begin{bmatrix}I_{1,\rho,t}^{\mathrm{S}} & \cdots & I_{I,\rho,t}^{\mathrm{S}}\end{bmatrix}^{T},\end{aligned} \tag{3.12}$$

$$U_{k,\rho,t}^{\mathrm{L}}=U_{k,\rho,t}^{\mathrm{TR}}-R^{\mathrm{P}}I_{k,\rho,t}^{\mathrm{TR}}, \tag{3.13}$$

$$I_{i,\rho,t}^{\mathrm{S}}=I_{i,\rho,t}^{\mathrm{SUB}}+\left(P_{\rho,t}^{\mathrm{SC}}+P_{\rho,t}^{\mathrm{BT}}\right)/U_{i,\rho,t}^{\mathrm{SUB}}, \tag{3.14}$$

$$J_{\rho,t}^{\mathrm{SUB}}=\sum\nolimits_{i=1}^{I}c_{\mathrm{SUB}}P_{i,\rho,t}^{\mathrm{SUB}}\Delta t, \tag{3.15}$$

$$J_{\rho,t}^{\mathrm{OM}}=c_{\mathrm{SC}}^{\mathrm{OM}}\left|P_{\rho,t}^{\mathrm{SC}}\right|\Delta t+c_{\mathrm{BT}}^{\mathrm{OM}}\left|P_{\rho,t}^{\mathrm{BT}}\right|\Delta t. \tag{3.16}$$

The aim of HESS control is to minimize the overall operation cost

$$\begin{aligned}&\min J_{\rho}=\sum\nolimits_{t=1}^{T}\left(J_{\rho,t}^{\mathrm{SUB}}+J_{\rho,t}^{\mathrm{OM}}\right),\\&\text{s.t. (3.2)-(3.16).}\end{aligned} \tag{3.17}$$

3.2.3.2 Sizing Optimization Model

Considering the match of converters and HESSs, the number of battery and supercapacitor modules in series is fixed. Hence, the sizing optimization problem aims to find the optimal number of battery and supercapacitor modules in parallel,

$$J^{\mathrm{INV}} = \left[c_{\mathrm{SC}}^{\mathrm{INV}} N_{\mathrm{SC}} + c_{\mathrm{BT}}^{\mathrm{INV}} N_{\mathrm{BT}} + c_{\mathrm{DC}}^{\mathrm{INV}} N_{\mathrm{DC}} \right] \cdot \eta^{\mathrm{CR}}, \tag{3.18}$$

$$J^{\mathrm{REP}} = \sum\nolimits_{rep=1}^{N^{\mathrm{REP}}} \frac{c_{\mathrm{BT}}^{\mathrm{INV}} N_{\mathrm{BT}} + c_{\mathrm{DC}}^{\mathrm{INV}} N_{\mathrm{DC}}}{\left(1+I^R\right)^{rep \cdot L_{\mathrm{BT}}}} \cdot \eta^{\mathrm{CR}}, \tag{3.19}$$

$$\eta^{\mathrm{CR}} = I^R \cdot \left(1+I^R\right)^L / \left[\left(1+I^R\right)^L + 1\right], \tag{3.20}$$

where J^{INV} and J^{REP} are the investment and replacement cost, respectively, η^{CR} is the capital recovery factor, $c_{\mathrm{SC}}^{\mathrm{INV}}$, $c_{\mathrm{BT}}^{\mathrm{INV}}$, and $c_{\mathrm{DC}}^{\mathrm{INV}}$ are the investment cost of supercapacitor, battery, and converter per module, respectively, N_{SC} and N_{BT} are the number of supercapacitor and battery modules, respectively, N_{DC} is the number of converter modules for HESS, N^{REP} is the replacement frequency, L_{BT} is the estimated battery life, L is the system lifetime, I^R is the interest rate.

Considering various operation uncertainties, such as passenger flow fluctuation and delays, a scenario-based method is adopted. With scenario ρ and occurrence probability f_ρ, the objective can be written as

$$\begin{aligned} &\min J^{\mathrm{LCC}} = \sum_{\rho} f_\rho J_\rho + J^{\mathrm{INV}} + J^{\mathrm{REP}} + J^{\mathrm{FIX}} \eta^{\mathrm{CR}}, \\ &\text{s.t. } (3.18) - (3.20). \end{aligned} \tag{3.21}$$

where J^{FIX} is other installation cost.

3.2.4 Impact Factors for Economic Traction Network Operation and Regenerative Braking Energy Utilization

3.2.4.1 Analysis Overview

In this subsection, the impacts of HESS control parameters and operation uncertainties on economic TN operation and RB energy utilization are analyzed. As an example, based on the same subway line described in subsection 2.4.1, four elevated stations (RJ, RC, TJN, and JH) are selected for this analysis. The timetable of these stations is shown in Table 3.1 and Fig. 3.4. The practical daily service pattern is listed in Table 3.2 [168], where there are 122 daily train services. We treat all headways during 5:30–9:00 and 16:00–19:00 as 350 s, 5:20–5:30 and 19:00–20:00 as 540 s, and 9:00–16:00 and 20:00–22:05 as 660 s. The train parameters are also listed in subsection 2.4.1, where the maximum capacity $N_{\max}^{\mathrm{B}}$ is 1500 [169], the start-up braking resistor voltage threshold U_1^{BR} =900 V, and the maximum allowed voltage U_2^{BR} =1000 V. The traction substation, TN [18] and HESS [166] parameters are listed in Table 3.3 and Table 3.4, respectively. The HESS is assumed to be installed in station 3.

Table 3.1 Planned running and dwell times of the studied sections.

Section	$T_{i,j,\rho}^{\mathrm{PL}}$ (s)	Length (m)	Direction	Dwell time (s)
RJ–RC (1–2)	104	1354	Down	30
RC–TJN (2–3)	165	2337		
TJN–JH (3–4)	151	2265		
JH–TJN (4–3)	151	2265	Up	
TJN–RC (3–2)	162	2337		
RC–RJ (2–1)	105	1354		

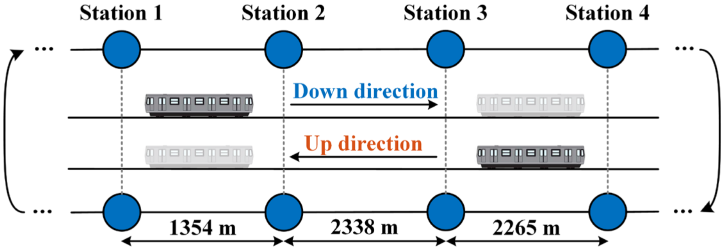


Fig. 3.4 Subway station data and illustration of train operation.

Table 3.2 Daily train service pattern.

Time	Headway (s)	Time	Headway (s)	Time	Headway (s)
5:20–5:30	535	9:00–16:00	660	19:00–20:00	540
5:30–9:00	390	16:00–19:00	350	20:00–22:05	660

Table 3.3 Traction substation and TN parameters.

Item	Value	Item	Value	Item	Value
$U_0^{\mathrm{SUB},1}, U_0^{\mathrm{SUB},2}$	860, 832 V	L	10 years	I^R	2.5%
$R^{\mathrm{SUB},1}, R^{\mathrm{SUB},2}$	0.0161, 0.0236 Ω	R^{P}	0.015 Ω	J^{FIX}	3.2×10^5 \$
I^{CRT}	672 A	R^{L}	0.016 Ω/km	c_{SUB}	0.11 \$/kWh

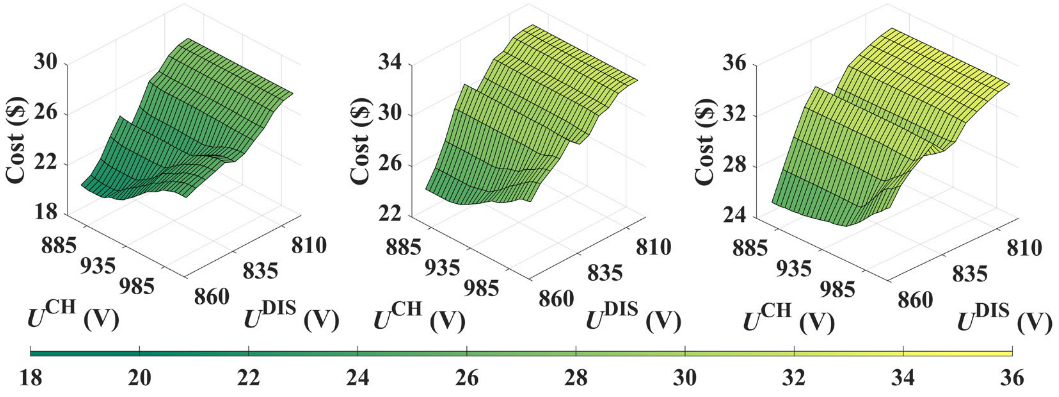
3.2.4.2 Impact of Hybrid Energy Storage System Control Parameters

1) Impact of Voltage Thresholds: In order to analyze the impact of HESS charge/discharge voltage thresholds in detail, the overall TN operation cost and RB energy utilization with respect to the change of voltage thresholds are calculated. In such analysis (Fig. 3.5–Fig. 3.6), all uncertainties of train operation are ignored. Specifically, the number of onboard passengers is fixed to the maximum train capacity $N_{\max}^{\mathrm{B}}$. No uncertain delays or train resistances are considered. Besides, the initial SoEs of supercapacitors and batteries are set as their maximum SoEs. Instead of using a dynamic ratio $\eta_{\rho,t}^{\mathrm{PA}}$, the power allocation strategy of the HESS follows a fixed allocation ratio. The allocation ratio is determined by the maximum battery power divided by the maximum HESS power, which equals 0.20. The thresholds are fixed with respect to the change of time.

Table 3.4 HESS parameters.

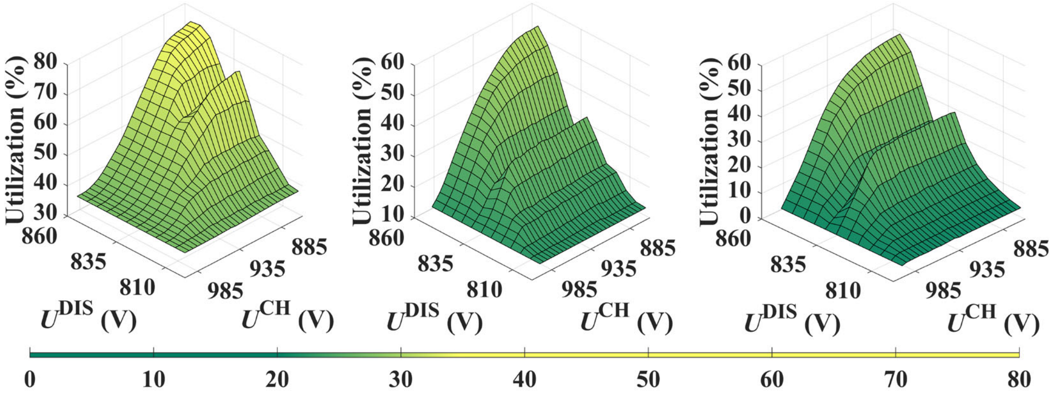
Battery module (LTO 20Ah)			
Item	Value	Item	Value
Nom. voltage	2.3 V	No. in series	292
Nom. capacity	20 Ah	No. in parallel	5
Max. discharge rate	5 C	$c_{\mathrm{BT}}^{\mathrm{OM}}$	1 \$/MWh
SoE range	0.2-0.8	$c_{\mathrm{BT}}^{\mathrm{INV}}$	31.51 \$/module
Supercapacitor module (BMOD00165P48)			
Item	Value	Item	Value
Nom. voltage	48 V	No. in series	14
Nom. capacity	165 F	No. in parallel	15
Nom. current	130 A	$c_{\mathrm{SC}}^{\mathrm{OM}}$	7.5 \$/MWh
Resistance	6.3×10^{-3} Ω	$c_{\mathrm{SC}}^{\mathrm{INV}}$	538 \$/module
SoE range	0.25-0.9		
Converter module			
Item	Value	Item	Value
Max. current	400 A	$c_{\mathrm{DC}}^{\mathrm{INV}}$	38500 \$/module
η^{BT}	0.8	η^{SC}	0.95

From the figure, it can be observed that, generally, the cost increases with a lower charge voltage threshold and higher discharge threshold. This is because, as HESS charge/discharge voltage thresholds are closer to the no-load voltage, more RB energy is reserved for the HESS rather than wasted. Thus, the reserved energy can be further utilized for energy saving. However, the cost drops when the discharge voltage threshold is close to 835 V, which shows a complex relationship between voltage threshold settings and economic operation. For RB energy utilization, similarly, it increases with a lower charge voltage threshold and higher discharge threshold.



(a) (b) (c)

Fig. 3.5 Impact of HESS charge/discharge voltage thresholds on overall TN operation cost under (a) 350s, (b) 540s, and (c) 660s headway.



(a) (b) (c)

Fig. 3.6 Impact of HESS charge/discharge voltage thresholds on RB energy utilization under (a) 350s, (b) 540s, and (c) 660s headway.

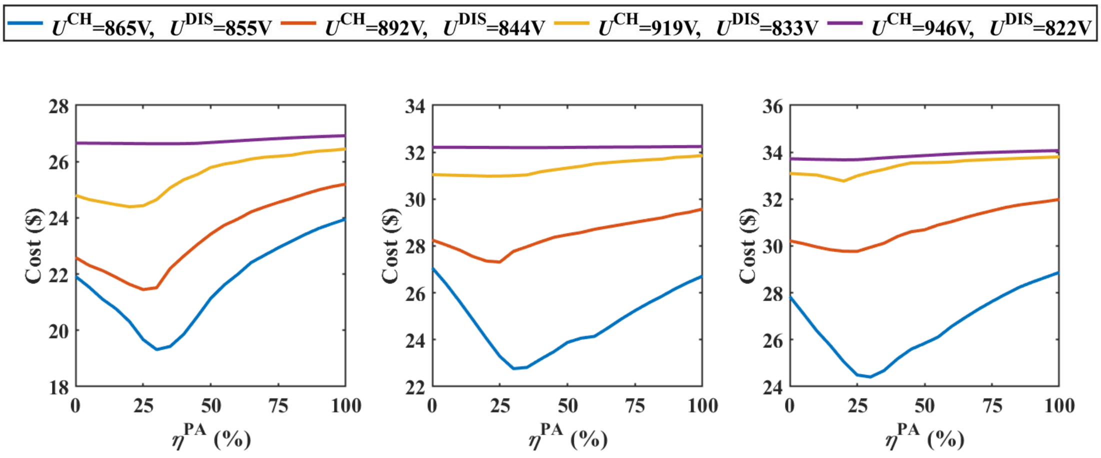
Table 3.5 summarizes the optimal voltage thresholds that have the lowest overall operation cost and highest RB energy utilization. It can be observed that, in most cases, the optimal voltage thresholds are closer to the no-load voltage. However, they vary with different train headways. For instance, when the headway is 540 s, the optimal charge voltage threshold is 892 V, which is closer to the start-up voltage threshold of braking resistors. Therefore, the HESS charge/discharge voltage thresholds should adapt to various train headways for operation cost minimization and RB energy utilization maximization.

Table 3.5 Optimal voltage thresholds in the fixed operation environment.

Objective	Headway (s)	Optimal U^{CH} (V)	Optimal U^{DIS} (V)
Overall operation cost	350	877	855
	540	892	855
	660	868	855
RB energy utilization	350	865	852
	540	877	855
	660	865	854

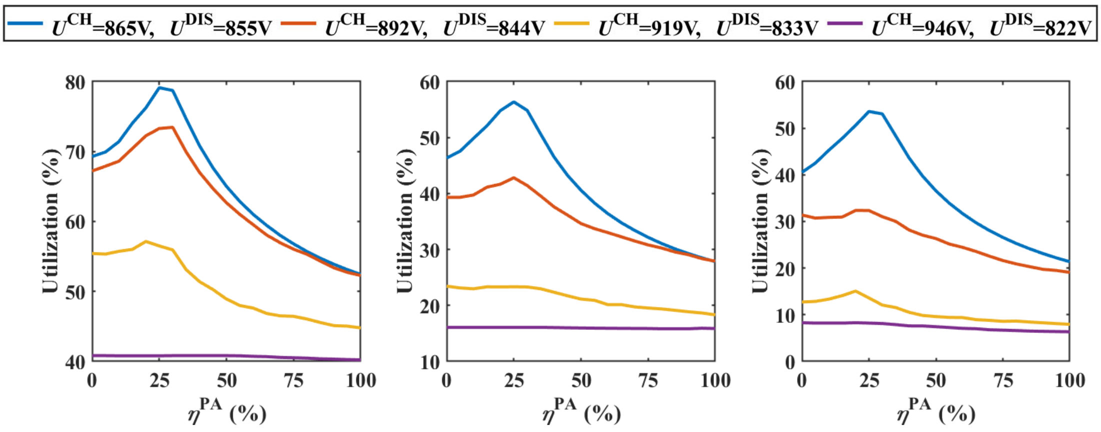
2) Impact of Power Allocation: The power allocation of the HESS can also influence the amount of charge/discharge energy. Similarly, the overall TN operation cost and RB energy utilization with respect to the change of power allocation ratio are calculated (Fig. 3.7–Fig. 3.8). In such analysis, all uncertainties of train operation are ignored. The power allocation ratio is fixed with respect to the change of time. From the figure, it can be observed that, generally, with the increase of power allocation ratio, the cost drops at first and then increases gradually. RB energy utilization similarly increases at first and then drops sharply. Besides, with the increase of the charge voltage threshold and the decrease of the discharge voltage threshold, the optimal power allocation ratio is smaller. Moreover, the optimal power allocation ratio is slightly larger than that in the conventional power allocation strategy, which is determined by the maximum battery power divided by the maximum HESS power (in this case, namely, 0.20). This is because the operation cost of the battery is much lower than that of the supercapacitor. Thus, an increased power allocation ratio will lead to a reduction in the operation cost.

Table 3.6 summarizes the optimal power allocation ratio that has the lowest overall operation cost and highest RB energy utilization. From the table, the optimal power allocation ratio varies with different train headways and objectives.



(a) (b) (c)

Fig. 3.7 Impact of HESS power allocation ratio on overall TN operation cost under (a) 350s, (b) 540s, and (c) 660s headway.



(a) (b) (c)

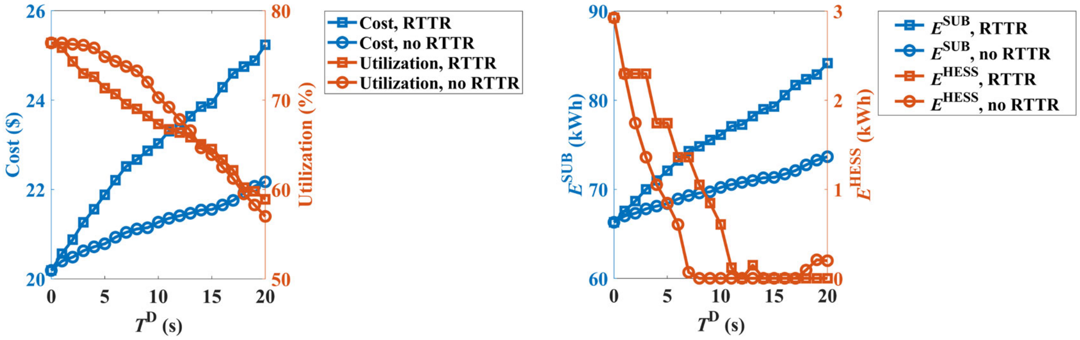
Fig. 3.8 Impact of HESS power allocation ratio on RB energy utilization under (a) 350s, (b) 540s, and (c) 660s headway.

Table 3.6 Optimal power allocation ratio in the fixed operation environment.

Objective	Headway (s)	Optimal η^{PA} (%)
Overall operation cost	350	30.0
	540	32.0
	660	28.5
RB energy utilization	350	27.5
	540	24.5
	660	27.0

3.2.4.3 Impact of Operation Uncertainties

1) Impact of Delays and RTTR: The uncertain delays and corresponding RTTR can influence the spatial-temporal traction load distribution. Generally, when a delay occurs, the railway operator will temporarily shorten the train running time in the next section. In order to analyze the impact of delays and RTTR, the overall TN operation cost and RB energy utilization with respect to the change of delay times are calculated (Fig. 3.9). In such analysis, all uncertainties of train operation except for delays are ignored. The delays are set to occur in the down direction of station RJ. Since delays generally occur during peak hours, only the train operation under 350 s headway is taken into account. The HESS control parameters are set as $\eta_{\rho,t}^{\mathrm{PA}}$=0.20, $U_{\rho,t}^{\mathrm{CH}}$=865 V, and $U_{\rho,t}^{\mathrm{DIS}}$=855 V. From Fig. 3.9(a), with the increase of delay times, the overall TN operation cost is growing while the RB energy utilization is dropping. This is because the overlapping time between accelerating and decelerating trains is reduced, which results in a larger amount of RB energy that is wasted. Meanwhile, part of the RB energy is absorbed by the HESS, as shown by the declining HESS energy output (Fig. 3.9(b)). It is worth noting that, although the total HESS energy output is zero under 15–20 s delays, the HESS power is not zero during train operation. Besides, the cost of implementing RTTR is significantly higher. This is because, in order to shorten the running time for RTTR, the traction energy consumption increases sharply, resulting in a growing traction energy usage.



(a) (b)

Fig. 3.9 Impact of delay times on (a) overall TN operation cost and RB energy utilization and (b) total substation and HESS energies.

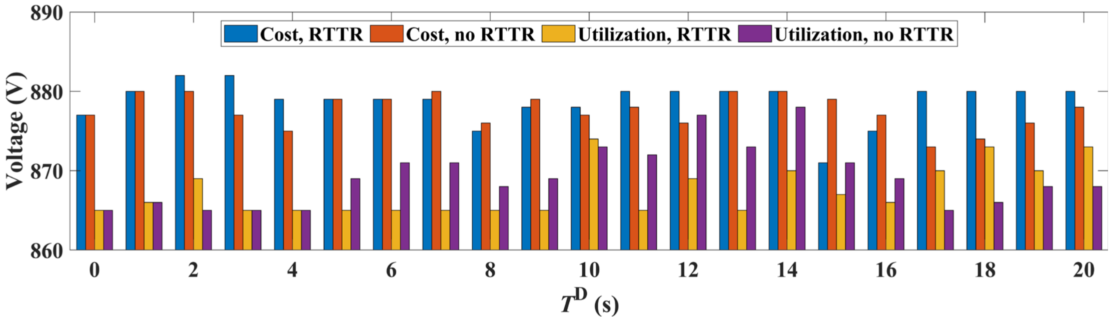


Fig. 3.10 Optimal charge voltage threshold under different delay times.

Then, the impact of delays and RTTR on optimal HESS control parameters is analyzed. As an example, the optimal charge voltage threshold with respect to the change of delay times is calculated and shown in Fig. 3.10. For convenience, other HESS control parameters are fixed as $\eta_{\rho,t}^{\mathrm{PA}}$ =0.20 and $U_{\rho,t}^{\mathrm{DIS}}$ =855 V. It can be observed that the optimal charge voltage threshold lays in the range of 860–890 V. Moreover, the law of optimal charge voltage threshold with the change of delay times is different with and without implementing RTTR. Therefore, the HESS control parameters should adapt to uncertain delays and RTTR.

2) Impact of Passenger Flows: The passenger flows can influence the amount of traction load and RB energy. As an example, the passenger flows in stations RC and TJN are analyzed. The overall TN operation cost and RB energy utilization with respect to the change of passenger flows are calculated (Fig. 3.11–Fig. 3.12).

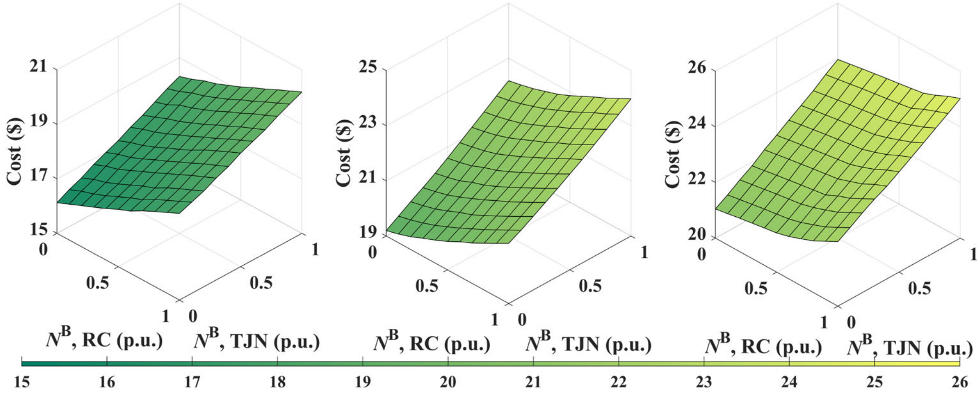


Fig. 3.11 Impact of passenger flows on overall TN operation cost under (a) 350s, (b) 540s, and (c) 660s headway.

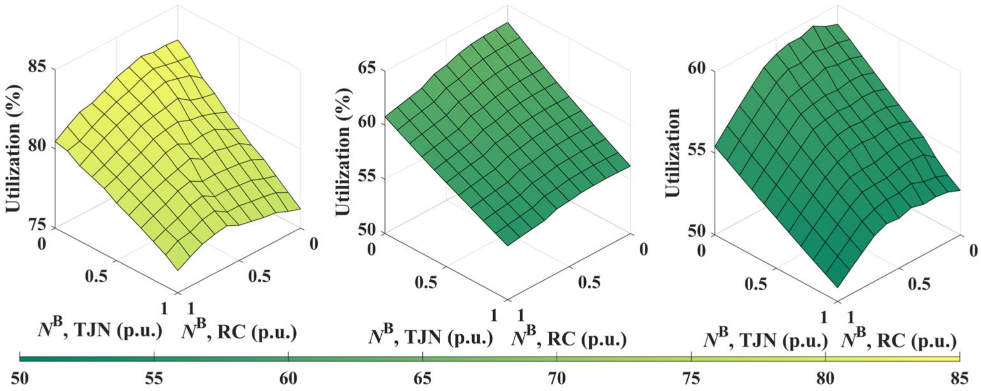


Fig. 3.12 Impact of passenger flows on RB energy utilization under (a) 350s, (b) 540s, and (c) 660s headway.

From the figure, it can be observed that, generally, with the growth of the passengers, the overall TN operation cost increases while the RB energy utilization decreases. In addition, the impact of passenger flows varies from station to station, where station TJN has a more profound influence on the change of cost and utilization. Therefore, it is necessary to model the passenger flows in each station for comprehensive sizing and control decision-making.

3) Impact of Initial SoEs: The initial SoEs of the HESS can also influence the amount of its charge/discharge energy. Due to its large energy capacity, the SoE of the battery only shows minor changes within seconds. Therefore, the subsection mainly focuses on the impact of initial supercapacitor SoE. Similarly, the overall TN operation cost and RB energy utilization with respect to the change of initial supercapacitor SoE are calculated (Fig. 3.13). In such analysis, all uncertainties of train operation are ignored. The HESS control parameters are set as $\eta_{\rho,t}^{\mathrm{PA}}$ =0.20, $U_{\rho,t}^{\mathrm{CH}}$ =865 V, and $U_{\rho,t}^{\mathrm{DIS}}$ =855 V.

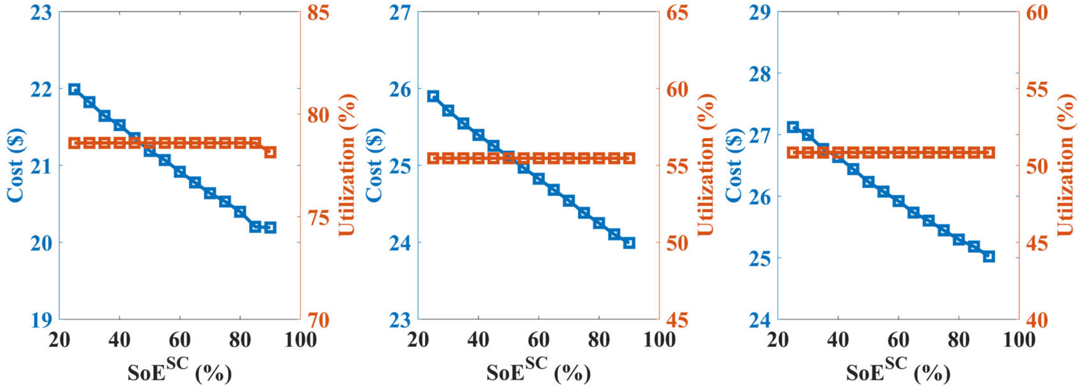


Fig. 3.13 Impact of initial supercapacitor SoE on overall TN operation cost and RB energy utilization under (a) 350s, (b) 540s, and (c) 660s headway.

From the figure, it can be observed that, with the increase of the initial supercapacitor SoE, the overall TN operation cost drops about 8.2%, 7.37%, and 7.74% under 350 s, 540 s, and 660 s headway, respectively. As shown in Fig. 3.14, this is because more HESS energy can be utilized to supply the traction loads, and the energy demands from external grids is reduced, leading to a decreased electricity trading cost. Nevertheless, the RB energy utilization is almost unchanged with the increase of the initial supercapacitor SoE. This indicates that the initial supercapacitor SoE only alters the RB energy distribution within the HESS, accelerating trains, and contact lines while having little impact on the start-up of braking resistors.

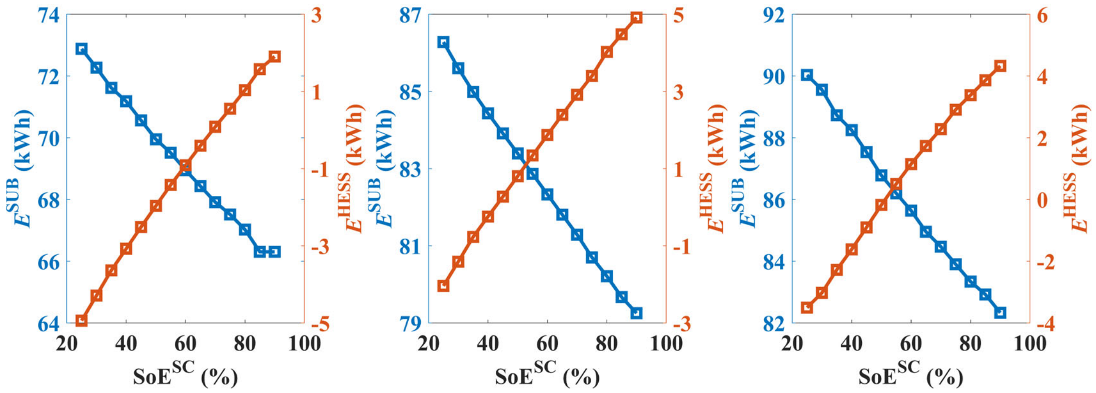


Fig. 3.14 Impact of initial supercapacitor SoE on total substation and HESS energies under (a) 350s, (b) 540s, and (c) 660s headway.

According to the above analysis, the HESS control parameters have a profound impact on the economic operation of TN and the energy utilization of RB. The optimal charge/discharge voltage thresholds and power allocation ratio of the HESS vary with train service patterns, including headways, delays, and RTTR, and their relationship is highly complex and nonlinear. It is necessary to limit the frequency of the start-up of the braking resistor to improve the network-wide cost and energy efficiency based on real-time traction loads and train positions. Meanwhile, the passenger flows and initial HESS SoEs can also substantially decrease the economy of TN operation, which become key factors in HESS sizing and control decision-making. Nevertheless, in the above derivation, the HESS control parameters are fixed with the change of time, and the control parameters that adapt to real-time operation uncertainties are difficult to acquire online under the complex energy interactions between multiple trains. Furthermore, it is necessary to take the sizing of HESS into account since its capacity can limit the reservation of RB energy, which results in energy waste. Therefore, in this chapter, a sizing and control optimization approach for HESS-integrated traction substation operation is developed, which mainly focuses on cost efficiency. The following chapters will extend the research into DHESSs-integrated TN operation and multiple objectives.

3.3 MTRL–SCO Approach

3.3.1 Approach Overview

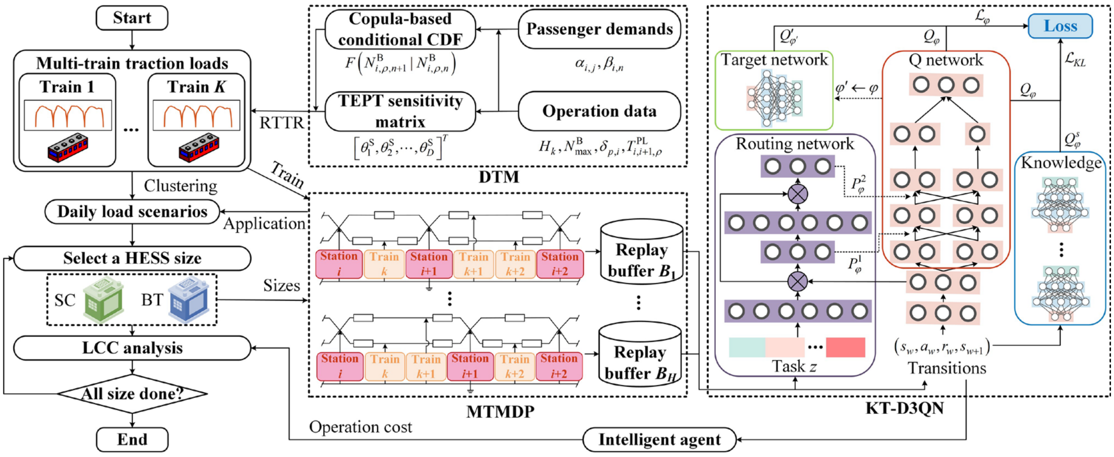


Fig. 3.15 Overview of MTRL–SCO.

The proposed approach (Fig. 3.15) contains the following steps: 1) Various traction load scenarios are randomly generated based on the DTM, and representative daily traction load scenarios are selected based on clustering algorithms. 2) A HESS size is selected from the size constraint set, and such a HESS control problem is reformulated as an MTMDP. 3) The proposed KT-D3QN algorithm solves the MTMDP and trains an intelligent agent for multi-task HESS control. 4) Based on daily service patterns, an

LCC analysis is performed, where the daily operation cost is calculated by the well-trained agent. 5) Repeat 2)–4) to traverse all sizes, and the optimal HESS size and control strategy is determined with the lowest LCC.

3.3.2 Dynamic Traffic Model

3.3.2.1 Copula-Based Passenger Flow Scenario Generation

Since passenger flow fluctuations can result in varying traction loads, the spatial-temporal uncertainty of passenger flows is quantified based on the Copula theory (Algorithm 3.1). First, we estimate the historical passenger data $N_{i,\rho,n}^{\mathrm{B}}$ according to OD and arrival rate tables, where $N_{i,\rho,n}^{\mathrm{B}}$ is the average onboard passengers between station *i* and *i*+1 at time step *n*. The calculation of $N_{i,\rho,n}^{\mathrm{B}}$ is illustrated in Appendix B. Then, for simplicity, we only consider the temporal correlation between two consecutive time steps, namely,

$$f\left(N_{i,\rho,n+1}^{\mathrm{B}} \mid N_{i,\rho,1}^{\mathrm{B}},\cdots,N_{i,\rho,n}^{\mathrm{B}}\right)=f\left(N_{i,\rho,n+1}^{\mathrm{B}} \mid N_{i,\rho,n}^{\mathrm{B}}\right), \tag{3.22}$$

where $f(\cdot)$ is the probability density function (PDF), $N_{i,\rho,n}^{\mathrm{B}}$ is the average onboard passengers at station *i* at prediction interval *n*.

Then, the conditional PDF in (3.22) can be obtained by (3.23)–(3.24), where the joint PDF is written by a Copula function.

$$\frac{f\left(N_{i,\rho,n+1}^{\mathrm{B}} \mid N_{i,\rho,n}^{\mathrm{B}}\right)}{f\left(N_{i,\rho,n+1}^{\mathrm{B}}\right)}=c\left(F_1,F_2\right)=\frac{\partial^2 C\left(F_1,F_2\right)}{\partial F_1 \partial F_2}, \tag{3.23}$$

$$F_1=F\left(N_{i,\rho,n+1}^{\mathrm{B}}\right),\ F_2=F\left(N_{i,\rho,n}^{\mathrm{B}}\right), \tag{3.24}$$

where $C(\cdot)$ and $c(\cdot)$ are the Copula function and its PDF, respectively, $F(\cdot)$ is the cumulative distribution function (CDF). Thus, multiple pseudo-observations can be drawn from the conditional CDF to generate sufficient scenarios.

3.3.2.2 TEPT-Based RTTR & Trajectory Selection Optimization

Different from the conventional method illustrated in subsection 3.2.4.3, a novel TEPT-based RTTR algorithm is proposed and adopted for further cost saving. Rather than focusing on delay time minimization, the TEPT-based RTTR aims to minimize traction energy consumption by shortening the running time of each section based on the energy sensitivity between each train speed profile. Thus, the objective is

$$\begin{aligned}
&\min \sum_{i=1}^{I-1} E_{d,\delta_{p,i}} \\
&s.t.\ \sum_{i=1}^{I-1} T_{i,i+1}^{\mathrm{RTTR}} = \sum_{i=1}^{I-1} T_{i,i+1}^{\mathrm{PL}}, \\
&d \in \{1,2,\cdots,D\}, i \neq I, p \neq P.
\end{aligned} \tag{3.25}$$

It is worth noting that since these delay times are short and do not cause any interruption of train services, the total running and dwell times are assumed unchanged [144]. Specifically, apart from passenger flows, the traction load changes in accordance with the running time. By dividing $N_{i,\rho,n}^{\mathrm{B}}$ into D intervals, the TEPT sensitivity θ^{S} can be written as

$$\begin{aligned}
&\theta^{\mathrm{S}} = \begin{bmatrix} \theta_1^{\mathrm{S}} \\ \theta_2^{\mathrm{S}} \\ \vdots \\ \theta_D^{\mathrm{S}} \end{bmatrix}, \quad \theta_d^{\mathrm{S}} = \begin{bmatrix} \theta_{d,\delta_{1,1}}^{\mathrm{S}} & \theta_{d,\delta_{1,2}}^{\mathrm{S}} & \cdots & \theta_{d,\delta_{1,I-1}}^{\mathrm{S}} \\ \theta_{d,\delta_{2,1}}^{\mathrm{S}} & \theta_{d,\delta_{2,2}}^{\mathrm{S}} & \cdots & \theta_{d,\delta_{2,I-1}}^{\mathrm{S}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{d,\delta_{P-1,1}}^{\mathrm{S}} & \theta_{d,\delta_{P-1,2}}^{\mathrm{S}} & \cdots & \theta_{d,\delta_{P-1,I-1}}^{\mathrm{S}} \end{bmatrix}, \\
&\theta_{d,\delta_{p,i}}^{\mathrm{S}} = \frac{\Delta E_{d,\delta_{p,i}}}{\Delta T}, \quad d = [1,2,\cdots,D], i \neq I, p \neq P,
\end{aligned} \tag{3.26}$$

where each section has P pre-programmed trajectories, $1 \le p \le P$, and the trajectories are ranked from the highest traction energy consumption to the lowest. $\delta_{p,i}$ denotes the pth trajectory of section $(i,i+1)$. $\theta_{d,\delta_{p,i}}^{\mathrm{S}}$ and $\Delta E_{d,\delta_{p,i}}$ are the sensitivity and energy difference between the pth and p+1th trajectory of section $(i,i+1)$ with passenger interval *d*. ΔT is the increment of train running time.

A TEPT-based RTTR algorithm can then be developed to obtain the rescheduled timetable and trajectories after the delay (Algorithm 3.1). The principle is that, at each iteration, we allocate a ΔT to the section with the lowest sensitivity. Then, the trajectory $\delta_{p,i}$ and sensitivity $\theta^{\mathrm{S}}_{d,\delta_{p,i}}$ for that section are updated, while $T^{\mathrm{D}}_{i,\rho} \leftarrow T^{\mathrm{D}}_{i,\rho} - \Delta T$. The allocation process ends when $T^{\mathrm{D}}_{i,\rho} = 0$. Thus, the rescheduled timetable and trajectories of each section are determined. The detailed steps are illustrated as follows.

Step 1: Initialize the delay scenario. Input the delayed station and the corresponding delay time $T^{\mathrm{D}}_{i,\rho}$. Set pre-programmed train trajectory set $\delta_{1,1},\cdots,\delta_{P-1,I-1}$ for each section, and generate the initial trajectory for each section according to the timetable $T^{\mathrm{PL}}_{1,2,\rho},\cdots,T^{\mathrm{PL}}_{I-1,I,\rho}$. Set the passengers $N^{\mathrm{B}}_{1,\rho,n},\cdots,N^{\mathrm{B}}_{I-1,\rho,n}$ at time interval n for each station. Calculate the traction energy consumption $E_{d,\delta_{p,i}}$ for each trajectory under each passenger flow, and generate the TEPT matrix θ^{S}.

Step 2: Initialize rescheduling. Following the conventional method, allocate $T^{\mathrm{D}}_{i,\rho}$ to the first section $(i,i+1)$ after the delayed station. If $T^{\mathrm{D}}_{i,\rho} > (p-1)\Delta T$, $T^{\mathrm{D}}_{i,\rho} \leftarrow T^{\mathrm{D}}_{i,\rho} - (p-1)\Delta T$, and then allocate $T^{\mathrm{D}}_{i,\rho}$ to section $(i+1,i+2)$. Repeat the process till $T^{\mathrm{D}}_{i,\rho} = 0$. Calculate the total energy consumption $E_0 = \sum_{i=1}^{I} E_{d,\delta_{p,i}}$.

Step 3: Select one row $\left[\theta^{\mathrm{S}}_{d,\delta_{p,1}},\cdots,\theta^{\mathrm{S}}_{d,\delta_{p,I-1}}\right]$ in matrix θ^{S} by current passenger interval d of each station and trajectories $\delta_{p,1},\cdots,\delta_{p,I-1}$ of each section.

Algorithm 3.1 Passenger Flow Scenario Generation and TEPT-Based RTTR in DTM

```
# Passenger flow scenario generation
Input: OD and arrival rate $\alpha$, $\beta$
For prediction interval $n = 1, N$ do
    Update arrival rate $\beta_{i,n}$ and operation data $H_k$, $K_n$
    Estimate historical passengers $N^{\mathrm{B}}_{i,\rho,n}$
Establish the joint conditional CDF by conditional Copula functions using (3.23)–(3.24), draw pseudo-observations from it to generate sufficient scenarios
Output: onboard passenger $N^{\mathrm{B}}$
# TEPT-based RTTR
Input: passenger $N^{\mathrm{B}}$, speed profile $\delta$, delay time $T^{\mathrm{D}}$, and running time $T^{\mathrm{PL}}$
Initialize TEPT matrix $\theta^{\mathrm{S}}$
Initialize rescheduling: allocate $T^{\mathrm{D}}_{i,\rho}$ to section $(i,i+1)$, $T^{\mathrm{D}}_{i,\rho} \leftarrow T^{\mathrm{D}}_{i,\rho} - (p-1)\Delta T$.
Then allocate $T^{\mathrm{D}}_{i,\rho}$ to section $(i+1,i+2)$. Repeat the process till $T^{\mathrm{D}}_{i,\rho} = 0$.
Calculate the total energy consumption $E_0 = \sum_{i=1}^{I} E_{d,\delta_{p,i}}$
While $T^{\mathrm{D}}_{i,\rho} \neq 0$ do
    Select $\left[\theta^{\mathrm{S}}_{d,\delta_{p,1}},\cdots,\theta^{\mathrm{S}}_{d,\delta_{p,I-1}}\right]$ by current interval $d$ and trajectory $\delta_{p,1},\cdots,\delta_{p,I-1}$
    Allocate $\Delta T$ to the target section with $\arg\min\left(\theta^{\mathrm{S}}_{d,\delta_{p,1}},\cdots,\theta^{\mathrm{S}}_{d,\delta_{p,I-1}}\right)$, update the trajectory and sensitivity of the target section
    $T^{\mathrm{D}}_{i,\rho} \leftarrow T^{\mathrm{D}}_{i,\rho} - \Delta T$
With the updated trajectories $\delta'_{p,1},\cdots,\delta'_{p,I-1}$, calculate $E'_0 = \sum_{i=1}^{I} E_{d,\delta'_{p,i}}$ and compare with $E_0$, select the timetable with lowest energy consumption
Output: rescheduled time $T^{\mathrm{RTTR}}$
```

Step 4: Allocate ΔT to the target section with the minimum energy sensitivity

$\arg\min\left(\theta_{d,\delta_{p,1}}^{\mathrm{S}},\cdots,\theta_{d,\delta_{p,I-1}}^{\mathrm{S}}\right)$, update the trajectory and sensitivity of the target section. Update the delay time by $T_{i,\rho}^{\mathrm{D}} \leftarrow T_{i,\rho}^{\mathrm{D}} - \Delta T$.

Step 5: Repeat step 3–4 until $T_{i,\rho}^{\mathrm{D}} = 0$. With the updated trajectories $\delta'_{p,1},\cdots,\delta'_{p,I-1}$, calculate the total energy consumption $E'_0 = \sum_{i=1}^{I} E_{d,\delta'_{p,i}}$. Compare E'_0 with E_0, select the timetable and trajectories with the lowest energy consumption.

3.3.2.3 Traction Load Calculation

The power of train *k* is calculated by (3.27),

$$P_{k,\rho,t}^{\mathrm{TR}} = \left[\left(M^{\mathrm{TR}} + N_{i,\rho,n}^{\mathrm{B}} M^{\mathrm{P}}\right) a_{k,\rho,t}^{\mathrm{TR}} - \mathcal{N}\left(F_{k,\rho,t}^{\mathrm{TR}}\right)\right] \frac{v_{k,\rho,t}^{\mathrm{TR}}}{\eta^{\mathrm{TR}}}, \tag{3.27}$$

where M^{TR} and M^{P} are the total vehicle mass and passenger mass per person, respectively, $v_{k,\rho,t}^{\mathrm{TR}}$ and $a_{k,\rho,t}^{\mathrm{TR}}$ are the train speed and acceleration, respectively, η^{TR} is the motor efficiency. Since total resistance $F_{k,\rho,t}^{\mathrm{TR}}$ can be uncertain due to weather and line conditions, it is subject to a truncated Normal distribution $\mathcal{N}(\cdot)$ [170], where the standard deviation is 5% of the mean value, and its variance is limited to 10%.

3.3.3 Multi-Task Markov Decision Process

3.3.3.1 Task Representation & MTMDP

Since different headways and rescheduled trajectories significantly change the spatial-temporal distribution of traction loads, each headway and each combination of trajectories between different stations is a specific task. The task set is $\mathcal{Z} = \{\mathcal{Z}_H, \mathcal{Z}_P\}$, where $\mathcal{Z}_H = \{z_1,\cdots,z_H\}$ contains H headway tasks in one-hot vectors, and $\mathcal{Z}_P = \left\{\{\delta_{1,1},\cdots,\delta_{1,I-1}\},\cdots,\{\delta_{P-1,1},\cdots,\delta_{P-1,I-1}\}\right\}$ contains $(P-1)(I-1)$ combination tasks of trajectories. Hence, the total number of tasks is $H(P-1)(I-1)$. Each task $z \in \mathcal{Z}$ can be formulated as a MDP introduced in subsection 1.3.1.1, and multiple tasks form a MTMDP with components $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma, \mathcal{Z} \rangle$.

3.3.3.2 State, Action, & Reward

State s_t contains two parts: **a) local traction substation operation status**, including supercapacitor SoE $\mathrm{SoE}_{\rho,t}^{\mathrm{SC}}$, battery SoE $\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}$, local traction substation outputs $U_{i,\rho,t}^{\mathrm{SUB}}$ and $I_{i,\rho,t}^{\mathrm{SUB}}$ (suppose HESS is in station *i*), **b) train operation status**, including position $x_{1,\rho,t}^{\mathrm{TR}},\cdots,x_{K,\rho,t}^{\mathrm{TR}}$, direction $d_{1,\rho,t}^{\mathrm{TR}},\cdots,d_{K,\rho,t}^{\mathrm{TR}}$, and power $P_{1,\rho,t}^{\mathrm{TR}},\cdots,P_{K,\rho,t}^{\mathrm{TR}}$. By this design, it is not necessary for the agent to exchange information with other traction substations for decision making, which reduces the communication burden.

Action a_t is the voltage thresholds $U_{\rho,t}^{\mathrm{CH}}$, $U_{\rho,t}^{\mathrm{DIS}}$ and power allocation $\eta_{\rho,t}^{\mathrm{PA}}$. $a_t = \left\{ U_{\rho,t}^{\mathrm{CH}}, U_{\rho,t}^{\mathrm{DIS}}, \eta_{\rho,t}^{\mathrm{PA}} \right\}$. According to (3.17), reward r_t is the minus of $J_{\rho,t}^{\mathrm{GRID}}$ and $J_{\rho,t}^{\mathrm{OM}}$, namely, $r_t = -\left(J_{\rho,t}^{\mathrm{GRID}} + J_{\rho,t}^{\mathrm{OM}} \right)$.

$$s_t = \begin{Bmatrix} \mathrm{SoE}_{\rho,t}^{\mathrm{SC}}, \mathrm{SoE}_{\rho,t}^{\mathrm{BT}}, U_{i,\rho,t}^{\mathrm{SUB}}, I_{i,\rho,t}^{\mathrm{SUB}}, x_{1,\rho,t}^{\mathrm{TR}}, \cdots, x_{K,\rho,t}^{\mathrm{TR}}, \\ d_{1,\rho,t}^{\mathrm{TR}}, \cdots, d_{K,\rho,t}^{\mathrm{TR}}, P_{1,\rho,t}^{\mathrm{TR}}, \cdots, P_{K,\rho,t}^{\mathrm{TR}}. \end{Bmatrix} \tag{3.28}$$

3.3.3.3 State Transition

The train operation status is updated by the selected trajectory p, and the network parameters, such as the contact line resistance, are updated accordingly. Then, the local traction substation outputs (e.g., traction substation voltage and current) are updated by power flow calculation according to a_t. According to the power flows, the SoEs of the HESS are updated by

$$\mathrm{SoE}_{\rho,t}^{\mathrm{SC}}=\left(\frac{U_{\rho,t}^{\mathrm{C}}}{U^{\mathrm{C,norm}}}\right)^2,\quad \mathrm{SoE}_{\rho,t}^{\mathrm{BT}}=\mathrm{SoE}_{\rho,t-1}^{\mathrm{BT}}-\frac{I_{\rho,t}^{\mathrm{BT}}\Delta t}{Q^{\mathrm{BT,norm}}},\tag{3.29}$$

where $U^{\mathrm{C,norm}}$ and $Q^{\mathrm{BT,norm}}$ are the nominal capacitor voltage and nominal battery capacity, respectively.

Furthermore, there are multiple states that are independent of a_t and have intrinsic uncertainties, such as passenger flows and delays. These uncertainties are updated by the generated scenario parameters in subsection 3.3.2.

3.3.4 KT-D3QN Algorithm Implementation

3.3.4.1 Dueling Double Deep Q Network (D3QN)

Since the complexity of calculating expected return Q^*, D3QN [45] approximates Q^* by Q_φ with parameter φ, and Q_φ is decoupled with a value estimation $VE(s_t)$ and an action advantage estimation $AE(s_t,a_t)$. This dueling architecture enables the agent to learn independent state values, which is useful in states where the actions have no effect on the environment,

$$Q_\varphi(s_t,a_t)=VE(s_t)+AE(s_t,a_t)-\frac{\sum_{a_{t+1}\in\mathcal{A}}AE(s_t,a_{t+1})}{|\mathcal{A}|}.\tag{3.30}$$

Totally H replay buffers are built for all headway tasks, and W transitions (s_w,a_w,r_w,s_{w+1}) are randomly sampled from each buffer for updating the multi-task D3QN. Thus, the original loss function of the D3QN in (2.31) is extended to a multi-task form, namely,

$$\mathcal{L}_\varphi=\frac{1}{HW}\sum_h\sum_w\left(y-Q_\varphi(s_w,a_w)\right)^2,\tag{3.31}$$

where $y=r_w+\gamma Q'_{\varphi'}\left(s_{w+1},\arg\max_{a_{w+1}}Q_\varphi(s_{w+1},a_{w+1})\right)$, $Q'_{\varphi'}$ and φ' are the target network and its parameter, respectively.

The loss updates can be written as

$$\varphi \leftarrow \varphi + \xi_q \nabla_{\varphi} \mathcal{L}_{\varphi}, \tag{3.32}$$

where ξ_q is the learning rate.

3.3.4.2 Knowledge Transfer & Policy Distillation Annealing

Considering the similarity of different trajectory tasks under a given headway, we develop a knowledge transfer method to rapidly and stably learn the multi-task policy incorporating common knowledge from task-specific agents by policy distillation. For each headway task with several sets of trajectory tasks, a single-task agent is first trained in a learning environment without delay. Then, the Kullback-Leibler divergence is adopted to measure the discrepancy between the policy distributions of single-task agents and the multi-task agent. An annealing strategy is utilized to gradually reduce the knowledge transfer for convergence.

$$\mathcal{L}_{KL} = \sum_{h}\sum_{w} \operatorname{softmax}\left(Q_{\varphi}^{s}\left(s_w, a_w\right)\right) \ln\left(\frac{\operatorname{softmax}\left(Q_{\varphi}^{s}\left(s_w, a_w\right)\right)}{\operatorname{softmax}\left(Q_{\varphi}\left(s_w, a_w\right)\right)}\right), \tag{3.33}$$

$$\mathcal{L} = \left(1-\lambda\right)\mathcal{L}_{\varphi} + \lambda \mathcal{L}_{KL}, \tag{3.34}$$

where Q_{φ}^{s} is the Q value of the corresponding single-task agent, λ decreases during training.

3.3.4.3 Soft Modulation with Conflict Gradient Projecting

As the difficulty of learning different tasks varies, soft modulation [69] (Fig. 3.15) is introduced to address this issue. The idea of soft modulation is to generate soft combinations between different neural network modules without explicitly specifying the policy structure for each task. To implement soft modulation, the network structure of D3QN is divided into multiple layers, and each layer contains a set of modules. A separate routing network with parameter φ_r is built to estimate the connection probability P_{φ}^{l} between modules in layer l and layer $l+1$ according to the task and current state. Hence, for different tasks, as the connection probability varies, each task will have different weighted combinations of shared network modules to construct its task-conditioned policy. This reconfiguration of the *Q* network improves flexibility in handling various tasks and ensures the quality of solutions. Moreover, in order to mitigate potential inter-task conflicts, the conflict gradient projecting technique [71] is adopted. It provides a simple solution to deal with task gradient interference in network updates by projecting the conflicting gradient onto the normal plane of the other. Thus, by implementing the above techniques, the multi-task learning performance of the multi-task D3QN can be enhanced. The training step of the algorithm is summarized in Algorithm 3.2.

Algorithm 3.2 KT-D3QN

```
Initialize Q network with $\varphi$, routing network with $\varphi_r$, target network with $\varphi' \leftarrow \varphi$, and replay buffers $B_1,\cdots,B_H$
For episode = 1, Max do
    Sample a task from the task set to initialize learning environment
    Receive the initial state $s_0$
    For control interval t = 1, T do
        Select $a_t$ with $\mu(a_t \mid s_t, z)$ and $\varepsilon$-greedy, obtain $r_t$ and $s_{t+1}$ in the environment
        Store transition $(s_t, a_t, r_t, s_{t+1})$ to $B_h$ based on headway task $z_h$
        Sample $W$ transitions $(s_w, a_w, r_w, s_{w+1})$ from each $B_h$
        Calculate loss $\mathcal{L}$ by (3.31) and (3.33)–(3.34), update $\varphi, \varphi_r, \lambda$
        Soft update: $\varphi' \leftarrow \tau\varphi + (1-\tau)\varphi'$
```

3.4 Case Study

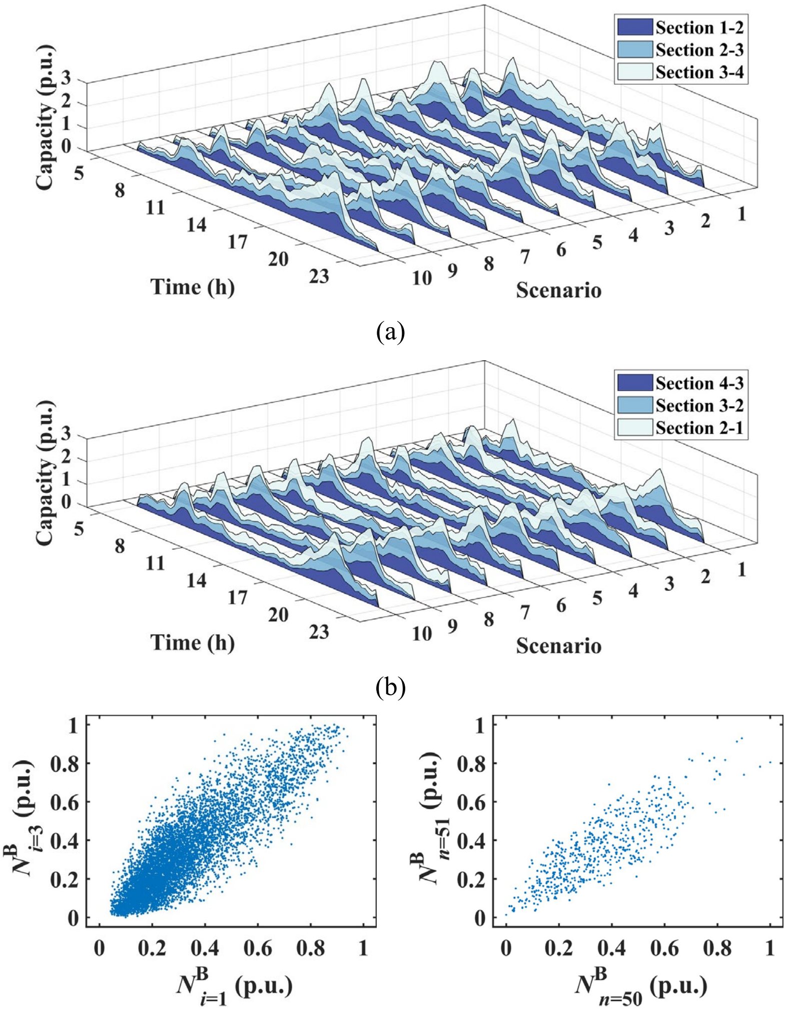
In this section, an analysis of the aforementioned formulations and algorithms is conducted. First, the optimal HESS control behaviors are investigated. On the one hand, the impact of different control schemes is demonstrated to verify the effectiveness of joint adjustments on voltage thresholds and power allocations of the HESS. On the other hand, the training and test performance of KT-D3QN on overall operation cost and RB energy utilization is examined by comparing it with other learning-based and non-learning-based algorithms. Furthermore, the optimal HESS configurations are demonstrated with an analysis of various traffic models and RTTR algorithms.

Table 3.7 Rescheduling settings.

Section	Direction	$T_{i,j,\rho}^{\text{RTTR}}$ (s)	Length (m)	Dwell time (s)
RJ–RC (1–2)	Down	[94, 104]	1354	30
RC–TJN (2–3)		[155, 165]	2337	
TJN–JH (3–4)		[141, 151]	2265	
JH–TJN (4–3)	Up	[141, 151]	2265	
TJN–RC (3–2)		[152, 162]	2337	
RC–RJ (2–1)		[95, 105]	1354	

3.4.1 Setup

The subway line data and HESS data used in case studies are illustrated in subsection 3.2.4.1. The rescheduling settings of these stations are shown in Table 3.7. Considering that batteries are not suitable for covering the large traction load power [121], their rated power is roughly taken as the average traction substation power during one train headway. To meet the peak traction power demand [171], the power difference between the average and peak traction substation power is roughly treated as the rated supercapacitor power. Thus, the optimal HESS size is searched in a range near the above empirical size setting, where the number of supercapacitors in parallel ranges from 15 to 25, and the number of batteries in parallel ranges from 5 to 10.



(a)

(b)

(c)

(d)

Fig. 3.16 Passenger flows in (a) down and (b) up directions. Time intervals 50–51 and sections RJ–RC with TJN–JH are used to show (c) spatial and (d) temporal passenger flow correlations.

For simplicity, the delays are only considered during peak hours (namely, 350 s headway), and only one delay occurs at a random station in each scenario. $T^{\mathrm{D}}_{i,\rho}$ is set between 5–20 s [32] using the log-Normal distribution [157] where the mean and variance are both 5 s. Hence, according to the delay time range, each station has 15 combination tasks of train trajectories with different planned running times. The trajectories are generated by the proposed SRL–EETTO in chapter 2, which achieves traction energy consumption minimization and meets multiple objectives of punctuality, safety, and ride comfort. $\Delta T = 1$ s. For multi-task control, the proposed algorithm is trained and tested by 1000 and 21 random traction load scenarios, respectively. The initial HESS SoEs are randomly generated. For sizing optimization, 122 random traction load scenarios are included in a daily operation scenario according to the service pattern, and 1000 such daily operation scenarios are generated. Then, to decrease the computational cost, 10 representative scenarios are retained by K-means clustering. The scenario probabilities are 0.112, 0.137, 0.126, 0.101, 0.096, 0.106, 0.077, 0.073, 0.100, and 0.072. The passenger flow fluctuations and their correlations in the representative scenarios are shown in Fig. 3.16, where historical OD and arrival rate tables are obtained from [156].

Table 3.8 KT-D3QN parameters.

Parameter	Value	Parameter	Value	Parameter	Value
ξ_q	10^{-4}	W	43	Optimizer	Adam
τ	5×10^{-4}	λ	1→0.05	Buffer capacity	2^{20}
γ	0.998	H	3	Exploration policy	0.5→0.01

The parameters of the KT-dD3QN are listed in Table 3.8. The Q network has two fully connected layers both with 128 units and ReLU non-linearity, and followed by 2 layers and 2 modules per layer. Each module has two hidden layers with 128 units and ReLU non-linearity. The routing network outputs 128 representations for connection probability per layer. The target network and task-specific agents have the same structure as the *Q* network. The exploration rate reduces linearly from 0.5 to 0.01 and remains constant at 0.01 after 2000 episodes. λ reduces linearly from 1 to 0.05 and remains constant at 0.05 after 2000 episodes. All simulations are performed by PyTorch 1.12.1 and Python 3.9.13 on the same device in subsection 2.4.1.

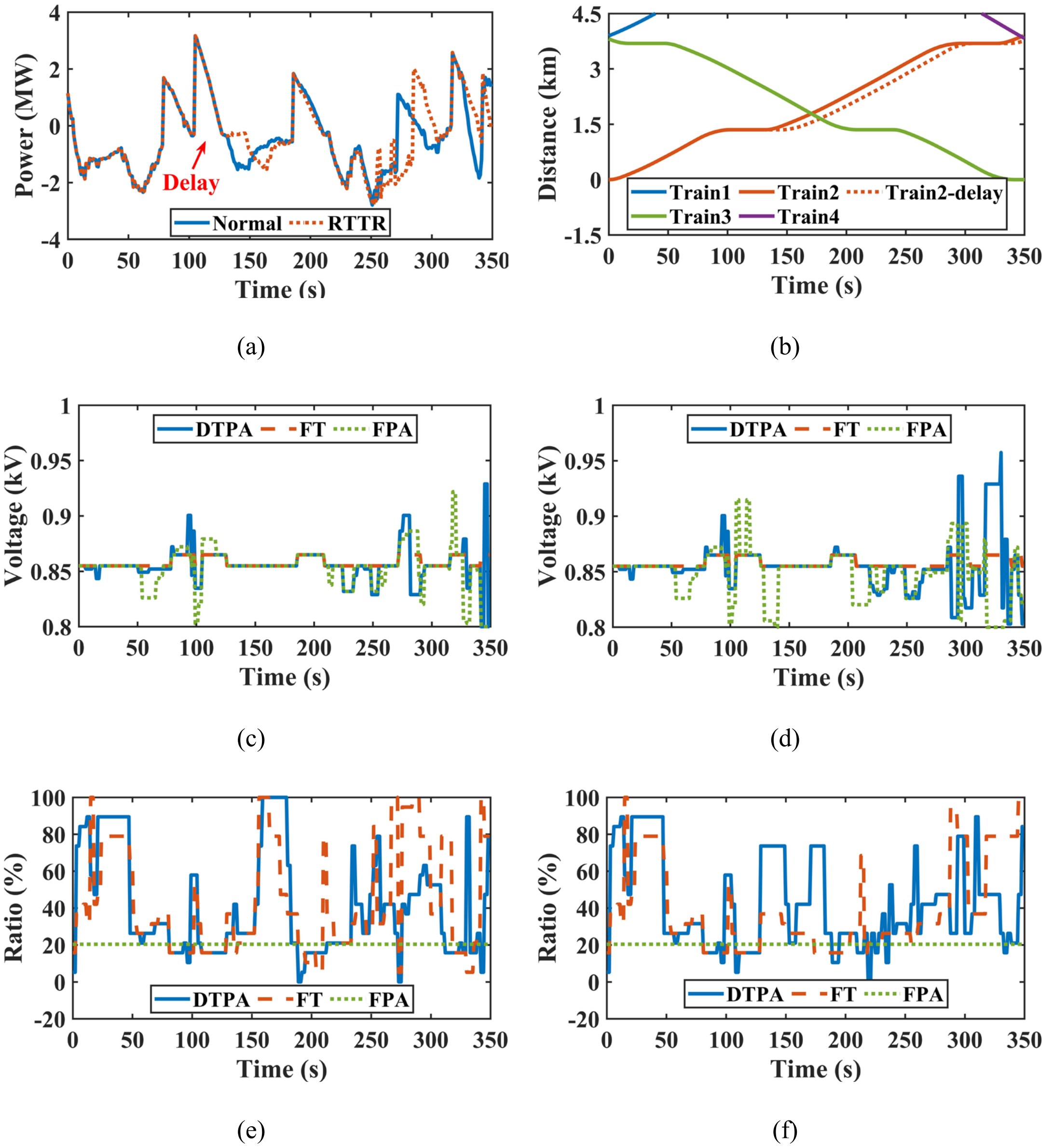
3.4.2 Analysis of Hybrid Energy Storage System Control Behaviors

3.4.2.1 Control Behaviors & Impact of Control Schemes

In this subsection, the impact of different control schemes on the overall operation cost and HESS control behavior are analyzed by the test set. The parallel numbers of 15 and 5 for supercapacitor and battery are taken, respectively. The following schemes are compared: *1) Dynamic threshold and power allocation (DTPA, proposed):* Both thresholds and the power allocation of the HESS can be dynamically adjusted. *2) Fixed threshold (FT):* This scheme aims to verify the effectiveness of threshold adjustments. $U_{\rho,t}^{\mathrm{CH}}$ =865 V, $U_{\rho,t}^{\mathrm{DIS}}$ =855 V. *3) Fixed power allocation (FPA):* This scheme aims to verify the effectiveness of power allocation adjustments. The fixed power allocation is the nominal battery power divided by nominal supercapacitor power. *4) Fixed threshold and power allocation (FTPA):* A conventional rule-based scheme [37], where thresholds and power allocation are fixed.

Table 3.9 Comparative costs of schemes 1–4.

Performance	DTPA	FT	FPA	FTPA
Overall operation cost ($)	351.57	358.48	368.58	373.56
Electricity trading cost ($)	337.53	344.94	353.95	356.92
HESS operation cost ($)	14.04	13.54	14.63	16.63



(a)

(b)

(c)

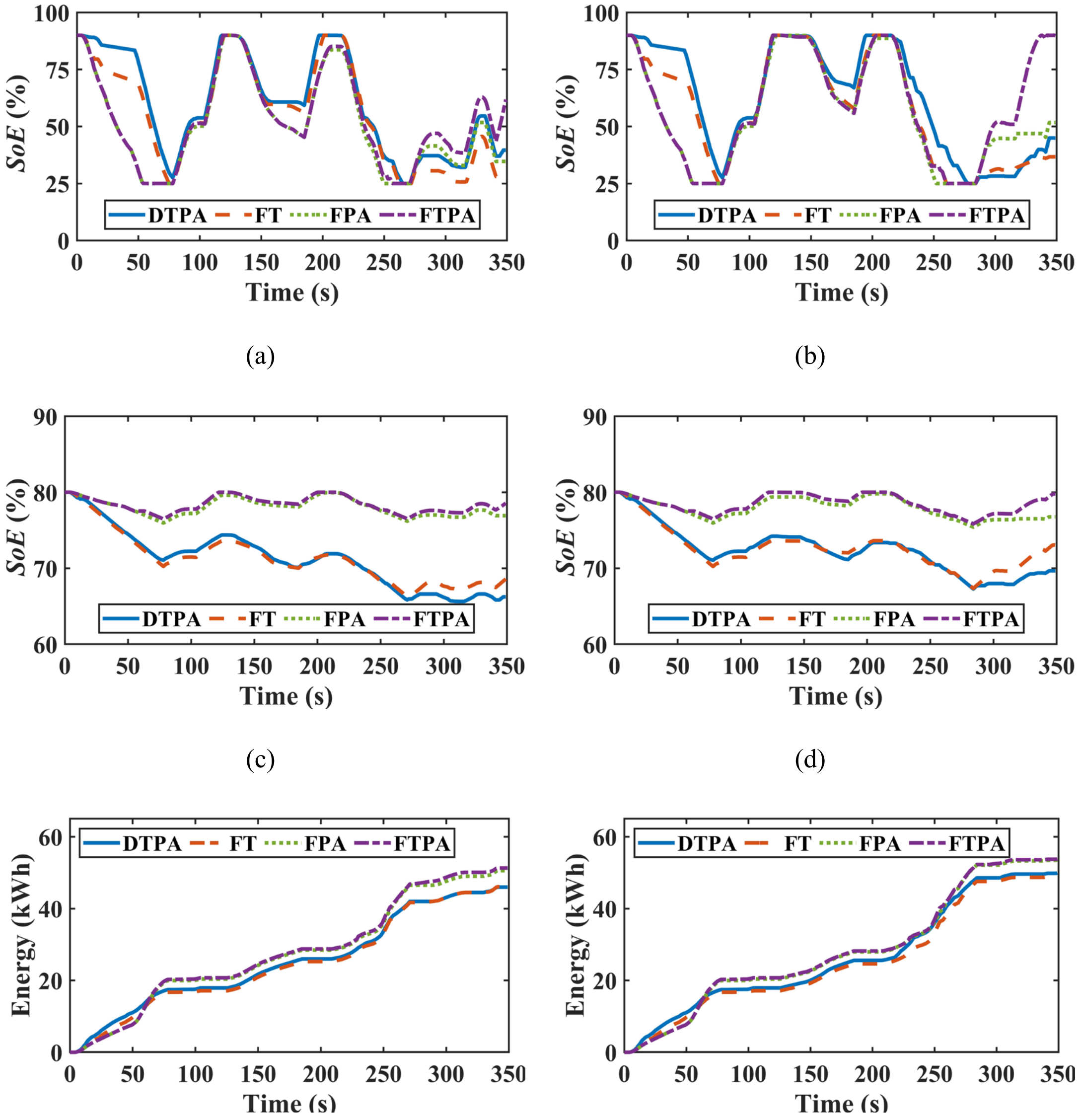
(d)

(e)

(f)

Fig. 3.17 Train operation and HESS control curves, including curves of (a) total train power, (b) train displacement, HESS voltage thresholds under (c) normal operation and (d) RTTR, and HESS power allocations under (e) normal operation and (f) RTTR.

Fig. 3.17(a) shows the total train power generation (+) and consumption (-). Fig. 3.17(b) shows the train displacement curves. When a delay occurs, the rescheduled energy-efficient speed profile prefers higher deceleration and braking power to avoid extra traction energy consumption due to the decreased running time. Hence, the overall power generation under RTTR is higher than that of normal operation. Fig. 3.17(c)–(f) show the HESS SoEs, control parameters, and traction substation energies under normal operation and RTTR, respectively. From Fig. 3.17(c) and (e), since the battery operation cost is lower than that of the supercapacitor, DTPA and FT utilize more battery capacity for cost-saving by leveraging a higher power allocation ratio than FPA.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 3.18 HESS SoEs and traction substation energy curves, including curves of supercapacitor SoEs under (a) normal operation and (b) RTTR, battery SoEs under (c) normal operation and (d) RTTR, and traction substation energy under (e) normal operation and (f) RTTR.

Besides, from Fig. 3.18(a) and (c), the DTPA and FT release less energy than FPA and FTPA during 50–75 s, which prevents the supercapacitor SoE from reaching its lowest limit. The continuous supercapacitor power supply of DTPA and FT decreases the traction substation energy consumption, as shown in Fig. 3.18(e). Moreover, compared with FT, other schemes maintain a reasonable supercapacitor SoE during 275–350 s, which can potentially utilize more supercapacitor energy for further usage. From Fig. 3.17(d) and (f), compared with normal operation, the power allocation ratio of DTPA and FT is closer to FPA. This is to fully utilize the available HESS power to absorb the higher braking power under RTTR. From Fig. 3.18(b)–(d), all schemes show similar performance in maintaining supercapacitor SoE as in normal operation. From Table 3.9, DTPA outperforms other schemes in decreasing the overall operation cost under normal operation and RTTR. The cost reduction is 1.93–5.89% on average.

To further demonstrate how online trip time adjustment affects the proposed approach, another RTTR scenario is analyzed (Fig. 3.19), where a train is informed to increase the trip time by 10 s after running 20 s at section RJ–RC. From the figures, the train speed at section RJ–RC is decreased after receiving the information to extend the trip time, and the RB power is reduced. If the control parameters of the HESS are unchanged, the direct RB power use by trains is reduced and more power is required from the substation to satisfy the traction demand, which increases the energy consumption and operation cost. Nevertheless, under the proposed approach, the HESS voltage thresholds and power allocation ratio increase, indicating less RB power is absorbed by the HESS and the direct RB power use by trains can be maintained. Thus, the proposed approach can effectively address the above energy and cost efficiency issue, validating its effectiveness under online trip time adjustment.

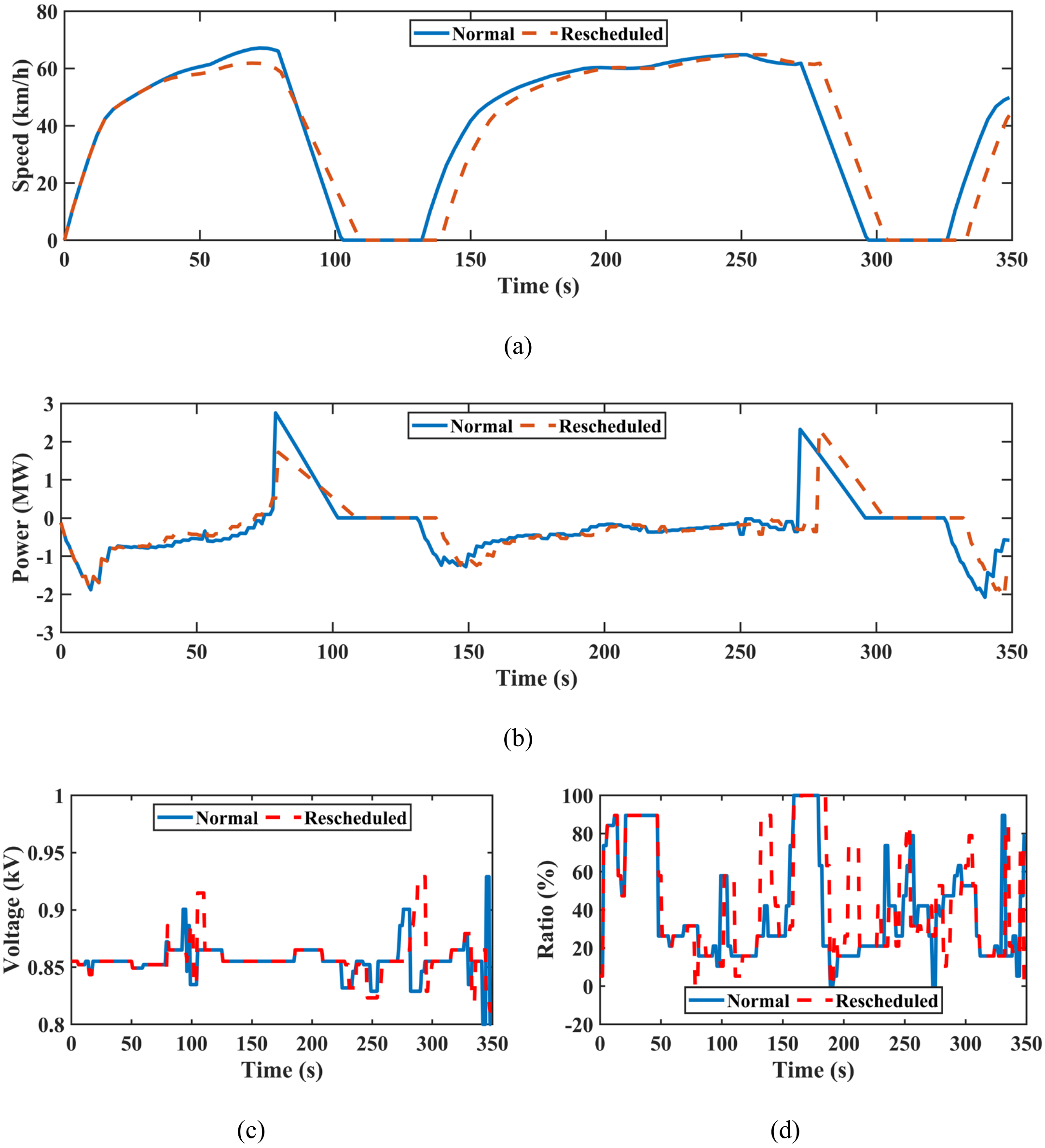


Fig. 3.19 HESS control behaviors under online trip time adjustment, including curves of (a) train trajectory adjustment, (b) train power, (c) HESS voltage thresholds, and (d) HESS power allocations.

3.4.2.2 Impact of Control Optimization Algorithms

In this subsection, the model performance of the proposed KT-D3QN is verified. Four learning-based and two non-learning-based algorithms are compared: *1) KT-D3QN:* proposed. *2) MT-D3QN*: the task set and the routing network are the same as KT-D3QN, while the knowledge transfer is removed. *3) ST-D3QN*: The routing network and the knowledge transfer are not included, and no task set is established. The changes in speed profiles and headways are treated as uncertainties. *4) MTMH-SAC*: The multi-task multi-head soft-actor-critic algorithm which uses an independent head for each task. We revised the realization in [69] to output discrete actions. The above methods are running with 4000 episodes and 3 random seeds. Besides, *5) Genetic algorithm (GA)*: GA is directly implemented on the test set, where Q^* is treated as the fitness function. To decrease the computational complexity, we perform GA for each test scenario individually. The population size is 40, the crossover fraction is 0.9, the mutation fraction is 0.1, and the maximum generation is 100. *6) FTPA:* as illustrated in subsection 3.4.2.1.

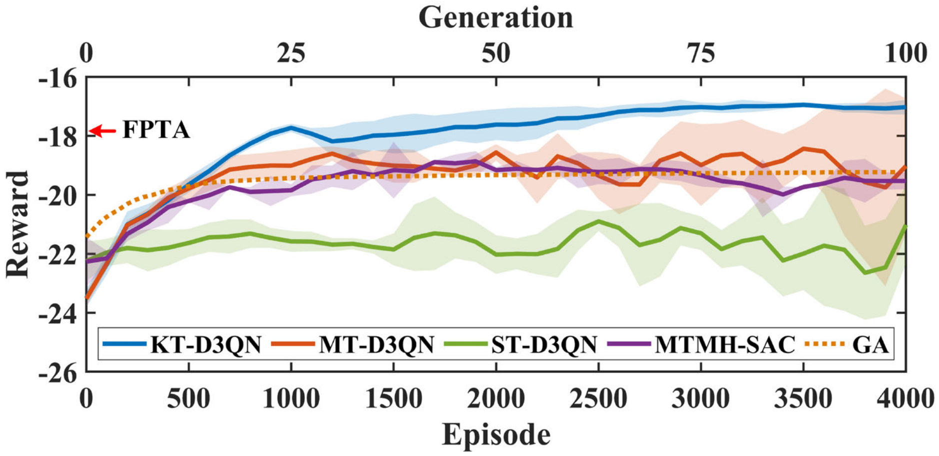


Fig. 3.20 Comparative reward curves of algorithms 1–6.

Fig. 3.20 shows the reward curves of the test set, where the bold line is the average value, the shaded area is one standard deviation, and the curves of learning-based algorithms are smoothed with a moving average smoothing factor of 0.1 for visual clarity. ST-D3QN gains the lowest reward and shows little improvement with episodes, which indicates that a single-task learning framework is insufficient to handle different headways and multi-source operation uncertainties. KT-D3QN achieves a stable performance and finds a near-optimal control policy after 3000 episodes. It obtains the highest reward.

Table 3.10 shows the RB energy utilization and overall operation cost of the test set, along with the best performance for each algorithm. Although FTPA achieves the highest RB energy utilization, its cost is higher than KT-D3QN and MT-D3QN. This is because the improved RB energy recovery of HESS also increases its operation cost. Hence, due to the multi-task learning framework and knowledge transfer, KT-D3QN outperforms other algorithms in improving economic benefits by 4.04%-13.06%, respectively, which verifies its effectiveness.

Table 3.10 Comparative RB energy utilizations and costs of algorithms 1–6.

Performance	KT-D3QN	MT-D3QN	ST-D3QN
Braking energy (MWh)	25.41	25.41	25.41
Braking loss (MWh)	7.98	6.09	9.03
RB energy (MWh)	17.43	19.32	16.38
Utilization (%)	68.60	76.03	64.46
Cost ($)	351.57	366.39	404.37
Performance	MTMH-SAC	GA	FTPA
Braking energy (MWh)	25.41	25.41	25.41
Braking loss (MWh)	8.41	9.45	5.67
RB energy (MWh)	17.00	15.96	19.74
Utilization (%)	66.90	62.81	77.69
Cost ($)	387.54	403.83	373.56

3.4.3 Analysis of Hybrid Energy Storage System Configurations

3.4.3.1 Configuration Results & Impact of Traffic Models

In this subsection, the effectiveness of the proposed DTM and the optimal configuration of HESS are investigated. The following traffic models are compared: *1) Dynamic traffic model:* proposed*.* *2) Static traffic model:* Only one most common traction load scenario is generated for each headway, and one daily operation scenario containing 122 such traction load scenarios with different headways is used for sizing optimization. Specifically, this daily operation scenario assumes no delays occur and the daily passenger flows follow the historical average daily passenger curve. The train resistance uncertainty is not considered. The initial HESS SoEs are set as the maximum SoE. *3) Static passenger model:* passenger uncertainty is not considered, and the historical average daily passenger curve is used for all daily operation scenarios. *4) No delay model:* the delays and RTTR are ignored, and only the normal operation scenarios in the traction load scenarios are adopted to establish daily operation scenarios. These traffic models are combined with different energy management strategies to optimize the HESS size. Specifically, we use framework F1–F4 to denote the results of combining KT-D3QN with traffic models 1)–4), respectively, and framework F5–F8 to denote the results of combining FTPA with traffic models 1)–4), respectively. F6 (FTPA and static traffic model) is the conventional approach and baseline.

Table 3.11 shows the LCC and optimal HESS size under various optimization frameworks. Compared with F1, F2–F4 lacks the consideration of spatial-temporal traction load characteristics on different degrees, which results in the LCC underestimation. Similarly, the LCCs of F6–F8 are lower than F5 due to the lack of the proposed dynamic traffic model. Besides, the LCCs of F5–F8 are significantly higher than F1–F4, which further verifies the effectiveness of KT-D3QN. Compared with the conventional approach F6, the proposed framework F1 reduces the HESS LCC, capacity, and power by 2.65%, 12.29%, and 17.63%, respectively, while increasing the battery life by 86.22%.

Table 3.11 Comparative LCCs and optimal HESS sizes of frameworks 1–8.

Performance	F1	F2	F3	F4
LCC ($)	1283.53	1184.99	1234.59	1209.40
Supercapacitor capacity (kWh)	14.78	17.74	14.78	17.74
Battery capacity (kWh)	107.46	80.59	107.46	80.59
Supercapacitor power (kW)	1.75	2.10	1.75	2.10
Battery power (kW)	0.54	0.40	0.54	0.40
Battery life (year)	10.00	10.00	10.00	10.00
Performance	F5	F6	F7	F8
LCC ($)	1340.22	1318.48	1301.45	1327.83
Supercapacitor capacity (kWh)	17.74	18.48	17.74	17.74
Battery capacity (kWh)	107.46	120.89	107.46	120.89
Supercapacitor power (kW)	2.10	2.18	2.10	2.10
Battery power (kW)	0.54	0.60	0.54	0.60
Battery life (year)	5.40	5.25	5.37	5.64

3.4.3.2 Impact of RTTR Algorithms

In this subsection, the effectiveness of the energy-saving-oriented RTTR based on TEPT sensitivity is verified. The following RTTR methods are compared using the test set. It is worth noting that only delay scenarios are included in this comparison. *1) Method 1 (M1):* Proposed TEPT-based RTTR. *2) Method 2 (M2):* A conventional rescheduling method aiming to minimize the delay time, which is illustrated in line 8 of algorithm 3.1. First, it tries to allocate the delay time $T_{i,\rho}^{\mathrm{D}}$ to the first section after the delayed station. Then, if there is remaining $T_{i,\rho}^{\mathrm{D}}$, allocate it to the second section. Repeat the process till $T_{i,\rho}^{\mathrm{D}}=0$. *3) Method 3 (M3):* no RTTR is implemented.

Table 3.12 Comparative overall operation costs and traction substation energies of RTTR methods 1–3.

Performance	M1	M2	M3
Overall operation cost ($)	301.73	311.54	294.53
substation energy (kWh)	987.89	1021.01	964.00

Table 3.12 shows the overall operation cost and traction substation energy outputs under different RTTR methods. M3 achieves the lowest cost and energy consumption following the original train trajectories and running times. However, the delay time has not been reduced. Although both M1 and M2 can minimize the delay time, M1 achieves better cost and energy efficiency due to the consideration of energy sensitivity in switching train trajectories.

3.5 Summary

In this chapter, an MTRL–SCO approach is proposed for enhancing the economic operations of HESSs and their integrated traction substations under dynamic spatial-temporal URT traffic. The research mainly includes the following aspects.

The configuration-specific HESS control problem under various spatial-temporal traction load distributions is formulated as an MTMDP, and an iterative sizing optimization approach considering daily service patterns is devised to minimize the HESS LCC. Then, a DTM composed of a Copula-based passenger flow generation method and a traction energy sensitivity-based RTTR algorithm is developed to characterize multi-train traction load uncertainty. Furthermore, a KT-D3QN algorithm is proposed to simultaneously learn a generalized multi-task HESS control policy from knowledge of annealing task-specific agents and operation environments. Finally, comparative studies have validated the effectiveness of the proposed approach for LCC reduction of HESS operation under URT traffic.

The key findings of the designated case study are summarized as follows: 1) With the joint optimization of voltage thresholds and power allocations in the MTMDP to effectively adjust SoEs, the operation cost can be reduced by 5.89% compared with conventional rule-based strategies using fixed thresholds and power allocations. 2) Leveraging the multi-task learning framework and knowledge transfer, the proposed KT-D3QN algorithm shows superior economic performance compared to benchmark learning-based and non-learning-based methods, decreasing the average overall operation cost by 4.04–13.06%. 3) The lack of consideration of spatial-temporal traction load characteristics can result in substantial LCC underestimation up to 6.69% for optimal HESS configuration. Compared with the conventional approach, the proposed optimization framework reduces the HESS LCC by 2.65% while increasing the battery life by 86.22%.

Chapter 4: Multi-Time Scale Energy Management for Distributed Hybrid Energy Storage System-Integrated Traction Network Operation Based on Multi-Task Multi-Agent Reinforcement Learning

Nomenclature in this chapter

A. Multi-Task Multi-Agent Reinforcement Learning Elements

a , s , o , r , z	Action, state, observation, reward, and task
B	Replay buffer
l_r	Length of history trajectory used for loss updates
Q_i , Q_{tot}	Q value of agent *i* and joint action-value function
$\mathcal{I},\mathcal{S},\mathcal{O},\mathcal{A},\mathcal{R},\mathcal{P}$	Sets of agents, states, observations, actions, rewards, and transitions
$\mathcal{Z}$, $\mathcal{Z}_H$, $\mathcal{Z}_P$	Sets of tasks, headway tasks, and combination tasks of trajectories
$\mathcal{L}$, $\mathcal{L}_\varphi$, $\mathcal{L}_{KL}$	Weighted *Q* loss, *Q* loss, and knowledge transfer loss
μ	Agent policy
κ , $\boldsymbol{\kappa}$, $\boldsymbol{a}$	Observation history, joint observation history, and joint action
γ	Discount factor
τ	Soft update rate
λ	Weight of knowledge transfer
ξ_q	Learning rate of neural networks
ς	Hidden state of the RNN
φ , φ' , φ_r , φ_m	Parameters of *Q,* target *Q*, routing, and mixing networks
Δr	PV-battery energy utilization reward

B. Indices

$h \in \{1,2,\cdots,H\}$ Index of train headways

$i \in \{1,2,\cdots,I\}$ Index of stations or traction substations

$j \in \{1,2,\cdots,J\}$ Index of stations or traction substations except for station or substation i

$k \in \{1,2,\cdots,K\}$ Index of trains

$t_r \in \{1,2,\cdots,T_r\}$ Index of the intraday time horizon

$w \in \{1,2,\cdots,W\}$ Index of transitions sampled from replay buffer

C. Time Scales

Δn, n, N Increment, current time step, and time horizon on a long time scale for economic dispatch and prediction (e.g., sub-hourly or hourly)

Δt, t, T Increment, current time step, and time horizon on a short time scale for real-time train and HESS control (e.g., sub-minutely)

D. Variables

E^{ref}, ΔE^{ref} Referential and available PV-battery energies (kWh)

I^{SC}, $I^{\mathrm{PV\text{-}BT}}$, I^{SUB} Currents of supercapacitor, PV-battery, and traction substation (A)

J^{DA}, J^{INT} Objectives of day-ahead and intraday scheduling

J^{D} Cost for any deviation from $E^{\mathrm{ref,INT}}$ (\$)

J^{SUB}, J^{CA}, J^{CUR} Costs of electricity trading, carbon trading, and PV curtailment (\$)

J^{OM}, J^{SC}, J^{BT}, J^{PV} Costs of PV-HESS, supercapacitor, battery, and PV operation (\$)

N^{B} Passengers who are onboard

$P^{\mathrm{PV\text{-}BT}}$ Scheduled power of PV-battery (W)

P^{PV}, P^{CUR} PV power and its curtailment (W)

P^{TR}, P^{SUB} Powers of train and traction substation (W)

$P^{\text{SC,CH}}, P^{\text{SC,DIS}}$	Charging and discharging powers of supercapacitor (W)
$P^{\text{BT,CH}}, P^{\text{BT,DIS}}$	Charging and discharging powers of battery (W)
R^{PV}	Solar irradiance (W/m^2)
$\text{SoE}^{\text{SC}}, \text{SoE}^{\text{BT}}$	SoEs of supercapacitor and battery (%)
$U^{\text{SC}}, U^{\text{C}}$	Terminal and capacitance voltages of supercapacitor (V)
$U^{\text{SUB}}, U^{\text{PV-BT}}$	Voltages of traction substation and PV-battery (V)
$U^{\text{CH}}, U^{\text{DIS}}$	Charge and discharge voltage thresholds of DHESS (V)
ρ, f_ρ	Scenario and its probability
ψ	Binary indicator of battery charging/discharging
ζ	DHESS operation mode

E. Parameters

A^{PV}	Area of PV arrays (m^2)
c_p, c_s, c_b	Weight coefficients of intraday objective
$c_{\text{SC}}^{\text{OM}}, c_{\text{BT}}^{\text{OM}}$	Unit costs of supercapacitor and battery operation ($/MWh)
$c_{\text{PV}}^{\text{OM}}, c_{\text{CUR}}$	Unit costs of PV operation and curtailment ($/kWh)
c_{SUB}, c_{CO_2}	Unit costs of electricity and carbon trading ($/kWh)
C^{SC}	Equivalent capacitance of supercapacitor (F)
k_e, k_q	Coefficient of carbon emission and quote (kg/kWh)
R^{SC}	Resistance of supercapacitor (Ω)
$\eta^{\text{SC}}, \eta^{\text{BT}}, \eta^{\text{PV}}$	Efficiencies of supercapacitor, battery, and PV

4.1 Background

The integration of distributed photovoltaics (PVs), regenerative braking (RB) techniques, and energy storage devices has become crucial to promote energy conservation and emission reduction for a sustainable future of urban rail transit (URT)

TNs. Nowadays, different aspects of distributed hybrid energy storage system (DHESS)-integrated traction network (TN) energy management have been studied in previous works (literature reviews in section 1.3). However, a multi-time scale energy management strategy to fully utilize inherent temporal and spatial operational flexibilities of TNs regarding energy generation (both PV and RB energy) and storage (DHESSs) for promoting renewable energy utilization and low-carbon economic TN operation requires comprehensive consideration (network-level challenges in section 1.4). Therefore, this chapter focuses on developing a multi-task multi-agent reinforcement learning-based multi-time scale energy management (MTMARL–MTSEM) approach for the economic and low-carbon operation of TNs with PV–RB DHESSs at the 3rd (network) level. Specifically, the main contributions of this chapter are outlined as follows:

- A tri-level MTSEM framework is proposed for optimal synergies of TN operation with solar generation to minimize the overall cost. A two-stage stochastic scheduling approach is developed on a long time scale to minimize daily operation and carbon trading costs at the upper level and correct day-ahead scheduling deviations at the middle level. Compared with previous frameworks, a lower level is added to coordinate RB energy recycling and PV energy consumption optimally on a short time scale for further cost reduction.
- A MTMARL-based real-time energy management algorithm (MTMARL–RTEMA) is proposed to optimize PV–RB power flow and promote its utilization by coordinating DHESSs, and the DHESS control problem is formulated as a decentralized partially observable Markov decision process (Dec-POMDP). An MTMARL algorithm based on monotonic value function

factorization, recurrent experience replay (RER), and knowledge transfer (KT) is developed to solve the Dec-POMDP effectively. A generalized and decentralized control scheme can be formed and adapted to different train service patterns and network uncertainties without knowing precise system models and uncertainty parameters.

- A Copula-based spatial-temporal dependency model is devised to characterize uncertainties of PVs, multi-station passenger flows, and multi-train traction loads. Latin hypercube sampling (LHS) is employed to generate typical daily TN operation scenarios for enhancing day-ahead and intraday decisions.

Finally, comparative studies are implemented to validate the effectiveness of the proposed approach. Section 4.2 illustrates the problem formulation, including the structure, operation modes, and modeling of the studied TN. Section 4.3 presents the proposed MTMARL–MTSEM approach, including detailed energy management methods at the upper, middle, and lower levels. Section 4.4 reports case studies and their results. Section 4.5 gives the summary.

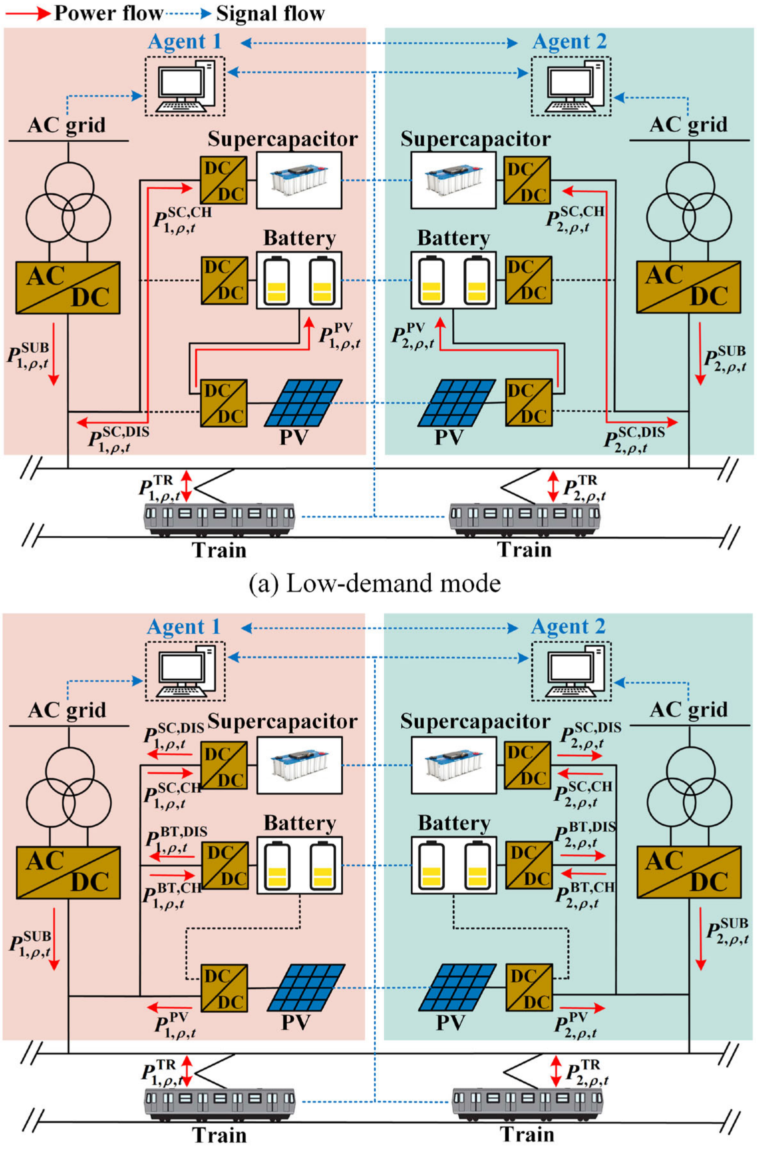
4.2 Problem Formulation

4.2.1 Structure and Operation Modes of Traction Networks with Distributed Hybrid Energy Storage Systems

To consider the impact of PV integration, a PV access scheme is proposed [20] (Fig. 4.1(a)–(b)). The PV is connected to the TN and battery via a single-input dual-output DC/DC converter. The battery and supercapacitor are connected to the traction network via a bidirectional DC/DC converter.

Due to the changes in daily train service patterns and the significant temporal mismatch between peak PV output and peak passenger demand in the morning/evening, the cost-effectiveness of PV integration can be undermined. For example, the peak PV output generally occurs at noon, which is the off-peak hour of urban rail operation. During this off-peak hour, the train headway is large, which means that fewer train services will be provided. Besides, the number of passengers is also smaller. Hence, it is reasonable to store excessive PV energy during periods of low passenger demand for later use to maximize PV usage. Besides, considering the uncertainty of PV output, energy storage devices are also needed to facilitate continuous and stable renewable energy supply. Batteries that possess large capacities can be utilized for these functions. However, the timing of storing PV energy should be considered to minimize the overall daily cost of URTs. Therefore, we designed two different operation modes for the distributed PV-battery system to manage the timing for PV energy storage and utilization, namely, the low-demand mode (LDM) and the high-demand mode (HDM).

In the LDM, the output portal from the PV to the TN is off. The PV can only charge the battery. The traction substation rectifier and supercapacitor provide energy for trains. The charging/discharging of the supercapacitor is determined based on the traction substation voltage [132]. It charges when the traction substation voltage rises above the charge voltage threshold and discharges when it drops below the discharge voltage threshold. In the HDM, the output portal from PV to the TN is on, and the supercapacitor operation is similar to LDM. The PV and battery can supply the TN when the traction substation voltage drops below the discharge voltage threshold. Considering its limited power, the battery will not absorb energy from the network to avoid excess charging and extend its lifetime. Therefore, the PV-battery operation is similar to LDM when the traction substation voltage rises. In this chapter, a binary signal $\zeta \in \{0,1\}$ distinguishes LDM and HDM, where 0 represents the LDM, and represents the HDM. Considering the limited roof area of the station, the installed PV capacity generally cannot satisfy the traction energy demand alone. Hence, the energy flow from PV to the battery can be ignored in the HDM, and the output portal from PV to the battery can be treated as closed.


Power flow
Signal flow
Agent 1
Agent 2
AC grid
Supercapacitor
Battery
PV
AC
DC
DC
DC
$P^{\mathrm{SC,CH}}_{1,\rho,t}$
$P^{\mathrm{SC,CH}}_{2,\rho,t}$
$P^{\mathrm{SC,DIS}}_{1,\rho,t}$
$P^{\mathrm{SC,DIS}}_{2,\rho,t}$
$P^{\mathrm{BT,CH}}_{1,\rho,t}$
$P^{\mathrm{BT,CH}}_{2,\rho,t}$
$P^{\mathrm{BT,DIS}}_{1,\rho,t}$
$P^{\mathrm{BT,DIS}}_{2,\rho,t}$
$P^{\mathrm{PV}}_{1,\rho,t}$
$P^{\mathrm{PV}}_{2,\rho,t}$
$P^{\mathrm{SUB}}_{1,\rho,t}$
$P^{\mathrm{SUB}}_{2,\rho,t}$
$P^{\mathrm{TR}}_{1,\rho,t}$
$P^{\mathrm{TR}}_{2,\rho,t}$
Train

(a) Low-demand mode

(b) High-demand mode

Fig. 4.1 (a) Low-demand and (b) high-demand operation modes.

Through the above analysis, it can be concluded that LDM and HDM can be realized by controlling the on/off of the portals of the single-input dual-output DC/DC converter. It is worth noting that, for pure optimization purposes, LDM can be regarded as a special situation of HDM where all PV-battery system outputs are zero. Nevertheless, considering their physical realization and control signal settings in the actual operation, we distinguish these two operation modes in this work.

4.2.2 Modeling of Traction Networks with Distributed Hybrid Energy Storage Systems

The DHESS-integrated TN model is shown in Fig. 4.2, where traction substation and train models are the same as (3.2)–(3.3). Nevertheless, considering the operation modes in subsection 4.2.1, the PV-battery system is modeled as an equivalent constant current source [172] to replace the battery model in (3.7)–(3.8). Besides, due to its large capacity, the battery generally does not reach the state-of-energy (SoE) limits within seconds. Therefore, the battery SoE limits are considered with the long time scale Δn. For supercapacitor, (3.9)–(3.10) are modified as

$$U_{i,\rho,t}^{\mathrm{SC}} = U_{i,\rho,t}^{\mathrm{C}} - I_{i,\rho,t}^{\mathrm{SC}} R^{\mathrm{SC}}, \quad U_{i,\rho,t}^{\mathrm{C}} = U_{i,\rho,t-1}^{\mathrm{C}} - I_{i,\rho,t}^{\mathrm{SC}} \Delta t / C^{\mathrm{SC}}, \tag{4.1}$$

$$I_{i,\rho,t}^{\mathrm{SC}} = \begin{cases} \dfrac{U_{i,\rho,t}^{\mathrm{C}} - \sqrt{\left(U_{i,\rho,t}^{\mathrm{C}}\right)^2 - 4R^{\mathrm{SC}} P_{i,\rho,t}^{\mathrm{SC}} / \eta^{\mathrm{SC}}}}{2R^{\mathrm{SC}}}, & P_{i,\rho,t}^{\mathrm{SC}} = P_{i,\rho,t}^{\mathrm{SC,DIS}}, \\ -\dfrac{U_{i,\rho,t}^{\mathrm{C}} - \sqrt{\left(U_{i,\rho,t}^{\mathrm{C}}\right)^2 - 4R^{\mathrm{SC}} \eta^{\mathrm{SC}} P_{i,\rho,t}^{\mathrm{SC}}}}{2R^{\mathrm{SC}}}, & P_{i,\rho,t}^{\mathrm{SC}} = P_{i,\rho,t}^{\mathrm{SC,CH}}. \end{cases} \tag{4.2}$$

The SoE of the supercapacitor is constrained by

$$\mathrm{SoE}_{i,\rho,t}^{\mathrm{SC}} = \left(\frac{U_{i,\rho,t}^{\mathrm{C}}}{U^{\mathrm{C,norm}}} \right)^2, \quad \mathrm{SoE}_{\min}^{\mathrm{SC}} \le \mathrm{SoE}_{i,\rho,t}^{\mathrm{SC}} \le \mathrm{SoE}_{\max}^{\mathrm{SC}}, \tag{4.3}$$

The power flow of such TNs is calculated by (3.11)–(3.13) and (4.4), where we use (4.4) to replace (3.14) for obtaining nodal current under the situation of DHESSs.

$$I_{i,\rho,t}^{\mathrm{S}} = I_{i,\rho,t}^{\mathrm{SUB}} + \left(P_{i,\rho,t}^{\mathrm{SC}} + P_{i,\rho,t}^{\mathrm{PV\text{-}BT}} \right) / U_{i,\rho,t}^{\mathrm{SUB}}, \tag{4.4}$$

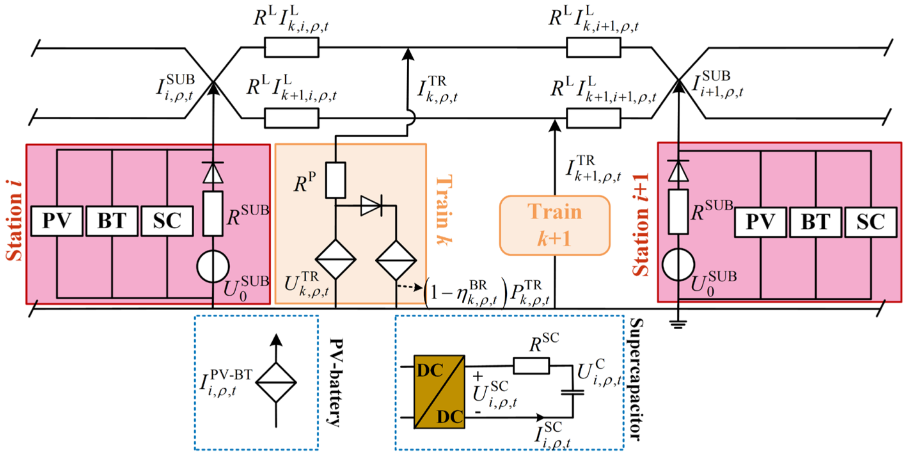

$R^{\mathrm{L}} I_{k,i,\rho,t}^{\mathrm{L}}$
$R^{\mathrm{L}} I_{k,i+1,\rho,t}^{\mathrm{L}}$
$I_{i,\rho,t}^{\mathrm{SUB}}$
$R^{\mathrm{L}} I_{k+1,i,\rho,t}^{\mathrm{L}}$
$I_{k,\rho,t}^{\mathrm{TR}}$
$R^{\mathrm{L}} I_{k+1,i+1,\rho,t}^{\mathrm{L}}$
$I_{i+1,\rho,t}^{\mathrm{SUB}}$
Station *i*
PV
BT
SC
R^{SUB}
U_0^{SUB}
R^{P}
$U_{k,\rho,t}^{\mathrm{TR}}$
Train *k*
$(1-\eta_{k,\rho,t}^{\mathrm{BR}})P_{k,\rho,t}^{\mathrm{TR}}$
$I_{k+1,\rho,t}^{\mathrm{TR}}$
Train *k*+1
Station *i*+1
R^{SUB}
PV
BT
SC
U_0^{SUB}
$I_{i,\rho,t}^{\mathrm{PV\text{-}BT}}$
PV-battery
DC
DC
R^{SC}
+
$U_{i,\rho,t}^{\mathrm{SC}}$
-
$U_{i,\rho,t}^{\mathrm{C}}$
$I_{i,\rho,t}^{\mathrm{SC}}$
Supercapacitor

Fig. 4.2 Equivalent circuit model of TN with PV–RB DHESSs.

4.3 MTMARL–MTSEM Approach

4.3.1 Approach Overview

The proposed approach is essentially a tri-level energy management framework (Fig. 4.3). At the upper level, a day-ahead scheduling plan of operation mode ζ and referential PV-battery energy E^{ref} is formulated to minimize the daily operation cost. Meanwhile, the day-ahead referential traction substation power P^{SUB}, supercapacitor SoEs $\mathrm{SoE}^{\mathrm{SC}}$, and battery $\mathrm{SoE}^{\mathrm{BT}}$ are generated. At the middle level, an intraday scheduling plan of intraday referential PV-battery energy $E^{\mathrm{ref,INT}}$ is formulated. to minimize the operation deviation from the day-ahead schedule. Specifically, the deviation includes P^{SUB}, $\mathrm{SoE}^{\mathrm{SC}}$, and $\mathrm{SoE}^{\mathrm{BT}}$, and the optimization is performed repeatedly with a finite time horizon. At the lower level, according to $E^{\mathrm{ref,INT}}$, the real-time PV–RB energy utilization is optimized by formulating the DHESS control problem into a Dec-POMDP and solving it with MTMARL. A penalty term $J^{\mathrm{D}}_{i,\rho,T}$ and a reward term Δr are utilized to follow the intraday schedule. The optimal solution consists of the real-time charge/discharge voltage thresholds of each HESS and the real-time power allocation between each PV-battery and supercapacitor. For upper and middle levels, the time scale $\Delta n = 15$ min. For the lower level, the time scale $\Delta t = 1$ s.

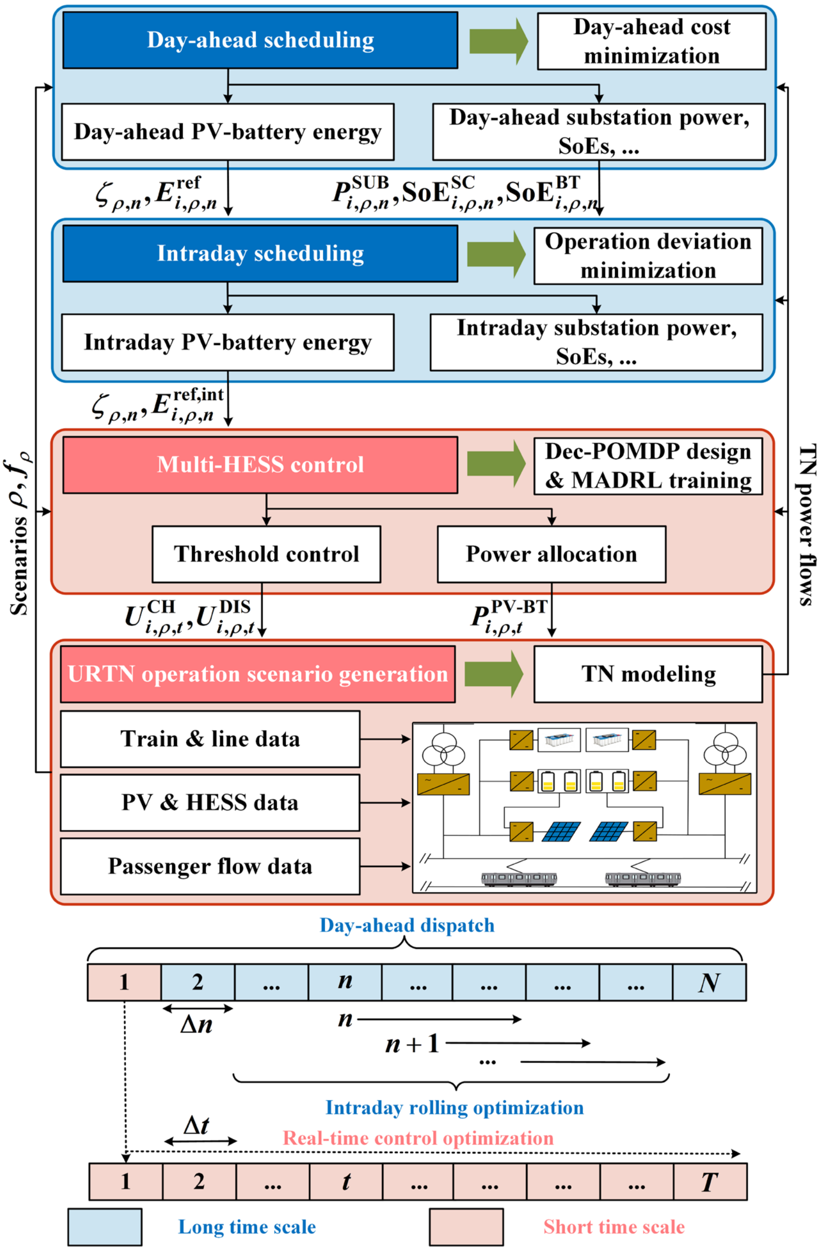

Day-ahead scheduling
Day-ahead cost minimization
Day-ahead PV-battery energy
Day-ahead substation power, SoEs, ...
$\zeta_{\rho,n}, E^{\mathrm{ref}}_{i,\rho,n}$
$P^{\mathrm{SUB}}_{i,\rho,n}, \mathrm{SoE}^{\mathrm{SC}}_{i,\rho,n}, \mathrm{SoE}^{\mathrm{BT}}_{i,\rho,n}$
Intraday scheduling
Operation deviation minimization
Intraday PV-battery energy
Intraday substation power, SoEs, ...
$\zeta_{\rho,n}, E^{\mathrm{ref,int}}_{i,\rho,n}$
Multi-HESS control
Dec-POMDP design & MADRL training
Threshold control
Power allocation
$U^{\mathrm{CH}}_{i,\rho,t}, U^{\mathrm{DIS}}_{i,\rho,t}$
$P^{\mathrm{PV\text{-}BT}}_{i,\rho,t}$
URTN operation scenario generation
TN modeling
Train & line data
PV & HESS data
Passenger flow data
Scenarios ρ, f_{ρ}
TN power flows
Day-ahead dispatch
1
2
...
n
...
...
...
...
N
Δn
n
$n+1$
...
Intraday rolling optimization
Δt
Real-time control optimization
1
2
...
t
...
...
...
...
T
Long time scale
Short time scale

Fig. 4.3 Tri-level framework of MTMARL–MTSEM.

4.3.2 Upper Level

4.3.2.1 Objective

The overall daily cost is (4.5), and each scenario ρ has an occurrence probability f_ρ. Specific cost terms are demonstrated in (4.6)–(4.12), including electricity trading cost $J_{\rho,n}^{\mathrm{SUB}}$, supercapacitor operation cost $J_{\rho,n}^{\mathrm{SC}}$, battery operation cost $J_{\rho,n}^{\mathrm{BT}}$, PV operation cost $J_{\rho,n}^{\mathrm{PV}}$, PV curtailment cost $J_{\rho,n}^{\mathrm{CUR}}$, and carbon trading cost $J_{\rho,n}^{\mathrm{CA}}$. Based on the current carbon trading mechanism [173], c_{CO_2} is the carbon trading price, k_q and k_e are the carbon emission coefficient and carbon quote coefficient, respectively. $P_{i,\rho,n}^{\mathrm{SUB}}$, $P_{i,\rho,n}^{\mathrm{SC,CH}}$, $P_{i,\rho,n}^{\mathrm{SC,DIS}}$, $P_{i,\rho,n}^{\mathrm{BT,CH}}$, $P_{i,\rho,n}^{\mathrm{BT,DIS}}$, $P_{i,\rho,n}^{\mathrm{PV}}$, and $P_{i,\rho,n}^{\mathrm{CUR}}$ denote the purchased power, supercapacitor charging/discharging power, battery charging/discharging power, PV power, and PV curtailment, respectively. N and I are the number of time steps with Δn and the number of traction substations, respectively. c_{SUB}, $c_{\mathrm{SC}}^{\mathrm{OM}}$, $c_{\mathrm{BT}}^{\mathrm{OM}}$, $c_{\mathrm{PV}}^{\mathrm{OM}}$, and c_{CUR} are the prices for energy purchasing, supercapacitor operation, battery operation, PV operation, and PV curtailment, respectively.

$$\min \sum_{\rho} \sum_{n} J_{\rho,n}^{\mathrm{DA}} f_\rho, \tag{4.5}$$

$$J_{\rho,n}^{\mathrm{DA}} = J_{\rho,n}^{\mathrm{SUB}} + J_{\rho,n}^{\mathrm{SC}} + J_{\rho,n}^{\mathrm{BT}} + J_{\rho,n}^{\mathrm{PV}} + J_{\rho,n}^{\mathrm{CUR}} + J_{\rho,n}^{\mathrm{CA}}, \tag{4.6}$$

$$J_{\rho,n}^{\mathrm{SUB}} = \sum\nolimits_{i=1}^{I} c_{\mathrm{SUB}} P_{i,\rho,n}^{\mathrm{SUB}} \Delta n, \tag{4.7}$$

$$J_{\rho,n}^{\mathrm{SC}} = \sum\nolimits_{i=1}^{I} c_{\mathrm{SC}}^{\mathrm{OM}} (P_{i,\rho,n}^{\mathrm{SC,CH}} + P_{i,\rho,n}^{\mathrm{SC,DIS}}) \Delta n, \tag{4.8}$$

$$J_{\rho,n}^{\mathrm{BT}} = \sum\nolimits_{i=1}^{I} c_{\mathrm{BT}}^{\mathrm{OM}} (P_{i,\rho,n}^{\mathrm{BT,CH}} + P_{i,\rho,n}^{\mathrm{BT,DIS}}) \Delta n, \tag{4.9}$$

$$J_{\rho,n}^{\mathrm{PV}} = \sum\nolimits_{i=1}^{I} c_{\mathrm{PV}}^{\mathrm{OM}} P_{i,\rho,n}^{\mathrm{PV}} \Delta n, \tag{4.10}$$

$$J_{\rho,n}^{\mathrm{CUR}} = \sum\nolimits_{i=1}^{I} c_{\mathrm{CUR}} P_{i,\rho,n}^{\mathrm{CUR}} \Delta n, \tag{4.11}$$

$$J_{\rho,n}^{\mathrm{CA}} = \sum\nolimits_{i=1}^{I} c_{CO_2} (k_e - k_q) P_{i,\rho,n}^{\mathrm{SUB}} \Delta n. \tag{4.12}$$

4.3.2.2 Energy & Power Balance Constraints

$\zeta_{\rho,n}$ and $E^{\text{ref}}_{i,\rho,n}$ in (4.13) can be utilized by MTMARL–RTEMA as the reward in (4.33) to determine $P^{\text{PV-BT}}_{i,\rho,t}$ at the lower level. Then, the power flow is solved with $P^{\text{PV-BT}}_{i,\rho,t}$. Finally, the energy terms in (4.7)–(4.8) can be obtained by summing up their counterparts with Δt by (4.14)–(4.15). The power balance is shown in (4.16)–(4.17). T is the number of time steps with Δt.

$$\zeta_{\rho,n}E^{\text{ref}}_{i,\rho,n} = \sum\nolimits_{t=1}^{T} P^{\text{PV-BT}}_{i,\rho,t}\Delta t, \tag{4.13}$$

$$P^{\text{SUB}}_{i,\rho,n}\Delta n = \sum\nolimits_{t=1}^{T} U^{\text{SUB}}_{i,\rho,t}I^{\text{SUB}}_{i,\rho,t}\Delta t, \tag{4.14}$$

$$\left(P^{\text{SC,CH}}_{i,\rho,n} + P^{\text{SC,DIS}}_{i,\rho,n}\right)\Delta n = \sum\nolimits_{t=1}^{T}\left(P^{\text{SC,CH}}_{i,\rho,t} + P^{\text{SC,DIS}}_{i,\rho,t}\right)\Delta t, \tag{4.15}$$

$$P^{\text{PV}}_{i,\rho,n} - P^{\text{CUR}}_{i,\rho,n} = \frac{\zeta_{\rho,n}E^{\text{ref}}_{i,\rho,n}}{\Delta n} + P^{\text{BT,CH}}_{i,\rho,n} - P^{\text{BT,DIS}}_{i,\rho,n}, \tag{4.16}$$

$$0 \le P^{\text{CUR}}_{i,\rho,n} \le P^{\text{PV}}_{i,\rho,n}. \tag{4.17}$$

4.3.2.3 DHESS Charge & Discharge Constraints

The supercapacitor behaviour is modeled by (4.1)–(4.3), where its SoEs are also constrained by (4.18)–(4.19). The battery behaviour is modeled by (4.20)–(4.22), where simultaneous charging/discharging event is prevented by (4.21). η^{BT} is the battery charging/discharging efficiency.

$$\text{SoE}^{\text{SC}}_{i,\rho,N} = \text{SoE}^{\text{SC}}_{i,\rho,1}, \tag{4.18}$$

$$\text{SoE}^{\text{BT}}_{i,\rho,N} = \text{SoE}^{\text{BT}}_{i,\rho,1}, \tag{4.19}$$

$$\text{SoE}^{\text{BT}}_{\min} \le \text{SoE}^{\text{BT}}_{i,\rho,n} \le \text{SoE}^{\text{BT}}_{\max}, \tag{4.20}$$

$$0 \le P^{\text{BT,DIS}}_{i,\rho,n} \le (1-\psi_{i,\rho,n})P^{\text{BT}}_{\max},\quad 0 \le P^{\text{BT,CH}}_{i,\rho,n} \le \psi_{i,\rho,n}P^{\text{BT}}_{\max}, \tag{4.21}$$

$$\mathrm{SoE}^{\mathrm{BT}}_{i,\rho,n}=\mathrm{SoE}^{\mathrm{BT}}_{i,\rho,n-1}+(\eta^{\mathrm{BT}}P^{\mathrm{BT,CH}}_{i,\rho,n}-\frac{P^{\mathrm{BT,DIS}}_{i,\rho,n}}{\eta^{\mathrm{BT}}})\Delta n. \tag{4.22}$$

4.3.3 Middle Level

The objective and constraints of the middle level are defined in (4.23)–(4.27). $P^{\mathrm{SUB}}_{i,\rho,n}$ is the day-ahead substation power. $\mathrm{SoE}^{\mathrm{SC}}_{i,\rho,n}$, and $\mathrm{SoE}^{\mathrm{BT}}_{i,\rho,n}$ are the day-ahead SoE of supercapacitor and battery, while $\mathrm{SoE}^{\mathrm{SC,INT}}_{i,\rho,n}$, and $\mathrm{SoE}^{\mathrm{BT,INT}}_{i,\rho,n}$ are the corresponding intraday variables, respectively. $P^{\mathrm{PV,INT}}_{i,\rho,n}$, $P^{\mathrm{CUR,INT}}_{i,\rho,n}$, $P^{\mathrm{SC,CH,INT}}_{i,\rho,n}$, $P^{\mathrm{SC,DIS,INT}}_{i,\rho,n}$, $P^{\mathrm{BT,CH,INT}}_{i,\rho,n}$, and $P^{\mathrm{BT,DIS,INT}}_{i,\rho,n}$ are the corresponding intraday PV and DHESS power variables. c_p, c_s, and c_b are weight coefficients. T_r is the time horizon. The intraday scheduling is performed by: **Step 1:** At time n, based on current states (PV outputs, energy storage states, and traffic flows), perform optimization for time $n+1$, …, $n+t_r$, …, $n+T_r$. **Step 2:** Apply results only for time $n+1$. At time $n+1$, update system states and repeat step 1.

$$\begin{aligned}\min J^{\mathrm{INT}}=&\sum\nolimits_{t_r=1}^{T_r}\sum\nolimits_{i=1}^{I}[c_p(P^{\mathrm{SUB,INT}}_{i,\rho,t_r}-P^{\mathrm{SUB}}_{i,\rho,t_r})^2\\&+c_s(\mathrm{SoE}^{\mathrm{SC,INT}}_{i,\rho,t_r}-\mathrm{SoE}^{\mathrm{SC}}_{i,\rho,t_r})^2+c_b(\mathrm{SoE}^{\mathrm{BT,INT}}_{i,\rho,t_r}-\mathrm{SoE}^{\mathrm{BT}}_{i,\rho,t_r})^2],\\ \text{s.t. }&(3.1)-(3.6),(3.11)-(3.14),(4.1)-(4.4),(4.18)-(4.22),\end{aligned} \tag{4.23}$$

$$\zeta_{\rho,t_r}E^{\mathrm{ref,INT}}_{i,\rho,t_r}=\sum\nolimits_{t=1}^{T}P^{\mathrm{PV\text{-}BT}}_{i,\rho,t}\Delta t, \tag{4.24}$$

$$P^{\mathrm{SUB,INT}}_{i,\rho,t_r}\Delta n=\sum\nolimits_{t=1}^{T}U^{\mathrm{SUB}}_{i,\rho,t}I^{\mathrm{SUB}}_{i,\rho,t}\Delta t, \tag{4.25}$$

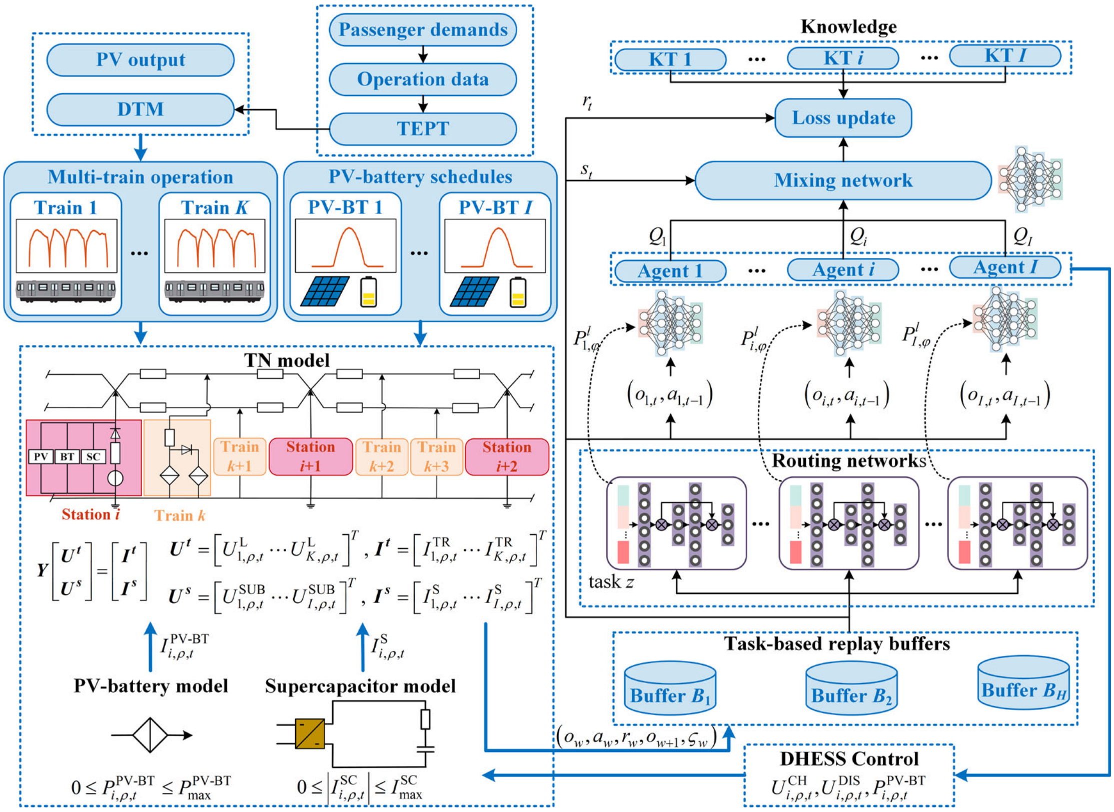
$$\left(P^{\mathrm{SC,CH,INT}}_{i,\rho,t_r}+P^{\mathrm{SC,DIS,INT}}_{i,\rho,t_r}\right)\Delta n=\sum\nolimits_{t=1}^{T}\left(P^{\mathrm{SC,CH}}_{i,\rho,t}+P^{\mathrm{SC,DIS}}_{i,\rho,t}\right)\Delta t, \tag{4.26}$$

$$P^{\mathrm{PV,INT}}_{i,\rho,t_r}-P^{\mathrm{CUR,INT}}_{i,\rho,t_r}=\frac{\zeta_{\rho,t_r}E^{\mathrm{ref,INT}}_{i,\rho,t_r}}{\Delta n}+P^{\mathrm{BT,CH,INT}}_{i,\rho,t_r}-P^{\mathrm{BT,DIS,INT}}_{i,\rho,t_r}. \tag{4.27}$$

4.3.4 Lower Level

4.3.4.1 Task Representation & Dec-POMDP

The illustration of the lower-level energy management, namely, the MTMARL–RTEMA, is shown in Fig. 4.4, where each HESS controller is considered an intelligent agent. First, the DHESS control problem is formulated as a Dec-POMDP. Generally, the Dec-POMDP contains $\langle \mathcal{I}, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$. Developed from the Markov game introduced in subsection 1.3.1.1, an observation set $\mathcal{O}$ is added to indicate that the environment is partially observable to the agents. The task representation is the same as subsection 3.3.3.1. Then, the Dec-POMDP is solved by the proposed MTMARL algorithm.



Passenger demands
Operation data
TEPT
PV output
DTM
Multi-train operation
Train 1
Train K
PV-battery schedules
PV-BT 1
PV-BT I
TN model
PV BT SC
Train $k+1$
Station $i+1$
Train $k+2$
Train $k+3$
Station $i+2$
Station i
Train k
$Y\begin{bmatrix}U^t\\U^s\end{bmatrix}=\begin{bmatrix}I^t\\I^s\end{bmatrix}$
$U^t=\left[U^{\mathrm{L}}_{1,\rho,t}\cdots U^{\mathrm{L}}_{K,\rho,t}\right]^T$, $I^t=\left[I^{\mathrm{TR}}_{1,\rho,t}\cdots I^{\mathrm{TR}}_{K,\rho,t}\right]^T$
$U^s=\left[U^{\mathrm{SUB}}_{1,\rho,t}\cdots U^{\mathrm{SUB}}_{I,\rho,t}\right]^T$, $I^s=\left[I^{\mathrm{S}}_{1,\rho,t}\cdots I^{\mathrm{S}}_{I,\rho,t}\right]^T$
$I^{\text{PV-BT}}_{i,\rho,t}$
$I^{\mathrm{S}}_{i,\rho,t}$
PV-battery model
Supercapacitor model
$0\le P^{\text{PV-BT}}_{i,\rho,t}\le P^{\text{PV-BT}}_{\max}$
$0\le\left|I^{\mathrm{SC}}_{i,\rho,t}\right|\le I^{\mathrm{SC}}_{\max}$
Knowledge
KT 1
KT i
KT I
r_t
Loss update
s_t
Mixing network
Q_1
Q_i
Q_I
Agent 1
Agent i
Agent I
$P^l_{1,\varphi}$
$P^l_{i,\varphi}$
$P^l_{I,\varphi}$
$(o_{1,t},a_{1,t-1})$
$(o_{i,t},a_{i,t-1})$
$(o_{I,t},a_{I,t-1})$
Routing networks
task z
Task-based replay buffers
Buffer B_1
Buffer B_2
Buffer B_H
$(o_w,a_w,r_w,o_{w+1},\varsigma_w)$
DHESS Control
$U^{\mathrm{CH}}_{i,\rho,t},U^{\mathrm{DIS}}_{i,\rho,t},P^{\text{PV-BT}}_{i,\rho,t}$

Fig. 4.4 Overview of MTMARL–RTEMA.

4.3.4.2 State & Observation

$o_{i,t}$ comprises two parts: **a) train operation status in the adjacent sections** $(i-1,i)$ and $(i,i+1)$, including position $x^{\mathrm{TR}}_{i-1,k,\rho,t}$ and $x^{\mathrm{TR}}_{i+1,k,\rho,t}$, direction $d^{\mathrm{TR}}_{i-1,k,\rho,t}$ and $d^{\mathrm{TR}}_{i+1,k,\rho,t}$, and power $P^{\mathrm{TR}}_{i-1,k,\rho,t}$ and $P^{\mathrm{TR}}_{i+1,k,\rho,t}$. If no train runs in adjacent sections, $o_{i,t}=\{0,0,0,0,0,0\}$. **b) local traction substation information**, including the local supercapacitor SoE $\mathrm{SoE}^{\mathrm{SC}}_{i,\rho,t}$, the local available PV-battery energy $\Delta E^{\mathrm{ref}}_{i,\rho,t}$, and the local traction substation output $U^{\mathrm{SUB}}_{i,\rho,t}$ and $I^{\mathrm{SUB}}_{i,\rho,t}$. $s\in\mathcal{S}$ includes all observations and other environmental information. $o_{i,t}$ is defined as

$$o_{i,t}=\begin{cases}\left\{x^{\mathrm{TR}}_{i-1,k,\rho,t},d^{\mathrm{TR}}_{i-1,k,\rho,t},P^{\mathrm{TR}}_{i-1,k,\rho,t},x^{\mathrm{TR}}_{i+1,k,\rho,t},d^{\mathrm{TR}}_{i+1,k,\rho,t},P^{\mathrm{TR}}_{i+1,k,\rho,t}\right\},\\ \left\{\mathrm{SoE}^{\mathrm{SC}}_{i,\rho,t},\Delta E^{\mathrm{ref}}_{i,\rho,t},U^{\mathrm{SUB}}_{i,\rho,t},I^{\mathrm{SUB}}_{i,\rho,t}\right\},\end{cases} \tag{4.28}$$

$$\Delta E^{\mathrm{ref}}_{i,\rho,t}=\frac{T}{N}E^{\mathrm{ref,INT}}_{i,\rho,n}-\sum\nolimits_{1}^{t}P^{\mathrm{PV\text{-}BT}}_{i,\rho,t}\Delta t. \tag{4.29}$$

4.3.4.3 Action

$a_{i,t}$ includes the charge/discharge voltage thresholds ($U^{\mathrm{CH}}_{i,\rho,t}$ and $U^{\mathrm{DIS}}_{i,\rho,t}$) and the power $P^{\mathrm{PV\text{-}BT}}_{i,\rho,t}$. Since the battery will not charge from the network, $P^{\mathrm{PV\text{-}BT}}_{i,\rho,t}=0$ when the traction substation voltage rises higher than the no-load voltage.

$$a_{i,t}=\left\{U^{\mathrm{CH}}_{i,\rho,t},U^{\mathrm{DIS}}_{i,\rho,t},P^{\mathrm{PB\text{-}BT}}_{i,\rho,t}\right\}. \tag{4.30}$$

4.3.4.4 Reward

The Dec-POMDP of the lower-level energy management is essentially a fully cooperative game. In such a game, all agents share the same reward r_t at each time step. Similar to subsection 3.3.3.2, r_t contains the minus of the total electricity trading cost $J^{\mathrm{SUB}}_{i,\rho,t}$ and operation cost $J^{\mathrm{OM}}_{i,\rho,t}$ at t. Besides, at the final time step T, a cost term $J^{\mathrm{D}}_{i,\rho,T}$ is added to penalize any deviation from the referential set-point $E^{\mathrm{ref,INT}}_{i,\rho,n}$. Nevertheless, $J^{\mathrm{D}}_{i,\rho,T}$ is a sparse cost term which may result in difficulty in connecting a long trajectory of actions to a distant reward [174]. Therefore, Δr is designed. It rewards the usage of PV-battery energy when $E^{\mathrm{ref,INT}}_{i,\rho,n}$ has not been reached while penalizing such usage when $E^{\mathrm{ref,INT}}_{i,\rho,n}$ has been reached.

$$r_t = \begin{cases} -\left(J^{\mathrm{SUB}}_{i,\rho,t} + J^{\mathrm{OM}}_{i,\rho,t} + J^{\mathrm{D}}_{i,\rho,T} + \Delta r\right), & t = T, \\ -\left(J^{\mathrm{SUB}}_{i,\rho,t} + J^{\mathrm{OM}}_{i,\rho,t} + \Delta r\right), & \text{Otherwise,} \end{cases} \tag{4.31}$$

$$J^{\mathrm{OM}}_{i,\rho,t} = \sum\nolimits_{i=1}^{I} \left[c^{\mathrm{OM}}_{\mathrm{SC}} \left| P^{\mathrm{SC}}_{i,\rho,t} \right| + (c^{\mathrm{OM}}_{\mathrm{BT}} + c^{\mathrm{OM}}_{\mathrm{PV}}) P^{\mathrm{PV\text{-}BT}}_{i,\rho,t} \right] \Delta t, \tag{4.32}$$

$$J^{\mathrm{D}}_{i,\rho,T} = \sum\nolimits_{i=1}^{I} c_{\mathrm{D}} \left| \Delta E^{\mathrm{ref}}_{i,\rho,T} \right|, \tag{4.33}$$

$$\Delta r = \sum\nolimits_{i=1}^{I} \begin{cases} c_{\mathrm{D}} P^{\mathrm{PV\text{-}BT}}_{i,\rho,t} \Delta t, & \Delta E^{\mathrm{ref}}_{i,\rho,t} > 0, \\ -c_{\mathrm{D}} P^{\mathrm{PV\text{-}BT}}_{i,\rho,t} \Delta t, & \Delta E^{\mathrm{ref}}_{i,\rho,t} \le 0. \end{cases} \tag{4.34}$$

4.3.4.5 State Transition

The train operation status $\left\{ x^{\mathrm{TR}}_{i-1,k,\rho,t}, d^{\mathrm{TR}}_{i-1,k,\rho,t}, P^{\mathrm{TR}}_{i-1,k,\rho,t}, x^{\mathrm{TR}}_{i+1,k,\rho,t}, d^{\mathrm{TR}}_{i+1,k,\rho,t}, P^{\mathrm{TR}}_{i+1,k,\rho,t} \right\}$ and observations that are dependent of $a_{i,t}$ (e.g., $\mathrm{SoE}^{\mathrm{SC}}_{i,\rho,t}$, $U^{\mathrm{SUB}}_{i,\rho,t}$, $I^{\mathrm{SUB}}_{i,\rho,t}$, and $\Delta E^{\mathrm{ref}}_{i,\rho,t}$) are updated as illustrated in subsection 3.3.3.3. Besides, in order to characterize the state transitions that are independent of $a_{i,t}$, a comprehensive spatial-temporal dependency model is devised to generate TN operation scenarios, as illustrated in subsection 4.3.4.7.

4.3.4.6 Training Algorithm

We develop an MTMARL algorithm based on monotonic value function factorization (QMIX) [39] with RER [175] and KT for training the RTEMA (Algorithm 4.1). The QMIX adopts a centralized training and decentralized execution mechanism with recurrent neural networks (RNNs). In training, apart from individual agent networks, a centralized mixing network is added to evaluate the joint action-value function Q_{tot} of all agents. In execution, the centralized mixing network is removed, and each agent works with its well-trained agent network. Q_{tot} is obtained by

$$\underset{\boldsymbol{a}}{\arg\max}\, Q_{\text{tot}}(\boldsymbol{\kappa}_t, \boldsymbol{a}_t) = \begin{pmatrix} \arg\max Q_1(\kappa_{1,t}, a_{1,t}) \\ \vdots \\ \arg\max Q_I(\kappa_{I,t}, a_{I,t}) \end{pmatrix}, \tag{4.35}$$

$$\frac{\partial Q_{\text{tot}}}{\partial Q_i} \geq 0,\ \forall i, \tag{4.36}$$

where Q_i represents the Q value of agent *i*. $\kappa_{i,t} = \{o_{i,0}, \cdots, o_{i,t}\}$ is the observation history, $\boldsymbol{\kappa} = \{\kappa_1, \cdots, \kappa_I\}$ and $\boldsymbol{a} = \{a_1, \cdots, a_I\}$ are the joint action-observation history and joint action, respectively. The loss function of QMIX can be written as

$$\mathcal{L}_\varphi = \frac{1}{W}\sum_w \min\left(y - Q_{\text{tot}}\left(\boldsymbol{\kappa}_w, \boldsymbol{a}_w, s_w\right)\right)^2, \tag{4.37}$$

where $y = r + \gamma Q'_{\text{tot}}\left(\boldsymbol{\kappa}_w, \arg\max Q_{\text{tot}}\left(\boldsymbol{\kappa}_{w+1}, \boldsymbol{a}_{w+1}, s_{w+1}\right), s_{w+1}\right)$ is the training target of the target network in RL. W is the sample batch size from the replay buffer.

For RNNs, their recurrent state ς is stored in the RER and used to initialize the network when extracted from the buffer. A warm-up process is adopted where a portion of the history trajectory will not be used for network update. This is to mitigate the inaccurate outputs of RNNs in the first few time steps due to representational drift and recurrent state staleness. Therefore, considering the task representation and the warm-up process, the loss of QMIX in (4.37) is extended to a task-based loss

$$\mathcal{L}_\varphi = \frac{1}{HW}\sum_h \sum_w \min\left(y' - Q_{\text{tot}}\left(\boldsymbol{\kappa}_w, \boldsymbol{a}_w s_w; l_r\right)\right)^2, \tag{4.38}$$

where $y' = r + \gamma Q'_{\text{tot}}\left(\boldsymbol{\kappa}_{w+1}, \arg\max Q_{\text{tot}}\left(\boldsymbol{\kappa}_{w+1}, \boldsymbol{a}_{w+1}, s_{w+1}; l_r\right), s_{w+1}; l_r\right)$, l_r is the length of the history trajectory used for updates.

Moreover, to deal with multiple tasks, the knowledge transfer approach with several learning tricks is applied to rapidly and stably learn the decentralized DHESS control policy incorporating common knowledge. These skills are illustrated in subsection 3.3.4.2–3.3.4.3.

Algorithm 4.1 MTMARL–RTEMA Training.

```
Initialize Q network with $\varphi$, routing network with $\varphi_r$, mixing network with
    $\varphi_m$, target network with $\varphi' \leftarrow \varphi$, and replay buffers $B_1, \cdots, B_H$.
For episode = 1, Max do
    Sample a task from the task set to initialize learning environment
    Receive the initial global state $s_0$
    For control interval t=1, T do
        For agent=1, I do
            Receive observation $o_{i,t}$
            $\kappa_{i,t} \leftarrow \kappa_{i,t-1} \cup \{(o_{i,t}, a_{i,t-1})\}$
            Select $a_{i,t}$ with $\mu(a_{i,t} | o_{i,t}, z)$ and $\varepsilon$-greedy, obtain $r_t$ and $o_{i,t+1}$
            from environment
            Store $(\kappa_{i,t}, a_{i,t}, r_t, \kappa_{i,t+1}, \varsigma_{i,t})$ to $B_h$ based on headway task $z_h$
        Sample $W$ transitions $(\kappa_w, a_w, r_w, \kappa_{w+1}, \varsigma_w)$ from each $B_h$
        Obtain the hidden state $\varsigma_{w,l_r}$ with the first stored recurrent state $\varsigma_{w,0}$
        Obtain $Q_{\text{tot}}$ with mixing network by (4.35)–(4.36)
        Calculate loss $\mathcal{L}$ by (3.33)–(3.34) and (4.38), update $\varphi, \varphi_r, \varphi_m, \lambda$
        Soft update: $\varphi' \leftarrow \tau\varphi + (1-\tau)\varphi'$
```

4.3.4.7 Operation Scenario Generation Method

The TN operation scenario is generated by Fig. 4.5. In subsection 3.3.2, the DTM accounts for the intrinsic uncertainty of passenger flows, delays, delay-induced RTTR, and train resistance. In this subsection, we also quantify solar generation uncertainty using the Copula theory. The Copula model of solar generation is similar to (3.22)–(3.24), where DPV power $P_{i,\rho,n}^{\mathrm{PV}}$ replaces variable $N_{i,\rho,n}^{\mathrm{B}}$. Since PV power is closely related to solar irradiance R^{PV}, we simulate the PV power generation by PVsyst, a PV system simulation software with various PV module models.

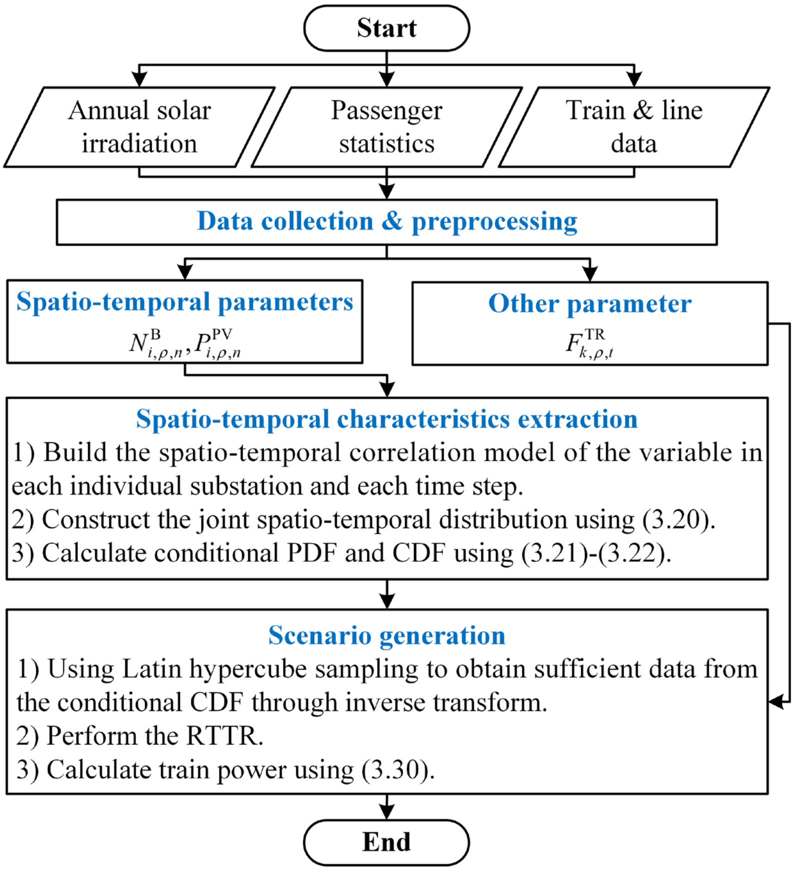

Start
Annual solar irradiation
Passenger statistics
Train & line data
Data collection & preprocessing
Spatio-temporal parameters
$N_{i,\rho,n}^{\mathrm{B}}, P_{i,\rho,n}^{\mathrm{PV}}$
Other parameter
$F_{k,\rho,t}^{\mathrm{TR}}$
Spatio-temporal characteristics extraction
1) Build the spatio-temporal correlation model of the variable in each individual substation and each time step.
2) Construct the joint spatio-temporal distribution using (3.20).
3) Calculate conditional PDF and CDF using (3.21)-(3.22).
Scenario generation
1) Using Latin hypercube sampling to obtain sufficient data from the conditional CDF through inverse transform.
2) Perform the RTTR.
3) Calculate train power using (3.30).
End

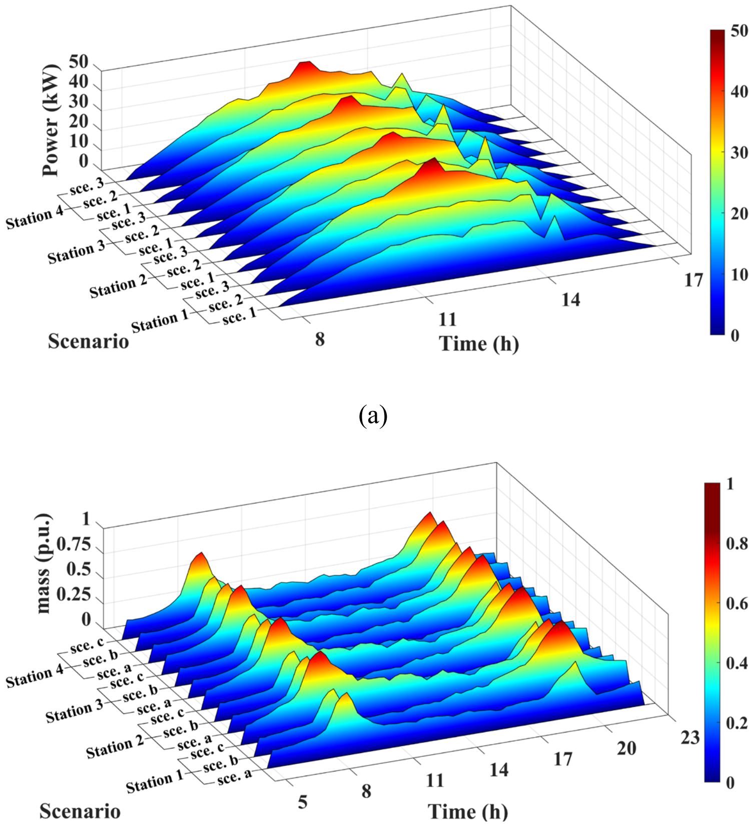
Fig. 4.5 Flowchart of scenario generation.

4.4 Case Study

In this section, a detailed analysis of the aforementioned formulations and algorithms is conducted. First, the configuration method of distributed PVs and DHESSs is illustrated. Then, the optimal day-ahead and intraday scheduling results are demonstrated to illustrate the effectiveness of energy management at the upper and middle levels. Several impact factors are investigated, including the impacts of energy management framework structures and prediction levels of PV and passenger flows. Besides, the optimal real-time control results are demonstrated to illustrate the effectiveness of energy management at the lower level. The impacts of control optimization algorithms and schemes are discussed. The training and test performance of MTMARL–RTEMA is verified by comparing it with other learning-based algorithms. In addition, the impact of different control schemes is also analyzed.

4.4.1 Setup

The description of the subway line and train data used in the case study is presented in subsection 3.2.4.1. The carbon emission quota and cost is obtained from [176], where k_e=1.303 kg/kWh, c_{CO_2}=6.15 \$/t. The day-ahead scheduling is solved by stochastic programming, and 1000 initial scenarios are generated. Since Δn=15 min, the time horizon N=96 to represent 24 h. To decrease the computational cost, K-means clustering is utilized for scenario reduction, and 9 scenarios are retained with probabilities 0.090, 0.107, 0.113, 0.089, 0.106, 0.112, 0.111, 0.132, and 0.140. The PV power and passenger flow profiles of each scenario are shown in Fig. 4.6, and their correlations are shown in Fig. 4.7. The configuration method for PVs is illustrated in subsection 4.4.2, and the passenger flows are generated by the historical OD and arrival rate tables. The intraday scheduling is performed by GA with T_r=1 h. The population size is 40, the crossover fraction is 0.9, and the mutation fraction is 0.1. The intraday PV output and passenger flow predictions are assumed to increase from the day-ahead prediction by 15%. They are added with a random error as the actual value. The random error is subject to Normal distribution with a 5% standard deviation. $c_p = c_s = c_b = 1$.



(a)

(b)

Fig. 4.6 Typical (a) PV output and (b) passenger flow scenarios.

Table 4.1 MTMARL–RTEMA parameters.

Parameter	Value	Parameter	Value	Parameter	Value
ξ_q	10^{-4}	H	3	Buffer capacity	2^{20}
τ	5×10^{-4}	l_r	30	Exploration policy	1→0.05
γ	0.998	W	43	Transition length	60
λ	1→0.05	Optimizer	Adam		

For MTMARL, the parameters of the MTMARL–RTEMA are listed in Table 4.1. The agent network is formulated as follows. Each agent network has two fully connected layers with 64 units and ReLU non-linearity to obtain the inputs of the routing network. Then, these layers are followed by two “soft” layers, where each “soft” layer consists of two modules. Each module has two hidden layers with 64 units and

ReLU non-linearity. These “soft” layers aim to handle the task-level coordination [69]. Moreover, a GRU layer with a 64-dimensional hidden state is added to deal with the partial observability issue of the Dec-POMDP. Finally, the dueling architecture [45] is implemented after the GRU layer to improve the network performance. As for the mixing network, it consists of a hypernetwork [39] and a single hidden layer of 256 units with an ELU nonlinearity. The hypernetwork also has a hidden layer of 256 units with ReLU nonlinearity, which conditions the weight of the mixing network on state s in an arbitrary manner, thus flexibly integrating state s into the Q value estimation.

The agent adopts ε-greedy exploration. Since the learning environment is much more complex than the HESS-integrated traction substation operation in chapter 3, a larger value of exploration rate is reserved at the end of training to encourage policy exploration. Thus, the exploration rate in this case study reduces linearly from 1 to 0.05 and remains constant at 0.05 after 5000 episodes. The knowledge transfer trade-off λ reduces linearly from 1 to 0.05 and remains constant at 0.05 after 2000 episodes. The time scale $\Delta t = 1$ s. Since train services are periodic, the time horizon T is equal to the time of the headway. To implement the RER, each transition that is randomly extracted from the replay buffer is clipped to have the same total time steps of 60. l_r is set as 30 to balance the time steps for recovering the start hidden state and network updates. The initial scenarios are also utilized as the training set for MTMARL–RTEMA, while 400 random scenarios other than the training scenarios are generated to construct the test set. We generate 50 random scenarios for each headway under LDM and HDM, respectively. The task and delay settings are the same as in subsection 3.4.1.

The stochastic programming is conducted by Gurobi 10.0.0, while the genetic algorithm is performed with a Python library PYGAD 3.3.1 using parallel computing.

The MTMARL is implemented by PyTorch 1.12.1. All simulations are performed with Python 3.9.13 on the same device in subsection 2.4.1.

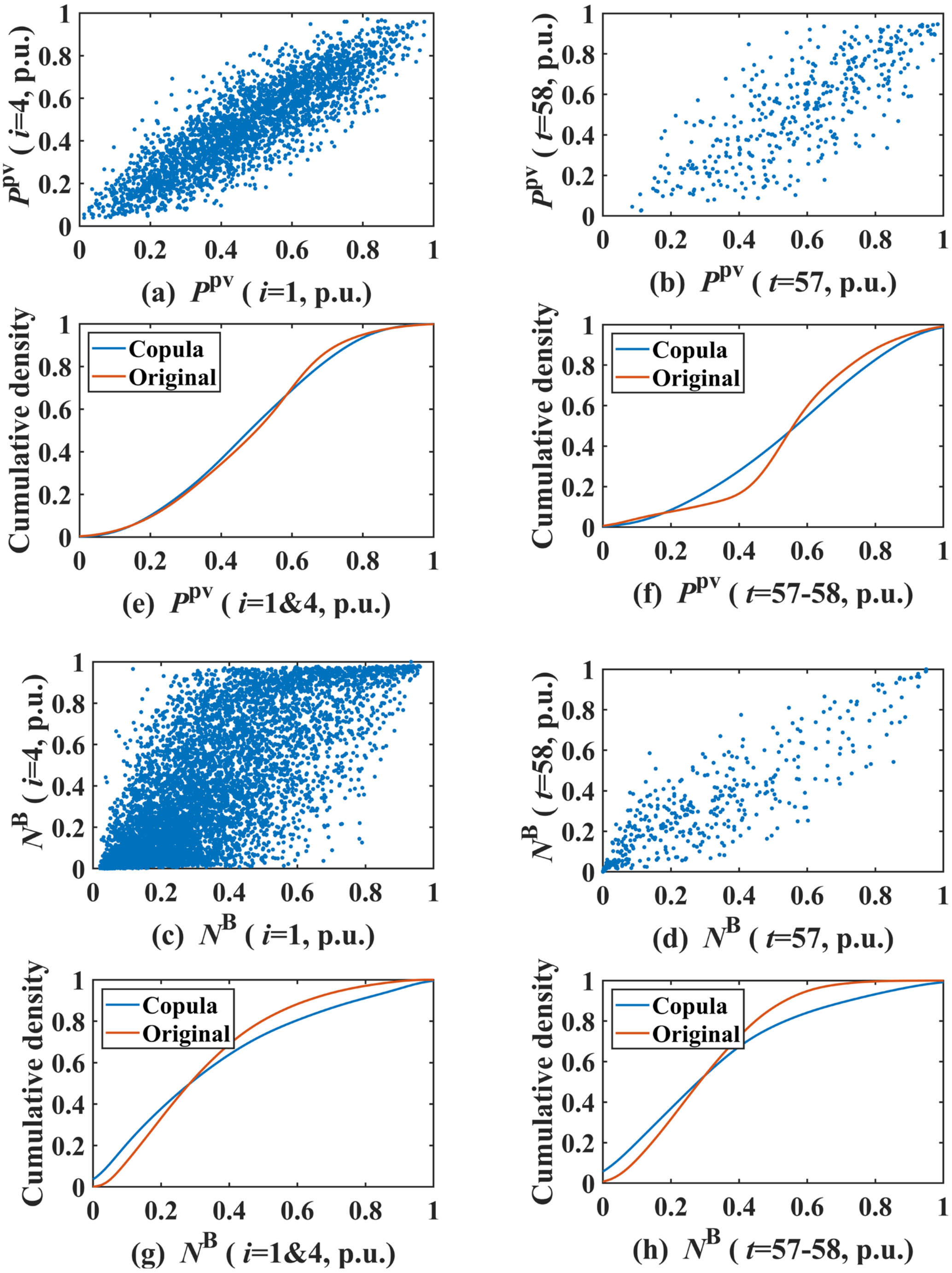
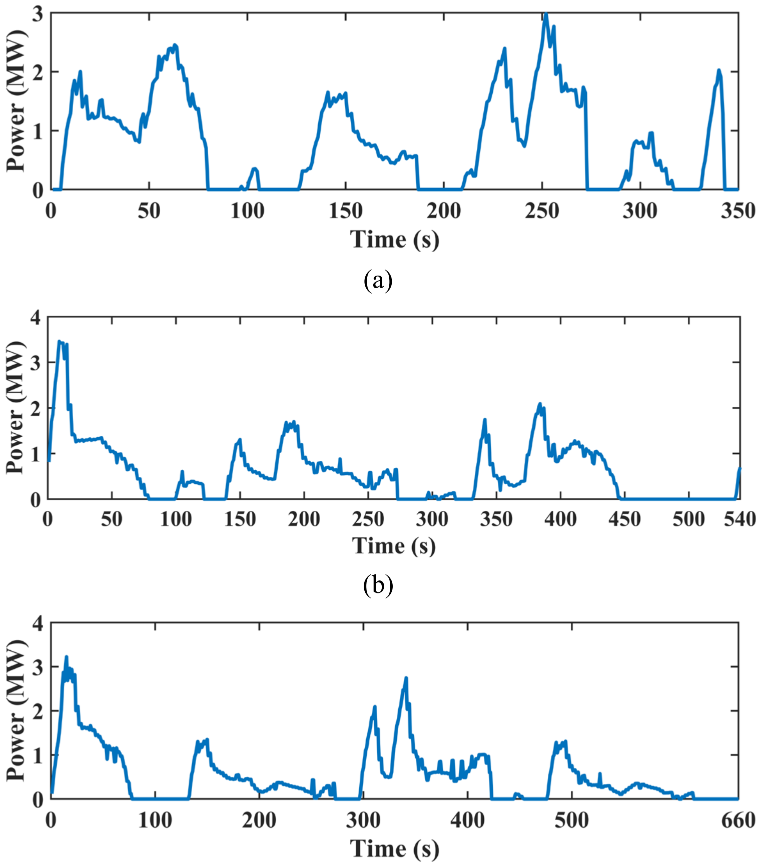


Fig. 4.7 (a)–(b) are PV spatial-temporal correlations, (c)–(d) are passenger spatial-temporal correlations, (e)–(h) are marginal CDFs for (a)–(d). The demonstration uses traction substations 1 and 4, and time periods 57 and 58.

4.4.2 Configuration Method for Distributed Photovoltaics and Hybrid Energy Storage Systems

For simplicity, distributed PVs and DHESSs are assumed to be installed at each station with the same configuration. The configuration of distributed PVs is illustrated as follows. Since no PVs are actually installed in the studied subway line, a reasonable PV size for each station should be considered. We first obtained the latest hourly solar irradiance R^{PV} of nearby areas from a public database [177]. Then, a PV generation model is established. The azimuth angle is set as 0°, and the array is tilted at 30°. The parameters of the PV array are listed in Table 4.2, where the maximum efficiency of the assigned inverter module is 97%. Next, the available area for PV installation is determined according to a recent survey [10], in which the rooftop PV potential assessment of an elevated metro station in a nearby area was studied. Thus, the total rooftop area of all elevated stations in the case study is set at around 2700 m^2, where the skylights account for around 30% of the total rooftop area. Taking the laying gaps and maintenance aisles into account, we assume that 16.4% of the total rooftop area is occupied by PV arrays. Finally, based on the above PV generation model, the average daily distributions of PV power generation at each station can be obtained, and their correlations are calculated. However, considering the spatial grid cell size of solar irradiance data is 2 km×2 km, there are some distances between the subway station and the nearest data point. The spatial correlation calculated from these data points may not represent the correlation at the station's location. Hence, we use an empirical correlation equation [178] to calculate the spatial correlations according to the distances of stations from each other. The economic parameters of PV are obtained from [103, 179].



(a)

(b)

(c)

Fig. 4.8 Total power of all traction substations without DHESSs in typical operation scenarios under (a) 350s, (b) 540s, and (c) 660s headway.

The configuration of DHESSs is illustrated as follows. The economic and technical parameters of supercapacitor and battery modules are listed in subsection 3.2.4.1. Empirically [121], the power of DHESSs should be large enough for peak traction power shaving. Since the supercapacitor generally provides most of the power of the HESS, the peak traction power of the traction substation is roughly selected as the rated power of the supercapacitor. Fig. 4.8 shows traction substation power curves of different train headways in typical operation scenarios without DHESSs. According to the figure, the highest traction substation power occurs when the headway is 540 s, and its value is 3.5 MW. Considering that the battery will also provide power, the total rated power of supercapacitors are set as 2.8 MW, where each station has a supercapacitor of 700 kW. As the rated power is determined, other parameters of the supercapacitor can be obtained according to its type.

Then, since the battery capacity should be large enough to store the daily PV energy, an iterative method is developed, and the procedure is illustrated as follows: 1) Set an initial battery capacity, which is a relatively small value, e.g., 10 kWh. 2) Perform the day-ahead optimization to obtain the PV curtailment powers in each scenario. 3) If the curtailment occurred, increase the battery capacity in steps, and repeat 1)–2). Following the above iterations, the marginal battery capacity that avoids PV curtailment can be obtained. Finally, the DC-DC converter data is obtained from [154]. To match with the converter, the parallel number of supercapacitor and battery modules is 14 and 292, respectively. The maximum C-rate of the battery is limited to 1.25.

Table 4.2 Technical and economic parameters of HESS and PV.

Item	Parameter	
Battery	$E_{\max}^{\mathrm{BT}}$ =160 kWh	$P_{\max}^{\mathrm{BT}}$ =200 kW
	$E_{\max}^{\mathrm{ref}}$ =7.5 kWh	$E_{\min}^{\mathrm{ref}}$ =0.75 kWh
	$\mathrm{SoE}_{\max}^{\mathrm{BT}}$ =0.8	$\mathrm{SoE}_{\min}^{\mathrm{BT}}$ =0.2
	$c_{\mathrm{BT}}^{\mathrm{OM}}$ =1 \$/MWh	η^{BT} =0.8
Supercapacitor	$E_{\max}^{\mathrm{SC}}$ =6 kWh	$P_{\max}^{\mathrm{SC}}$ =700 kW
	$\mathrm{SoE}_{\max}^{\mathrm{SC}}$ =0.9	$\mathrm{SoE}_{\min}^{\mathrm{SC}}$ =0.25
	R^{SC} =11 mΩ	C^{SC} =94 F
	η^{SC} =0.95	$I_{\max}^{\mathrm{SC}}$ =1040 A
		$c_{\mathrm{SC}}^{\mathrm{OM}}$ =7.5 \$/MWh
PV	A^{PV} =443 m^2	η^{PV} =13.58 %
	$C_{\mathrm{D}}=\begin{cases} c_{\mathrm{CUR}}, & \Delta E_{i,\rho,t}^{\mathrm{ref}}>0, \\ c_{\mathrm{SUB}}, & \Delta E_{i,\rho,t}^{\mathrm{ref}}\le 0. \end{cases}$	$c_{\mathrm{PV}}^{\mathrm{OM}}$ =4.5 \$/MWh
		c_{CUR} =0.26 \$/kWh

4.4.3 Analysis of Day-Ahead and Intraday Scheduling Results

4.4.3.1 Day-Ahead & Intraday Scheduling Results

In this subsection, the optimal day-ahead and intraday schedules are demonstrated. First, Fig. 4.9 shows the day-ahead power balance. It can be observed that the maximum traction power demand occurs during peak hours 8:00–9:00 and 17:00–19:00. Generally, DHESSs charge during off-peak hours between 9:00–16:00 and discharge during peak hours between 16:00–19:00. As a result, the peak substation power during evening peak hours can be reduced. Besides, a substantial amount of train braking energy is directly utilized to supply traction load.

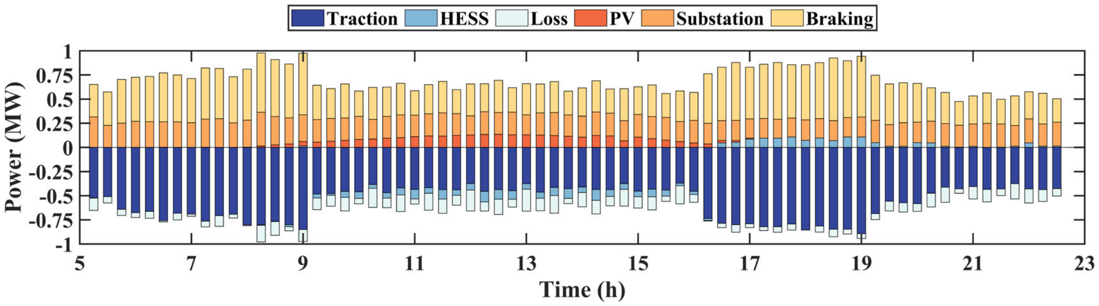
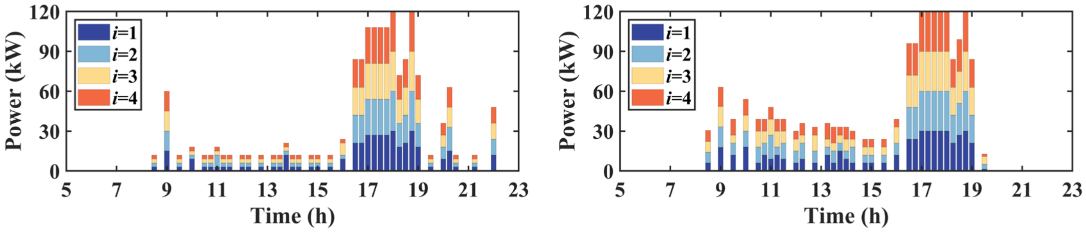


Fig. 4.9 Day-ahead power balance.

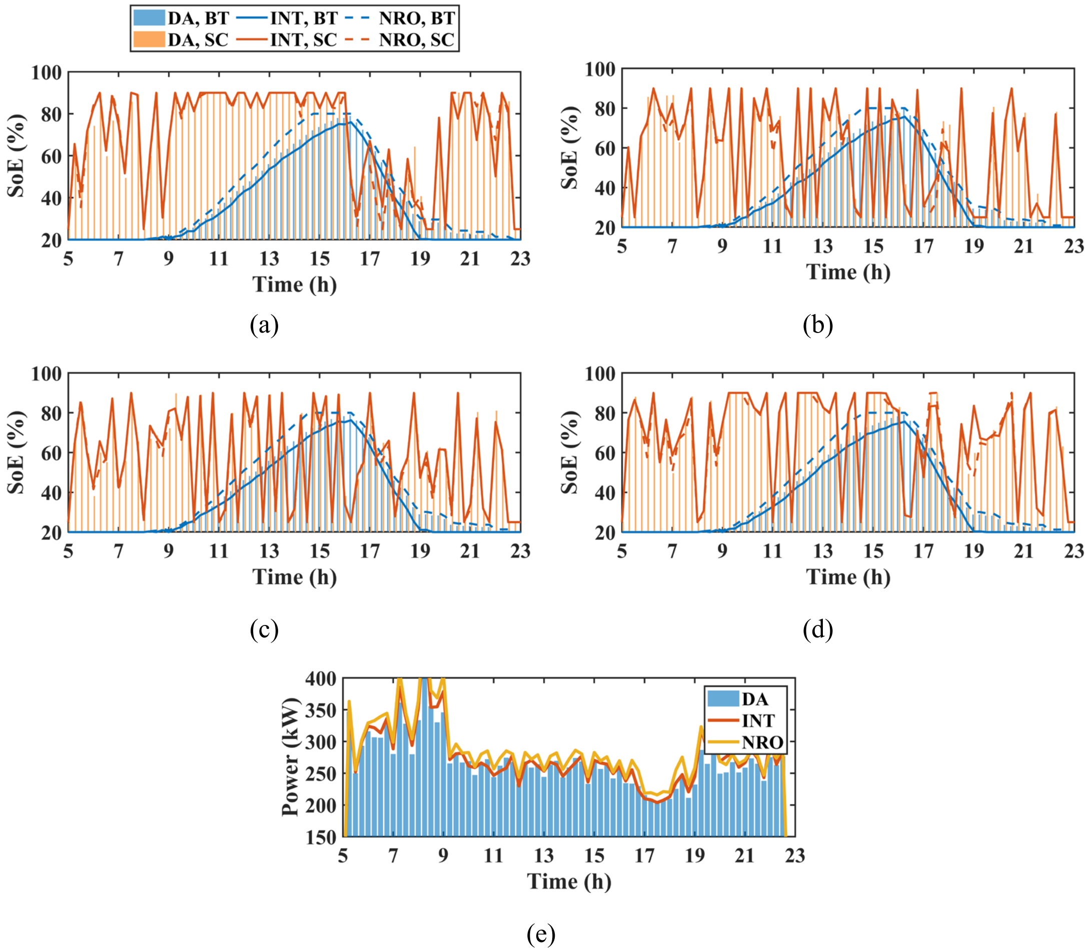


(a)

(b)

Fig. 4.10 Optimal (a) day-ahead and (b) intraday PV-battery powers.

Then, the optimal intraday schedule is demonstrated, where Fig. 4.10 illustrates the day-ahead and intraday referential PV-battery energy E^{ref} at different traction substations. Since the intraday prediction of PV outputs is different from the day-ahead perdition, if the operation of the PV-battery is based on the day-ahead schedule, curtailment of PV power can occur, with consequent increasing operation cost. From the figure, it can be observed that, according to the correction, E^{ref} during peak hours are generally larger than during off-peak hours to meet the load demand. E^{ref} is zero in some time steps, which represents LDM. HDM is in effect for other situations. LDM and HDM are switched optimally to minimize the overall daily operation cost. Since the intraday predictions increase, the intraday E^{ref} is larger than its day-ahead counterparts in general to avoid PV curtailment.



(a)

(b)

(c)

(d)

(e)

Fig. 4.11 SoEs at traction substations 1–4 (a)–(d), and (e) total substation power[1].

[1] DA: day-ahead, INT: intraday, NRO: no rolling optimization, BT: battery, SC: supercapacitor.

Table 4.3 Comparative correction performances.

Scenario	MRAE of SoE		MRAE of substation power	
	INT (%)	NRO (%)	INT (%)	NRO (%)
1	3.90	6.55	6.09	6.37
2	4.27	6.68	6.59	7.35
3	4.33	6.78	7.05	8.02
4	4.21	6.80	5.35	6.24
5	4.48	7.03	5.81	7.00
6	4.35	7.27	6.67	7.82
7	4.70	7.21	5.38	6.11
8	4.94	7.61	6.34	7.09
9	4.80	7.28	6.21	7.51
Avg.	4.44	7.02	6.17	7.06

To further illustrate the performance of the intraday operation deviation correction, Fig. 4.11 compares the intraday correction terms (DHESS SoEs and total substation power) with and without the middle-level energy management. From the figure, the operation deviations resulting from passenger flow and PV generation uncertainties are smoothed by the intraday correction. The mismatch between actual energy consumption and the schedule plan is compensated. In order to quantify the correction performance, the mean relative absolute error (MRAE) is introduced to calculate the average of the absolute value of the difference between the intraday and day-ahead variables.

$$\mathrm{MRAE_{SoE}} = \frac{1}{N}\left(\sum_{i=1}^{I}\sum_{n=1}^{N} \frac{\left|\mathrm{SoE}_{i,\rho,n}^{\mathrm{SC,INT}} - \mathrm{SoE}_{i,\rho,n}^{\mathrm{SC}}\right|}{\mathrm{SoE}_{i,\rho,n}^{\mathrm{SC}}} + \sum_{i=1}^{I}\sum_{n=1}^{N} \frac{\left|\mathrm{SoE}_{i,\rho,n}^{\mathrm{BT,INT}} - \mathrm{SoE}_{i,\rho,n}^{\mathrm{BT}}\right|}{\mathrm{SoE}_{i,\rho,n}^{\mathrm{BT}}} \right), \tag{4.39}$$

$$\mathrm{MRAE}_{P^{\mathrm{SUB}}} = \frac{1}{N}\sum_{i=1}^{I}\sum_{n=1}^{N} \frac{\left|P_{i,\rho,n}^{\mathrm{GRID,INT}} - P_{i,\rho,n}^{\mathrm{GRID}}\right|}{P_{i,\rho,n}^{\mathrm{GRID}}}. \tag{4.40}$$

From Table 4.3, the proposed rolling optimization in the intraday can effectively track day-ahead scheduling plans. The MRAE of SoE and traction substation power can be reduced by 36.75% and 12.61% compared with non-rolling optimization.

4.4.3.2 Impact of Energy Management Framework Structures

In this subsection, the effectiveness of the proposed tri-level structure is verified, where the following frameworks are compared: *1) The proposed framework (PF). 2) Real-time PV-battery output control (RPOC):* A real-time energy management strategy similar to [131], while we add PVs and passenger flows for comparison purposes. *3) Day-ahead scheduling (DAS):* Similar to [180], while their model is extended to involve multiple stations for comparison purposes. To simulate the situation of no control strategy, the charge/discharge thresholds are fixed and close to the no-load voltage, which maximizes the DHESS usage. $U_{i,\rho,t}^{\mathrm{CH}}$ =865 V, $U_{i,\rho,t}^{\mathrm{DIS}}$ =855 V. Meanwhile, $P_{i,\rho,t}^{\text{PV-BT}} = P_{\max}^{\mathrm{BT}}$ regardless of operation modes. *4) Day-ahead and intraday scheduling (DAIS):* Similar to [134], while we add passenger flows and train service patterns for comparison purposes. *5) Maximum PV-battery output (MPO):* the basic framework which utilizes the maximum PV-battery output at each time step.

Table 4.4 Comparative performances of frameworks 1–5 under prediction level 1.

Item	PF	RPOC	DAS	DAIS	MPO
Overall cost ($)	595.81	642.50	703.15	674.89	682.95
Purchasing cost ($)	502.15	515.02	610.01	601.82	599.21
Carbon trading cost ($)	14.18	14.54	17.22	16.99	16.92
O&M cost ($)	58.13	55.21	46.10	46.10	55.23
Curtailment cost ($)	21.35	57.03	29.81	9.99	11.59
PV–RB energy utilization (%)	80.81	78.49	67.31	68.45	68.66

Table 4.5 Comparative performances of frameworks 1–5 under prediction level 2.

Item	PF	RPOC	DAS	DAIS	MPO
Overall cost ($)	610.16	650.17	727.33	696.76	695.40
Purchasing cost ($)	513.74	524.62	623.40	614.44	609.62
Carbon trading cost ($)	14.50	14.81	17.60	17.35	17.21
O&M cost ($)	59.11	56.49	47.66	47.66	56.99
Curtailment cost ($)	22.81	54.25	38.67	17.30	11.57
PV–RB energy utilization (%)	80.56	78.55	67.51	68.71	69.25

Table 4.6 Comparative performances of frameworks 1–5 under prediction level 3.

Item	PF	RPOC	DAS	DAIS	MPO
Overall cost ($)	551.92	591.59	628.75	626.18	636.62
Purchasing cost ($)	473.15	481.70	571.32	569.87	566.98
Carbon trading cost ($)	13.36	13.60	16.13	16.09	16.01
O&M cost ($)	51.48	48.81	38.75	38.75	47.00
Curtailment cost ($)	13.94	47.48	2.54	1.47	6.63
PV–RB energy utilization (%)	83.01	80.38	67.38	67.47	67.47

Table 4.7 Comparative performances of frameworks 1–5 under prediction level 4.

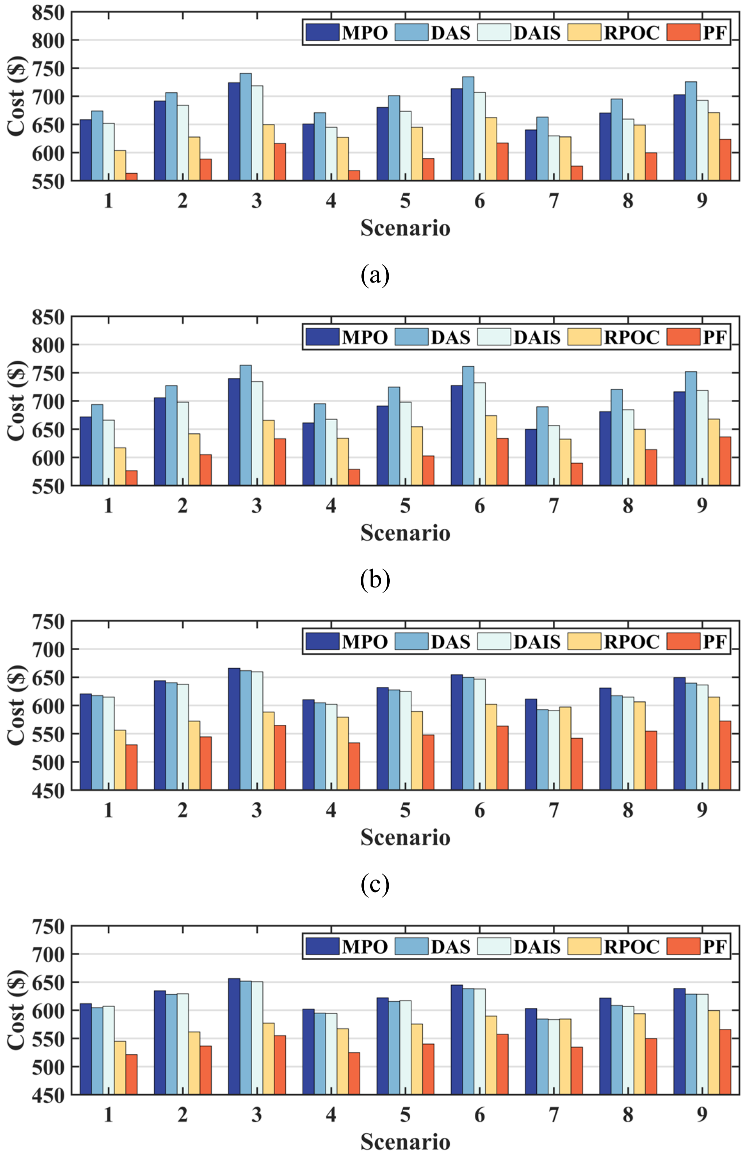
Item	PF	RPOC	DAS	DAIS	MPO
Overall cost ($)	544.61	579.12	618.14	618.10	627.27
Purchasing cost ($)	466.74	473.74	563.20	563.00	560.05
Carbon trading cost ($)	13.18	13.38	15.90	15.90	15.81
O&M cost ($)	50.21	47.79	37.53	37.53	45.57
Curtailment cost ($)	14.48	44.21	1.50	1.68	5.83
PV–RB energy utilization (%)	83.14	80.63	67.09	67.13	67.13

In Fig. 4.12(a). compared with MPO, frameworks 1–4 achieve average cost-savings of 12.76%, 5.92%, -2.96%, and 1.18%, respectively. For DAS and DAIS, due to the lack of consideration of DHESS control in the short time scale, their average operation cost is at least 13.27% higher than PF. Besides, the lack of day-ahead and intraday scheduling plans in RPOC will also significantly increase the operation cost compared with PF. Thus, leveraging the synergistic consideration of energy management on both time scales, the proposed approach achieves more economic and low-carbon TN operation than other frameworks.

4.4.3.3 Impact of Prediction Levels

In this subsection, the impact of various prediction levels of PV outputs and passenger flows is analyzed. Based on the common range of prediction errors of PV outputs [181] and passenger flows [182], four prediction levels are set for comparison. For prediction levels 1–4, the intraday predictions increase from the day-ahead predictions by 15%, 20%, -15%, -20%, respectively. Then, a random error is added to generate the actual values. The random error is subject to Normal distribution with a 5% (for levels 1 and 3) and 10% (for levels 2 and 4) standard deviation, respectively.

As shown in Table 4.4–Table 4.7 and Fig. 4.12, DAS and MPO generally have the worst performance because they cannot deal with real-time uncertainties effectively. Besides, PF achieves the lowest overall daily cost across all prediction levels, which is 11.98% (11.72–12.43%) lower compared with DAIS. According to the figures, it can also be observed that PF has the best economic performance in all scenarios. In addition, PF also achieves the highest PV–RB energy utilization and is 13.94% (11.85–16.01%) higher than DAIS on average. Furthermore, the PV–RB utilization of RPOC and PF is significantly higher than other approaches, indicating the effectiveness of real-time PV–RB power flow optimization. Thus, the effectiveness of the proposed approach under various perdition levels of PV outputs and passenger flows is verified.



(a)

(b)

(c)

(d)

Fig. 4.12 Comparative operation costs of frameworks 1–5 under prediction levels 1–4 (a)–(d).

4.4.4 Analysis of Real-Time Control Results

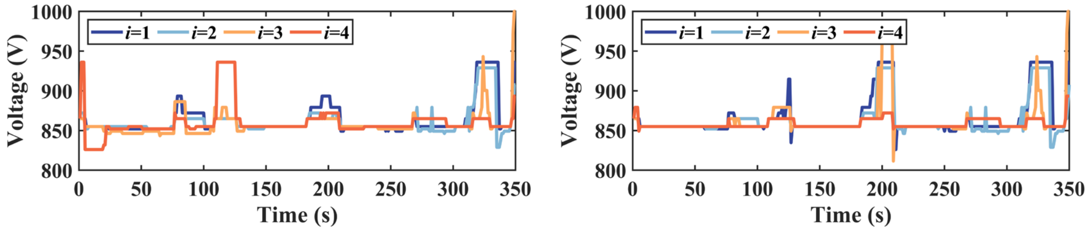
4.4.4.1 Real-Time Control Results

In this subsection, the real-time control behaviors of the TN are demonstrated. First, the DHESS and PV operation under the normal operation condition is analyzed, where Fig. 4.13 shows the DHESS charge/discharge voltage thresholds, and Fig. 4.14

presents the charging (-) and discharging (+) powers of each HESS, as well as the traction substation powers. For illustration purposes, the initial DHESS SoEs are set as the minimum SoE. It can be observed that the voltage thresholds and PV-battery power of each HESS are adjusted dynamically according to real-time TN operation states. The DPVs and DHESSs release energy to reduce traction substation output and shave peak power during 0–50 s, 140–170 s, 220–260 s, and 330–350 s. Especially under HDM, the PV power complements the deficiency of RB power generation during 0–50 s, further reducing the traction substation power supply. In addition, a more significant amount of supercapacitor energy is released during 225–260 s under HDM. This is because the use of PV-battery power during 0–200 s reserves more available capacity of the supercapacitor.

To further analyze the capacity reservation of supercapacitors under HDM, Fig. 4.15 demonstrates the SoE curves of supercapacitors. From the figure, due to the use of PV-battery power under HDM, the SoEs of supercapacitors at stations 3–4 increase sharply during 100-200 s, indicating a large amount of energy is reserved for later use. Nevertheless, the SoEs at stations 1–2 decrease. To explain the reason, the train power and displacement curves are drawn in Fig. 4.16. From the figure, during this time period, train 1 is braking at sections 3–4. Train 2 stops at station 2 and then accelerates. Train 3 accelerates at sections 3–2. Therefore, the RB energy is mainly distributed between stations 3 and 4, while the traction demand is mainly distributed between stations 2 and 3. Considering the contact line loss due to distance, it is reasonable to deliver the RB energy to the supercapacitor at station 4 and supply the traction demand by supercapacitors at stations 1–3. As PV-battery power covers part of the traction demand under HDM, more RB energy is delivered to the supercapacitor at stations 3 and 4.

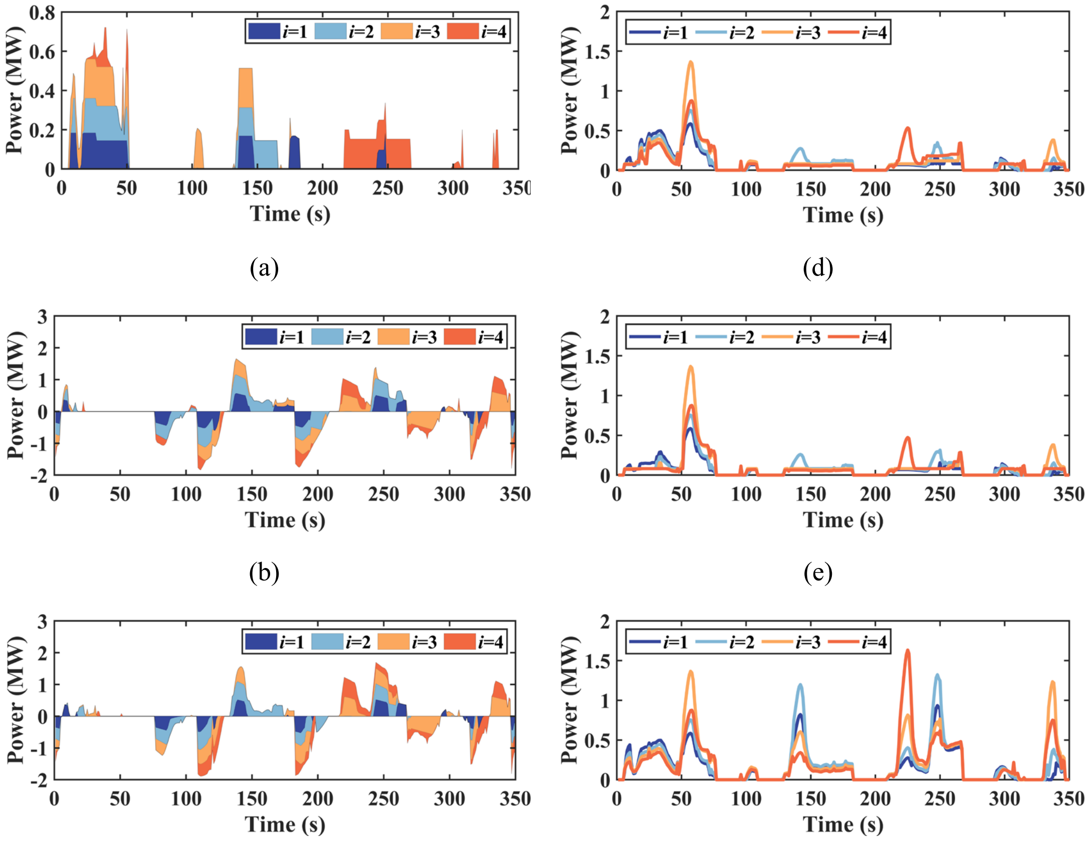
From the above analysis, by implementing the proposed real-time control over supercapacitors and PB-batteries, the DHESS capacity is flexibly adjusted to reserve more RB energy, satisfy the traction load demands, and reduce contact line losses, which ensures the network-wide cost efficiency.



(a)

(b)

Fig. 4.13 Voltage thresholds of HESSs under (a) LDM and (b) HDM.



(a)

(d)

(b)

(e)

(c)

(f)

Fig. 4.14 DHESS and substation powers. (a) PV-battery powers under HDM (under LDM, the powers are always zero), supercapacitor powers under (b) LDM and (c) HDM, substation powers under (d) LDM, (e) HDM, and (f) no PV-DHESS.

4.4.4.2 Impact of Control Optimal Algorithms

In this subsection, the effectiveness of the MTMARL–RTEMA on single-task and multi-task learning is discussed. First, the following algorithms are compared to illustrate the model performance on single-task learning. *1) MTMARL–RTEMA-1:* Coordinated DHESS control is considered, while only adjacent train information and local traction substation information are utilized. RER is used (Proposed). *2) QMIX:* the original QMIX is used while the RER is removed. *3) MTMARL–RTEMA-2:* Similar to MTMARL–RTEMA-1, however, traction substation information can be exchanged and shared by all agents, as shown in (4.41). *4) IQL:* No mixing network to coordinate DHESSs, and each HESS independently takes actions based on local observations [39]. For single-task learning, each algorithm trains on a single task till convergence and uses well-trained agents to evaluate the model performance on this single task. Then, new training starts with another task and generates new agents for evaluation on this task. Repeat the above process till all tasks are evaluated. Finally, the evaluation of each task is averaged to represent the performance of each algorithm. To better illustrate the results, delay scenarios are not included in the single-task environment but are added in the multi-task environment.

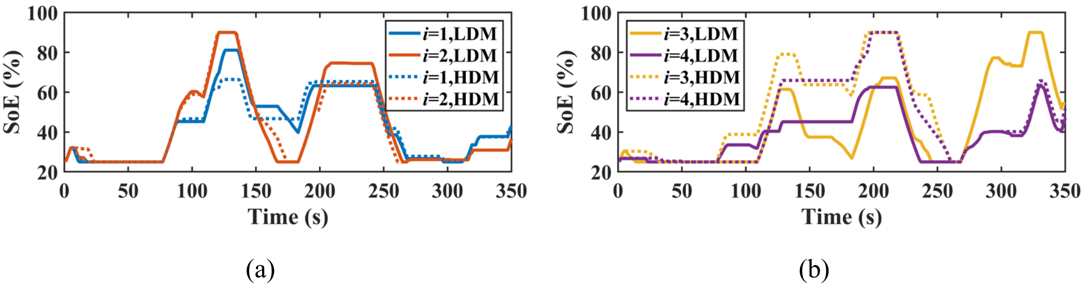
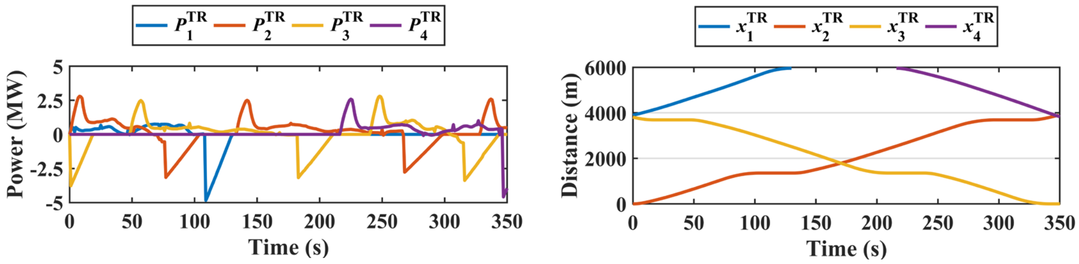


Fig. 4.15 Supercapacitor SoEs under LDM and HDM, at stations (a) 1–2 and (b) 3–4.



(a) (b)

Fig. 4.16 Curves of train (a) power and (b) displacement.

$$o_{i,t}=\begin{cases}\left\{x_{i-1,k,\rho,t}^{\mathrm{TR}},d_{i-1,k,\rho,t}^{\mathrm{TR}},P_{i-1,k,\rho,t}^{\mathrm{TR}},x_{i+1,k,\rho,t}^{\mathrm{TR}},d_{i+1,k,\rho,t}^{\mathrm{TR}},P_{i+1,k,\rho,t}^{\mathrm{TR}}\right\},\\ \left\{\begin{matrix}\mathrm{SoE}_{1,\rho,t}^{\mathrm{SC}},\ldots,\mathrm{SoE}_{I,\rho,t}^{\mathrm{SC}},\Delta E_{1,\rho,t}^{\mathrm{ref}},\ldots,\Delta E_{I,\rho,t}^{\mathrm{ref}},\\ U_{1,\rho,t}^{\mathrm{SUB}},\ldots,U_{I,\rho,t}^{\mathrm{SUB}},I_{1,\rho,t}^{\mathrm{SUB}},\ldots,I_{I,\rho,t}^{\mathrm{SUB}}\end{matrix}\right\}.\end{cases} \tag{4.41}$$

The performance of algorithms 1–4 in the single-task learning environment is depicted in Table 4.8 and Fig. 4.17. The figure shows the training episodic reward curves, where five random runs are performed. Generally, QMIX obtains a much lower reward compared with MTMARL–RTEMA-1, which verifies the effectiveness of using the RER. Besides, IQL obtains the second lowest reward, and its curve shows a declining trend at the end. This is because IQL only focuses on local observation while disregarding the actions of other agents, which causes training instability. Meanwhile, the other three algorithms overcome the stability issue by coordinately considering the actions of all agents. MTMARL–RTEMA-1 obtains a similar high reward and utilization compared to MTMARL–RTEMA-2. It indicates that high-quality decisions can be made without global traction substation information, which verifies the effectiveness of the Dec-POMDP design. From the table, MTMARL–RTEMA-1 achieves the highest PV–RB energy utilization. Thus, the proposed Dec-POMDP design, the necessity of DHESS coordination, and the performance of MTMARL–RTEMA-1 under the single-task learning environment are verified.

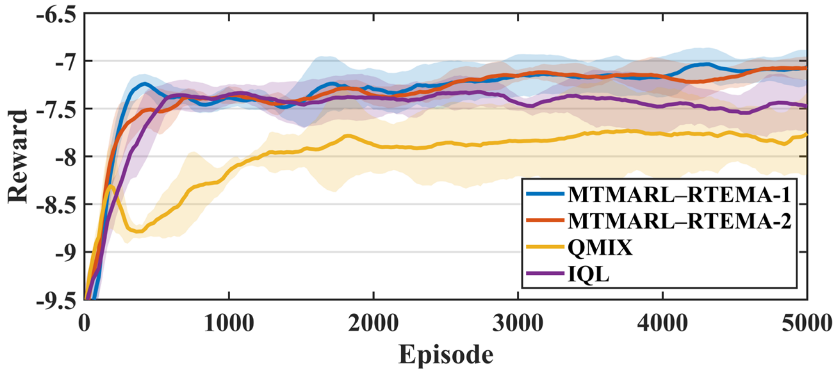


Fig. 4.17 Comparative rewards of algorithms 1–4 in single-task learning environment. Solid lines are moving averages over 200 episodes. Shaded areas are one standard deviation ranges.

Table 4.8 Comparative performances of algorithms 1–4.

Item	Algorithm	Avg.	Headway (s)		
			350	540	660
Braking & PV energy (kWh)	-	92.43	88.54	94.87	93.87
Braking resistor loss (kWh)	MTMARL–RTEMA-1	22.94	7.73	27.45	33.65
	MTMARL–RTEMA-2	22.58	7.26	24.79	35.70
	QMIX	24.45	9.33	28.94	35.08
	IQL	29.09	11.26	37.02	38.98
Curtailment (kWh)	MTMARL–RTEMA-1	0.24	0.00	0.00	0.72
	MTMARL–RTEMA-2	0.35	0.00	0.89	0.17
	QMIX	1.62	0.00	0.55	4.32
	IQL	3.90	2.88	6.48	2.33
PV–RB energy utilization (%)	MTMARL–RTEMA-1	75.24	91.27	71.07	63.38
	MTMARL–RTEMA-2	75.51	91.80	72.93	61.79
	QMIX	72.13	89.46	68.91	58.03
	IQL	64.72	84.03	54.15	55.99

Furthermore, the following algorithms are compared to illustrate the model performance on multi-task learning. *1) MTMARL–RTEMA-1:* (Proposed). *5) MTMARL–RTEMA-noKT:* The KT function is removed from the proposed algorithm to verify its effectiveness. *6) QMIX-MT:* Since the original QMIX algorithm can only be used in single-task learning. We extend the QMIX into the multi-task learning framework by adding the soft module and conflict gradient projecting techniques. Thus, the QMIX-MT is essentially the MTMARL–RTEMA-1 without RER and KT.

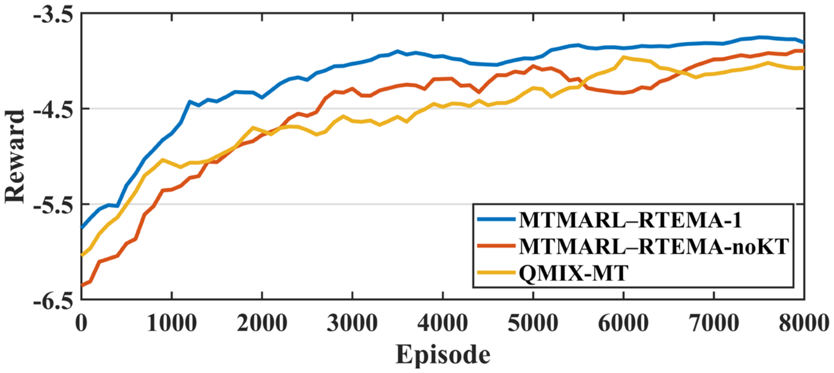


Fig. 4.18 Comparative rewards of algorithms 1, 5, and 6 in multi-task learning environment. Solid lines are moving averages over 200 episodes.

The performance of algorithms 1, 5, and 6 in the multi-task learning environment is depicted in Fig. 4.18. Similarly, the figure shows the training episodic reward curves. It can be observed that, generally, the rewards in the multi-task learning environment is higher than the single-task learning environment. This is because the added delay scenarios in the multi-task learning environment have short train headways (350 s) and their scenario costs are usually lower than 540 s and 660 s headways. Based on (4.31), since the average scenario cost of multi-task learning environment is lower, its reward is usually higher than the single-task learning environment. Moreover, due to the difficulty of learning from multiple tasks simultaneously, the algorithms take more episodes to converge. Furthermore, the performance of MTMARL–RTEMA-1 is superior to MTMARL–RTEMA-noKT and QMIX-MT, and QMIX-MT has the worst performance. Hence, the effectiveness of KT and RER under the multi-task learning environment is verified.

The scalability of the proposed approach is analyzed in Table 4.9. with the increase of agents, the total training time increases almost linearly, while the average test time per agent remains almost unchanged (<1 ms), validating excellent scalability of the proposed approach.

Table 4.9 Scalability analysis.

Number of agents	Total training time (h)	Average test time per agents (ms)
4	1.03	0.79
12	2.39	0.76
20	3.59	0.79

4.4.4.3 Impact of Control Schemes

In this subsection, the impact of different control schemes on the overall operation cost is investigated. Four control strategies are compared to evaluate the decision-making performance and generalization capability of MTMARL–RTEMA by the test set (Table 4.10). *1) Proposed control strategy (MTMARL–RTEMA)*. *2) Fixed threshold (FT):* The charge/discharge thresholds are fixed to be $U_{i,\rho,t}^{\mathrm{CH}}$=865 V and $U_{i,\rho,t}^{\mathrm{DIS}}$=855 V, which aims to encourage maximum DHESS usage. *3) Fixed power allocation (FPA):* The PV-battery power is set to be $P_{i,\rho,t}^{\text{PV-BT}} = P_{\max}^{\mathrm{BT}}$, which aims to use PV-battery power as much as possible. *4) No control strategy (NCS):* As illustrated in framework DAS. Besides, the situation with no PV and HESS (*NPH*) are also included for comparison.

From the table, since LDM does not involve any PV consumption, only threshold control is in effect under LDM. Hence, the performance of FPA and MTMARL–RTEMA under LDM is the same. MTMARL–RTEMA achieves the lowest average operation cost in the test set, which verifies its generalization capability on different train service patterns and network uncertainties. Compared with FT, MTMARL–RTEMA exerts coordinated control over the voltage threshold and PV–RB power allocation of the DHESS, thereby significantly reducing the deviation of real-time PV consumption to the intraday scheduling decision and slightly decreasing the operation cost. Since the excessive usage of the PV-battery results in a high deviation cost, the operation cost of FPA under HDM and NCS is higher than MTMARL–RTEMA.

Table 4.10 Comparative performances of control schemes 1–5.

LDM ($)	MTMARL–RTEMA	NCS	FPA	FT	NPH
Avg. operation cost	4.79	6.38	4.79	4.89	8.42
Avg. purchasing cost	4.31	3.27	4.31	4.39	8.42
Avg. O&M cost	0.48	0.36	0.48	0.50	0.00
Avg. deviation cost	0.00	2.75	0.00	0.00	0.00
Min. operation cost	3.11	4.71	3.11	3.21	6.52
Max. operation cost	6.69	8.30	6.69	6.55	11.00
HDM ($)	MTMARL–RTEMA	NCS	FPA	FT	NPH
Avg. operation cost	4.25	5.31	5.39	4.31	8.42
Avg. purchasing cost	3.74	3.27	3.53	3.66	8.42
Avg. O&M cost	0.51	0.49	0.47	0.54	0.00
Avg. deviation cost	0.00	1.55	1.39	0.11	0.00
Min. operation cost	2.76	3.61	3.61	2.83	6.52
Max. operation cost	6.12	7.01	7.01	6.18	11.00

4.5 Summary

In this chapter, an MTMARL–MTSEM approach is proposed for the economic and low-carbon operation of TNs integrated with PV–RB DHESSs. The research mainly includes the following aspects.

A tri-level energy management framework is developed, where the upper level minimizes daily operation and carbon trading costs, the middle level corrects day-ahead scheduling deviations against multi-source uncertainties, and the lower level proposes an MTMARL–RTEMA to address the DHESS control problem. A Copula-based spatial-temporal dependency model is devised to characterize uncertainties of PVs, passenger flows, and traction loads. Finally, comparative studies demonstrate the effectiveness of the proposed framework in terms of cost reduction and PV–RB energy utilization improvement.

The key findings of the designated case study are summarized as follows: 1) the proposed framework with multiple time scales can outperform other energy management frameworks on system operational economy (11.98% reduction) and renewable energy utilization (13.94% increase) compared to the conventional long-time-scale energy management framework. 2) The MTMARL–RTEMA can coordinate voltage threshold and PV–RB power allocation of DHESSs, and its PV–RB energy utilization can be increased by 10.31% compared to the uncoordinated control scheme. 3) The generalization capability of MTMARL–RTEMA is verified under train operation scenarios.

Chapter 5: Data-Driven Multi-Objective Configuration Optimization for Distributed Hybrid Energy Storage System-Integrated Traction Network Operation Based on Multi-Task Multi-Agent Reinforcement Learning

Nomenclature in this chapter

A. Multi-Task Multi-Agent Reinforcement Learning Elements

$\mathcal{Z}$, $\mathcal{Z}_H$, $\mathcal{Z}_P$, $\mathcal{Z}_D$, $\mathcal{Z}_{RB}$	Set of tasks, headway tasks, combination tasks of trajectories, dwell time tasks, and RB parameter tasks

B. NSGA-II Elements

F	Fitness function
g_{iter}	Generation
$N^{\text{ga}}, S^{\text{ga}}$	Number and set of solutions being dominated by a specific solution
N^{pop}	Number of solutions
$p^{\text{ga}}, q^{\text{ga}}$	Pareto solution and solution which is dominated by p^{ga}
PA, OS	Parent and off-spring populations
x_{ga}	Decision variable
P^{*}, χ, Θ	Optimal solution, weight term, and entropy of information of multi-criteria decision-making
P^{-}, P^{+}	Positive and negative ideal distances

C. EL Elements

$c_{\text{el},1}, c_{\text{el},2}$	XGBoost parameters
f	Functional space that represents a set of regression trees

T_{leaves}	Number of leaves
x_{el} , y_{el} , $\hat{y}_{\mathrm{el}}$	Feature, label, and estimated label of training samples
ω	Regularization term
$\mathcal{L}_{EL}$, $\mathcal{L}_{L}$	Losses of EL and regression tree
λ	Weight of leaves
ξ_e	Learning rate of ensemble models

D. Indices

$d \in \{1,2,\cdots,D\}$	Index of training samples
$h \in \{1,2,\cdots,H\}$	Index of train headways
$i \in \{1,2,\cdots,I\}$	Index of stations or traction substations
$j \in \{1,2,\cdots,J\}$	Index of stations or traction substations except for station or traction substation i
$k \in \{1,2,\cdots,K\}$	Index of trains
$m \in \{1,2,\cdots,M\}$	Index of regression trees
$o \in \{1,2,\cdots,O\}$	Index of objectives
$w \in \{1,2,\cdots,W\}$	Index of Pareto solutions

E. Time Scales

Δn , n , N	Increment, current time step, and time horizon on a long time scale for economic dispatch and prediction (e.g., sub-hourly or hourly)
Δt , t , T	Increment, current time step, and time horizon on a short time scale for real-time train and HESS control (e.g., sub-minutely)

F. Variables

A^{PV}	Area of PV array (m^2)
E^{ref}	Referential PV-battery energy (kWh)

$E^{\mathrm{SUB}}, E^{\mathrm{PV}}, E^{\mathrm{CUR}}$	Energies of traction substations, PV, and PV curtailment (kWh)
$E^{\mathrm{SC}}, E^{\mathrm{RB}}, E^{\mathrm{BR}}$	Energies of supercapacitors, RB, and braking resistor (kWh)
$J^{\mathrm{LCC}}, J^{\mathrm{EU}}, J^{\mathrm{TT}}$	Economic, energy utilization, and travel time indicators
$J^{\mathrm{INV}}, J^{\mathrm{REP}}, J^{\mathrm{FIX}}$	Costs of investment, replacement, and installation ($)
$J^{\mathrm{SUB}}, J^{\mathrm{CA}}, J^{\mathrm{CUR}}$	Costs of electricity trading, carbon trading, and PV curtailment ($)
$J^{\mathrm{OM}}, J^{\mathrm{SC}}, J^{\mathrm{BT}}, J^{\mathrm{PV}}$	Costs of PV-HESS, supercapacitor, battery, and PV operation ($)
L_{BT}	Estimated battery lifetime (year)
$N_{\mathrm{BT}}, N_{\mathrm{SC}}, N_{\mathrm{DC}}$	Number of battery, supercapacitor, and converter modules
$N_{\mathrm{BT}}^{\mathrm{P}}, N_{\mathrm{SC}}^{\mathrm{P}}$	Number of in-parallel battery and supercapacitor modules
$P^{\mathrm{SUB}}, P^{\mathrm{PV}}, P^{\mathrm{RB}}, P^{\mathrm{BR}}$	Powers of traction substation, PV, RB, and braking resistor (W)
$P^{\mathrm{SC,CH}}, P^{\mathrm{SC,DIS}}$	Charging and discharging powers of supercapacitor (W)
T^{DW}	Dwell time (s)
U_1^{BR}	Voltage threshold of braking resistor (V)
η^{BR}	Proportion of train braking power to traction network (%)

G. Parameters

$c_{\mathrm{SC}}^{\mathrm{INV}}, c_{\mathrm{BT}}^{\mathrm{INV}}, c_{\mathrm{DC}}^{\mathrm{INV}}$	Unit costs of investment of supercapacitor, battery, and converter ($/module)
$c_{\mathrm{PV}}^{\mathrm{INV}}$	Unit cost of investment of PV ($/kW)
$c_{\mathrm{SC}}^{\mathrm{OM}}$	Unit cost of supercapacitor operation ($/MWh)
c_{SUB}	Unit cost of electricity trading ($/kWh)
$I_{\max}^{\mathrm{SC}}, I_{\max}^{\mathrm{BT}}, I_{\max}^{\mathrm{DC}}$	Maximum currents of supercapacitor, battery, and converter modules (A)
$N_{\mathrm{BT}}^{\mathrm{S}}, N_{\mathrm{SC}}^{\mathrm{S}}$	Number of in-series battery and supercapacitor modules
η^{CR}	Capital recovery factor

5.1 Background

In order to comprehensively consider the balance between the economic benefits, photovoltaic–regenerative braking (PV–RB) energy utilization, and passenger demands associated with the installation of DHESSs, the DHESS capacities and train operation parameters should be flexibly configured based on factors such as line conditions, train service patterns, and energy management strategies. Nowadays, most studies on configuration optimization focused solely on supercapacitor-based ESSs (literature reviews in section 1.3). However, the sizes of DHESSs and distributed PVs and the parameters of train operation (e.g., RB parameter and timetable) have not been jointly optimized. Besides, it is necessary to consider the operation uncertainties (e.g., passenger flow fluctuations and delays) and a generalized model-free energy management strategy to handle them (network-level challenges in section 1.4). Therefore, furthering the 3rd (network) level, this chapter focuses on developing a multi-task multi-agent reinforcement learning-based data-driven multi-objective configuration optimization (MTMARL–DDMOCO) approach to optimize DHESS and train operation parameter configurations considering the proposed multi-time scale energy management framework, the electrothermal-coupled DHESS operation, and the spatial-temporal operation uncertainty for coordinating economy, energy efficiency, and passenger demands. Specifically, the main contributions of this chapter are outlined as follows:

- A multi-objective configuration optimization model of DHESS-integrated URT TNs considering electrothermal-degradation of batteries is formulated for balancing economy, energy efficiency, and passenger demands. The first stage optimizes DHESS and train operation parameter configurations to coordinate multiple objectives. The second stage implements the multi-time scale energy management approach developed in chapter 4 to minimize the daily operation cost, where its decentralized partially observable Markov decision process (Dec-POMDP) is reformulated to take first-stage decision variables as inputs.

- A non-dominated sorting genetic algorithm (NSGA-II) integrated with ensemble load prediction models is developed to solve the configuration optimization model effectively. The computational performance and configuration decisions of the proposed approach are thoroughly analyzed.

The remaining parts of this chapter are organized as follows. Section 5.2 illustrates the problem formulation of the two-stage configuration optimization model. Section 5.3 presents the proposed MTMARL–DDMOCO approach, including the ensemble load prediction model and the algorithm implementation. Section 5.4 reports case studies and their results. Section 5.5 gives the summary.

5.2 Problem Formulation

5.2.1 Overview

Similar to subsection 3.2.3, the configuration optimization model is formulated as a two-stage optimization problem. The aim of the first stage is formulated as a multi-objective function to optimize economy, energy efficiency, and passenger demands. The electrothermal-degradation relationship is especially considered for a more accurate replacement cost calculation. One of the first-stage decisions is to determine the capacities of DHESSs. As the specifications of the HESS and PV are illustrated in previous chapters and the in-series number of battery and supercapacitor modules is fixed, this decision is essential to determine the optimal in-parallel number of battery and supercapacitor modules, as well as the size of each rooftop PV. Another first-stage decision is to determine train operation parameters. Note that the start-up voltage threshold of train braking resistors determines the proportion of total braking energy delivered to the TN, which directly influences the RB energy utilization. Therefore, it is regarded as one key decision variable. Besides, the train timetable also influences the total energy consumption and RB energy utilization. To simplify the problem, we aim to fine-tune the dwell time of each station to adjust the timetable. The configuration decisions are then regarded as boundary parameters of the second stage. The second stage addresses the multi-time scale energy management issue in chapter 4 to minimize the daily operation cost. The Dec-POMDP is reformulated to take various train operation parameters as tasks. Finally, the objective and scheduling plan of the second stage is returned to the first stage for assessing configuration decisions.

5.2.2 Two-Stage Configuration Optimization Model

5.2.2.1 Battery Electrothermal-Degradation Model

Unlike the rainflow counting method focusing on the cyclic loading history, in order to consider more general cases where some traction substations may lack enough space for cooling system installation or utilize other thermal management measures (e.g., an optimal thermal management algorithm) to control the battery temperature, a more refined degradation model is provided to depict the electrothermal-degradation relationship of HESSs. The detailed steps to implement both rainflow counting and electrothermal-coupled methods for degradation estimation are illustrated in Appendix A.

5.2.2.2 Objectives and Constraints of the First Stage

The objective function can be written as

$$\min\left(J^{\text{LCC}}, 100\% - J^{\text{EU}}, J^{\text{TT}}\right). \tag{5.1}$$

where J^{LCC}, J^{EU}, and J^{TT} are the economic, energy utilization, and travel time indictor, respectively.

1) Economic Indicator: The life cycle cost (LCC) of the urban rail transit (URT) TN is utilized as the economic indicator, which consists of the investment cost J^{INV}, operation cost J^{OM}, replacement cost J^{REP}, and other installation cost J^{FIX}. The objective can be written as

$$J^{\text{LCC}} = J^{\text{INV}} + J^{\text{OM}} + J^{\text{REP}} + J^{\text{FIX}}\eta^{\text{CR}}, \tag{5.2}$$

where η^{CR} is the capital recovery factor.

Specifically, the investment cost J^{INV} is modified from (3.18) by adding a PV investment cost term

$$J^{\text{INV}} = \left[\left(c_{\text{SC}}^{\text{INV}} N_{\text{SC}} + c_{\text{BT}}^{\text{INV}} N_{\text{BT}} + c_{\text{DC}}^{\text{INV}} N_{\text{DC}}\right) + c_{\text{PV}}^{\text{INV}} P^{\text{PV,norm}}\right] \cdot \eta^{\text{CR}}, \tag{5.3}$$

where $P^{\text{PV,norm}}$ is the nominal power of the PV generator, $c_{\text{SC}}^{\text{INV}}$, $c_{\text{BT}}^{\text{INV}}$, and $c_{\text{DC}}^{\text{INV}}$ are the investment cost of supercapacitor, battery, and converter per module, respectively, $c_{\text{PV}}^{\text{INV}}$ is the investment cost of PV per kW. N_{SC} and N_{BT} are the number of supercapacitor and battery modules, respectively, N_{DC} is the number of converter modules for HESS.

As for the operation cost J^{OM}, it is derived from (4.6), namely,

$$J^{\text{OM}} = \sum_{n}\left(J_n^{\text{SUB}} + J_n^{\text{SC}} + J_n^{\text{BT}} + J_n^{\text{PV}} + J_n^{\text{CUR}} + J_n^{\text{CA}}\right), \tag{5.4}$$

where J_n^{SUB}, J_n^{SC}, J_n^{BT}, J_n^{PV}, J_n^{CUR}, and J_n^{CA} are the electricity trading, supercapacitor operation, battery operation, PV operation, PV curtailment, and carbon trading costs at time interval *n*, respectively. Specially, J_n^{SUB} and J_n^{SC} are obtained by summing up their counterparts on the short time scale,

$$J_n^{\mathrm{SUB}} = \sum\nolimits_{i=1}^{I} c_{\mathrm{SUB}} E_{i,n}^{\mathrm{SUB}} = \sum\nolimits_{t=1}^{T} \sum\nolimits_{i=1}^{I} c_{\mathrm{SUB}} P_{i,t}^{\mathrm{SUB}} \Delta t, \tag{5.5}$$

$$J_n^{\mathrm{SC}} = \sum\nolimits_{i=1}^{I} c_{\mathrm{SC}}^{\mathrm{OM}} E_{i,n}^{\mathrm{SC}} = \sum\nolimits_{t=1}^{T} \sum\nolimits_{i=1}^{I} c_{\mathrm{SC}}^{\mathrm{OM}} \left(P_{i,t}^{\mathrm{SC,CH}} + P_{i,t}^{\mathrm{SC,DIS}} \right) \Delta t, \tag{5.6}$$

where E^{SUB} and E^{SC} are the traction substation and supercapacitor energy, respectively. P^{SUB} is the traction substation power. $P^{\mathrm{SC,CH}}$ and $P^{\mathrm{SC,DIS}}$ are the supercapacitor charge and discharge power, respectively. c_{SUB} is the electricity trading cost per kWh, $c_{\mathrm{SC}}^{\mathrm{OM}}$ is the supercapacitor operation cost per MWh.

As for the replacement cost J^{REP}, since the lifetime of a PV array can be much longer than the system lifetime [183], it is the same as (3.19). The only difference is that the replacement frequency here is estimated based on the battery electrothermal-degradation model.

2) Energy Utilization Indicator: To improve energy efficiency, we aim to maximize the energy utilization related to PV and RB. From the impact factor analysis in section 0, the optimal HESS control parameters for cost-saving and RB utilization improvement are different. Thus, the energy utilization indicator J^{EU} is also included as one of the objectives, which can be defined as one minus the ratio of total PV curtailment plus total braking loss to the total PV and braking energy,

$$J^{\mathrm{EU}} = \left[1 - \sum_n \left(\frac{\sum\limits_i E_{i,n}^{\mathrm{CUR}} + \sum\limits_k E_{k,n}^{\mathrm{BR}}}{\sum\limits_i E_{i,n}^{\mathrm{PV}} + \sum\limits_k E_{k,n}^{\mathrm{RB}} + \sum\limits_k E_{k,n}^{\mathrm{BR}}} \right) \right] \times 100\%, \tag{5.7}$$

where $E_{i,n}^{\mathrm{PV}}$ and $E_{i,n}^{\mathrm{CUR}}$ are the PV energy and curtailment at substation *i* at time interval *n*, respectively, $E_{k,n}^{\mathrm{RB}}$ and $E_{k,n}^{\mathrm{BR}}$ are the RB and braking resistor energies of train k at time interval n, respectively.

$E_{i,n}^{\mathrm{PV}}$ and $E_{i,n}^{\mathrm{CUR}}$ are formulated by (4.13)–(4.17). For $E_{k,n}^{\mathrm{RB}}$ and $E_{k,n}^{\mathrm{BR}}$, they are formulated as the sum of their components on the short time scale, namely,

$$E_{k,n}^{\mathrm{RB}}=\sum_{t}P_{k,t}^{\mathrm{RB}}\Delta t,\quad E_{k,n}^{\mathrm{BR}}=\sum_{t}P_{k,t}^{\mathrm{BR}}\Delta t, \tag{5.8}$$

$$P_{k,t}^{\mathrm{BR}}=\left(1-\eta_{k,t}^{\mathrm{BR}}\right)P_{k,t}^{\mathrm{TR}}. \tag{5.9}$$

where $P_{k,t}^{\mathrm{RB}}$ is determined by the train trajectory profile. $P_{k,t}^{\mathrm{BR}}$ is determined by the operation status $\eta_{k,t}^{\mathrm{BR}}$ of the braking resistor..

3) Travel Time Indicator: Apart from economy and energy utilization, we also aim to minimize passenger travel time by timetable adjustments. For simplicity, only the dwell time adjustment is considered in this work. J^{TT} is defined as the average difference between the dwell time and its lower bound of each station.

$$J^{\mathrm{TT}}=\frac{1}{I}\sum_{i}\left(T_{i}^{\mathrm{DW}}-T_{\min}^{\mathrm{DW}}\right), \tag{5.10}$$

where T^{DW} represents the dwell time.

4) Constraints: For the configuration model, the decision variables include: the area of PV arrays, the in-parallel number of battery and supercapacitor modules, and the braking resistor start-up voltage threshold. Thus, the constraints can be written as

$$\begin{cases}N_{\mathrm{SC}}=N_{\mathrm{SC}}^{\mathrm{S}}\sum_{i=1}^{I}N_{\mathrm{SC},i}^{\mathrm{P}},\\ N_{\mathrm{BT}}=N_{\mathrm{BT}}^{\mathrm{S}}\sum_{i=1}^{I}N_{\mathrm{BT},i}^{\mathrm{P}},\\ N_{\mathrm{DC}}=\left\lceil\dfrac{I_{\max}^{\mathrm{SC}}\sum_{i=1}^{I}N_{\mathrm{SC},i}^{\mathrm{P}}}{I_{\max}^{\mathrm{DC}}}\right\rceil+\left\lceil\dfrac{I_{\max}^{\mathrm{BT}}\sum_{i=1}^{I}N_{\mathrm{BT},i}^{\mathrm{P}}}{I_{\max}^{\mathrm{DC}}}\right\rceil,\\ N_{\mathrm{SC,min}}^{\mathrm{P}}\le N_{\mathrm{SC},i}^{\mathrm{P}}\le N_{\mathrm{SC,max}}^{\mathrm{P}},\ N_{\mathrm{BT,min}}^{\mathrm{P}}\le N_{\mathrm{BT},i}^{\mathrm{P}}\le N_{\mathrm{BT,max}}^{\mathrm{P}},\\ A_{\min}^{\mathrm{PV}}\le A_{i}^{\mathrm{PV}}\le A_{\max}^{\mathrm{PV}},\\ U_{1,\min}^{\mathrm{BR}}\le U_{1}^{\mathrm{BR}}\le U_{1,\max}^{\mathrm{BR}},\\ T_{\min}^{\mathrm{DW}}\le T_{i}^{\mathrm{DW}}\le T_{\max}^{\mathrm{DW}},\end{cases} \tag{5.11}$$

where $N_{\mathrm{SC}}^{\mathrm{S}}$ and $N_{\mathrm{BT}}^{\mathrm{S}}$ are the number of supercapacitor and battery modules in series, respectively. $N_{\mathrm{SC}}^{\mathrm{P}}$ and $N_{\mathrm{BT}}^{\mathrm{P}}$ are the number of supercapacitor and battery modules in parallel, respectively. $I_{\max}^{\mathrm{SC}}$, $I_{\max}^{\mathrm{BT}}$, and $I_{\max}^{\mathrm{DC}}$ are the maximum current of supercapacitor, battery, and converter modules, respectively. A^{PV} is the area occupied by PV arrays, U_{1}^{BR} is the braking resistor start-up voltage threshold. $\lceil\,\rceil$ represents the ceiling function to round up to the nearest integer.

5.2.2.3 Objectives and Constraints of the Second Stage

The proposed multi-time scale energy management framework in chapter 4 is utilized at the second stage to minimize the daily operation cost. While its upper and middle-level formulation is the same as section 4.3, the Dec-POMDP at the lower level is reformulated to take first-stage decision variables as inputs. The task representation is modified to take different dwell time and RB parameter settings as tasks,

$$\mathcal{Z}=\{\mathcal{Z}_{H},\mathcal{Z}_{P},\mathcal{Z}_{D},\mathcal{Z}_{RB}\}, \tag{5.12}$$

where $\mathcal{Z}_{H}$, $\mathcal{Z}_{P}$, $\mathcal{Z}_{D}$, $\mathcal{Z}_{RB}$ are the headway, trajectory, dwell time, and RB parameter tasks, respectively. Through this modification, the agent aims to learn a generalized DHESS control policy adapting to different headway, trajectory, dwell time, and RB parameter tasks.

It is worth noting that the task set $\mathcal{Z}$ can be huge with the increase of stations and train headway numbers, and it is possible that not all tasks are seen by the agent during training. This problem can be properly addressed by combining meta-learning with RL or simply increasing the neural network size to improve performance on unseen tasks. Due to our limited computational resources, for simplicity, we only consider the worst-case representative scenario in subsection 4.4.1 for daily operation cost calculation. As the scenario is fixed, the headway and delay information can be known, where $\mathcal{Z}_H$ and $\mathcal{Z}_P$ include only a small amount of tasks. The detailed task set capacity and agent performance are illustrated in the case study.

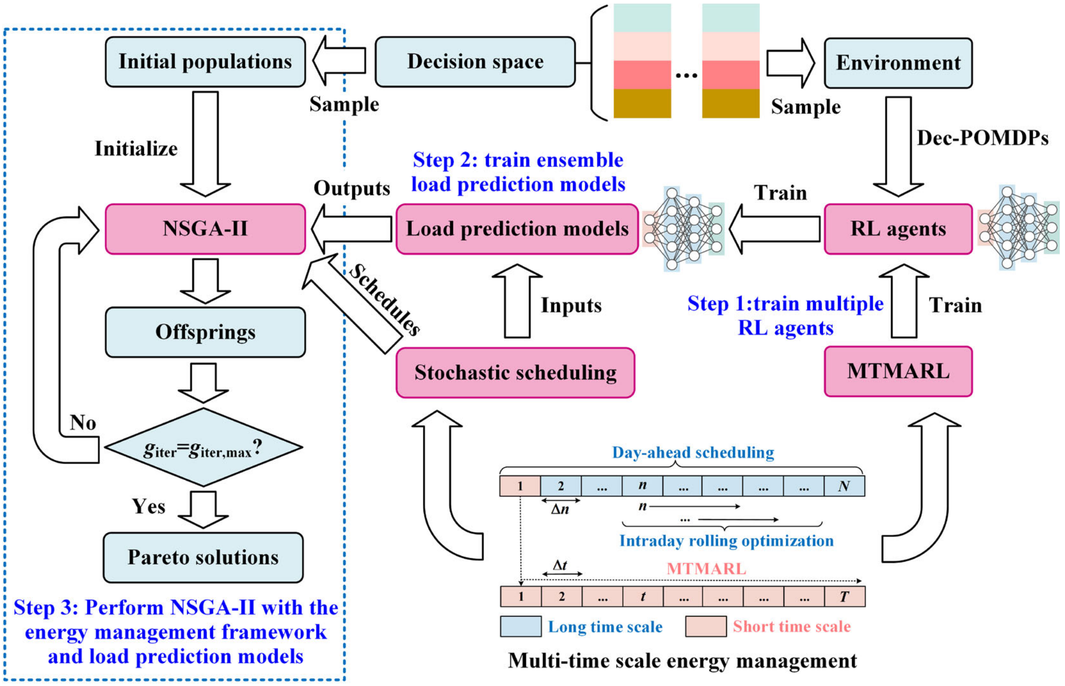


Fig. 5.1 Overview of MTMARL–DDMOCO.

5.3 MTMARL–DDMOCO Approach

5.3.1 Approach Overview

The two-stage configuration optimization model is solved by the following steps: 1) training multiple RL agents based on the modified Dec-POMDP and the MTMARL–RTEMA proposed in chapter 4 under different DHESS capacity and train operation parameter configurations. 2) training multiple ensemble load prediction models to estimate the total energy output of substations, supercapacitors, and RB on the short time scale, respectively, where the well-trained RL agents are utilized to generate the EL dataset. 3) Perform NSGA-II with the proposed multi-time scale framework and load prediction models to obtain the Pareto solutions. Specifically, the first-stage configuration decisions are provided by NSGA-II, and they are taken as inputs to optimize the scheduling plan on the long-time scale based on load prediction models. The calculated daily operation cost and other variables are returned to the first stage for assessing configuration decisions. A detailed illustration is provided in Fig. 5.1.

5.3.2 Ensemble Learning-Based Load Prediction

5.3.2.1 Principle of Ensemble Learning & XGBoost

Generally speaking, combining the predictions from several models has proven to be an effective approach for increasing the prediction accuracy of the models [184]. EL is a machine learning method that strategically combines different models (classifiers, prediction models, experts, etc.) to address regression or classification problems. Bagging, boosting, and stacking are the three main categories of EL. Among EL methods, boosting functions essentially as an approach to combine a series of weak models to develop a strong model, where weak models are sequentially ensembled to correct the predictions made by prior models, aiming to improve the prediction accuracy.

Extreme gradient boosting (XGBoost) [185] is a widely used boosting algorithm that shows efficiency and flexibility over other EL methods. It has an ensemble of M regression trees with individual predictions to construct the final output as a weighted sum of predictions. Suppose there are D samples with feature labels $\left(x_{\mathrm{el},d}, y_{\mathrm{el},d}\right)$. XGBoost is equipped with a loss function $\mathcal{L}_{EL}$ consisting of a training loss term and a regularization term

$$\mathcal{L}_{EL}(f) = \sum_{d=1}^{D} \mathcal{L}_L\left(y_{\mathrm{el},d}, \hat{y}_{\mathrm{el},d}\right) + \omega(f), \tag{5.13}$$

where $\mathcal{L}_L$ is the training loss function, $y_{\mathrm{el},d}$ is the real label, $\hat{y}_{\mathrm{el},d}$ is the estimated label, ω is the regularization term, f is the functional space that represents a set of regression trees used for boosting. The training loss $\mathcal{L}_L$ evaluates the model prediction performance of the regression tree m, while the regularization term ω mitigates the overfitting problem. As the final outcome of XGBoost is the sum of predictions of all trees, namely, $\hat{y}_d = \sum_{m=1}^{M} f_m\left(x_{\text{el},d}\right)$, the loss of the m th tree can be written as [186]

$$\mathcal{L}_{EL}(f_m) = \sum_{d=1}^{D} \mathcal{L}_L\left(y_{\text{el},d}, \hat{y}_{\text{el},d}^{m-1} + f_m\left(x_{\text{el},d}\right)\right) + \omega\left(f_m\right), \tag{5.14}$$

where $\hat{y}_{\text{el},d}^{m-1}$ represents the estimated label of the $\left(m-1\right)$ th tree.

Then, the Taylor approximation technique is utilized to transform the loss function to a function that can be solved by traditional optimization techniques, namely,

$$\mathcal{L}_{EL}(f_m) = \sum_{d=1}^{D} \mathcal{L}_L\left(y_{\text{el},d}, \hat{y}_{\text{el},d}^{m-1} + f^1 f_m\left(x_{\text{el},d}\right) + \frac{1}{2} f^2\left[f_m\left(x_{\text{el},d}\right)\right]^2\right) + \omega\left(f_m\right), \tag{5.15}$$

$$f^1 = \partial_{\hat{y}_{\text{el},d}^{m}} \mathcal{L}_L\left(y_{\text{el},d}, \hat{y}_{\text{el},d}^{m-1}\right), \quad f^2 = \partial_{\hat{y}_{\text{el},d}^{m}}^2 \mathcal{L}_L\left(y_{\text{el},d}, \hat{y}_{\text{el},d}^{m-1}\right), \tag{5.16}$$

The regularization term ω is calculated by

$$\omega\left(f\right) = c_{\text{el},1} T_{leaves} + \tfrac{1}{2} c_{\text{el},2} \left\|\lambda\right\|^2, \tag{5.17}$$

where T_{leaves} is the number of leaves, λ is the weight of leaves, $c_{\text{el},1}$ and $c_{\text{el},2}$ are coefficients with default values of 1 and 0, respectively.

With the regularization term, XGBoost can mitigate the overfitting problem. Meanwhile, in chapter 4, in order to generate the optimal schedules, it is necessary to 1) train RL agents under specific configuration decisions and 2) calculate power flows subject to agents' actions to calculate operation cost under different referential PV-battery energy. Nevertheless, on the one hand, due to the short time scale for DHESS control, each operation scenario requires hundreds of power flow calculations. On the other hand, each configuration solution requires retraining of the RL agents, which results in substantial computational time.

Therefore, in this work, several ensemble load prediction models are developed to directly predict energy terms in (5.5)–(5.6) and (5.8) instead of executing hundreds of power flow calculations. These models learn from environmental settings and behaviors of well-trained RL agents. More importantly, the generalized EL models can adapt to different DHESS and train operation configurations, which do not require retraining of the RL agents. The detailed approach to load prediction is illustrated as follows.

5.3.2.2 Construction of Ensemble Load Prediction Models

To simplify the training process of XGBoost, the aim of the proposed load prediction models is to estimate the total energy output of substations, supercapacitors, and RB of each time step under the specific operation scenario for configuration optimization purposes. In the scenario, several operation parameters, such as the daily passenger flow, train trajectory, dwell time, and delay setting, are fixed. Therefore, we do not aim to develop a comprehensive load prediction model that adapts to various TN operation scenarios.

Totally three load prediction models are trained. The scheduled referential PV-battery energy $E_{1,n}^{\mathrm{ref}},\cdots,E_{I,n}^{\mathrm{ref}}$, time step n, headway h, supercapacitor in-parallel module number $N_{\mathrm{SC},1}^{\mathrm{P}},\cdots,N_{\mathrm{SC},I}^{\mathrm{P}}$, battery in-parallel module number $N_{\mathrm{BT},1}^{\mathrm{P}},\cdots,N_{\mathrm{BT},I}^{\mathrm{P}}$, and train operation parameters $U_1^{\mathrm{BR}},T_1^{\mathrm{DW}},\cdots,T_I^{\mathrm{DW}}$ are chosen as the same input of these models. Accordingly, the total substation energy $\sum_{i=1}^{I}E_{i,n}^{\mathrm{SUB}}$, total supercapacitor energy $\sum_{i=1}^{I}E_{i,n}^{\mathrm{SC}}$, and total RB energy $\sum_{k=1}^{K}E_{k,n}^{\mathrm{RB}}$ are the outputs of prediction models #1, #2, and #3, respectively.

The training data are generated based on the fixed operation scenario and the trained RL agents, and the generation process is illustrated as follows. First, randomly set the supercapacitor and battery module numbers and train corresponding RL agents. Next, randomly generate the referential PV-battery energy schedule for each time step. Next, run the operation scenario with fixed daily passenger flow, train trajectory, dwell time, and delay settings, as well as the random referential PV-battery energy schedule. Noted that the well-trained RL agents will determine the DHESS charge/discharge voltage thresholds and PV-battery power on a short time scale. Finally, match the input and output to form a data point and repeat the process till sufficient data are obtained. 80% of the dataset is randomly taken for training, and the rest 20% is for testing.

5.3.3 NSGA-II Implementation with Multi-Time Scale Energy Management Framework and Load Prediction Models

5.3.3.1 Principle of NSGA-II

The genetic algorithm (GA) [187] is a heuristic search algorithm that emulates the processes of natural selection and evolutionary genetics observed in the biological realm. Although GA demonstrates strong applicability for single-objective optimization problems, it exhibits significant limitations when applied to multi-objective optimization challenges. For the configuration optimization problem in this work, the objectives are interrelated and exhibit conflicting interactions. For instance, increasing the PV–RB utilization can result in excessive usage of battery and increased replacement cost, which leads to a worse performance on LCC. The optimal solution to such a problem should consist of a set of non-dominated solutions, namely, Pareto optimal solutions [188]. By definition, for decision variables $x_{\mathrm{ga},1}$ and $x_{\mathrm{ga},2}$: 1) If for $\forall o \in \{1,2,\cdots,O\}$, the fitness function $F_o(x_{\mathrm{ga},1}) \le F_o(x_{\mathrm{ga},2})$, variable $x_{\mathrm{ga},1}$ dominates $x_{\mathrm{ga},2}$. 2) If for $\forall o \in \{1,2,\cdots,O\}$, $F_o(x_{\mathrm{ga},1}) \le F_o(x_{\mathrm{ga},2})$, and $\exists o \in \{1,2,\cdots O\}$,

$F_o(x_{\mathrm{ga},1}) < F_o(x_{\mathrm{ga},2})$, variable $x_{\mathrm{ga},1}$ weakly dominates $x_{\mathrm{ga},2}$. 3) If for arbitrary decision variable $x_{\mathrm{ga},2}$ and $\forall o \in \{1,2,\cdots,O\}$, $F_o(x_{\mathrm{ga},1}) \le F_o(x_{\mathrm{ga},2})$, variable $x_{\mathrm{ga},1}$ is the non-dominated solution for objective minimization. To effectively solve the multi-objective configuration optimization problem, the NSGA-II [189] is introduced as the backbone of the proposed algorithm. The features of NSGA-II are summarized as follows.

First, it utilizes a fast non-dominated sorting (NDS) method to significantly decrease computational complexity. In general, for the multi-objective optimization problem, it aims to calculate the number of solutions $N^{\mathrm{ga}}(p^{\mathrm{ga}})$ which dominate the solution p^{ga} and the solution set $S^{\mathrm{ga}}(p^{\mathrm{ga}})$ that the solution p^{ga} dominates. 1) Find the solution p^{ga} with $N^{\mathrm{ga}}(p^{\mathrm{ga}})=0$ and save them to a set $S^{\mathrm{ga}}(1)$, $S^{\mathrm{ga}}(1) \leftarrow S^{\mathrm{ga}}(1) \cup \{p^{\mathrm{ga}}\}$, rank it with $rank=1$. 2) For the solution q^{ga} in set $S^{\mathrm{ga}}(q^{\mathrm{ga}})$ which is dominated by solution p^{ga} in set $S^{\mathrm{ga}}(1)$, $N^{\mathrm{ga}}(p^{\mathrm{ga}}) \leftarrow N^{\mathrm{ga}}(p^{\mathrm{ga}})-1$. If $N^{\mathrm{ga}}(p^{\mathrm{ga}})$=0, save q^{ga} to a set $S^{\mathrm{ga}}(2)$, $S^{\mathrm{ga}}(2) \leftarrow S^{\mathrm{ga}}(2) \cup \{q^{\mathrm{ga}}\}$, rank it with $rank=2$. 3) Repeat the process till all solutions are saved. The non-dominated sets $S^{\mathrm{ga}}(1)$, $S^{\mathrm{ga}}(2)$, $S^{\mathrm{ga}}(3)$, … are generated with rank 1, 2, 3, ….

Then, it introduces the crowding distance metric to estimate the density of solutions around a particular solution and maintain the diversity in the population. Thus, after the above two steps, each solution is characterized by the non-dominated rank and crowding distance. The population selection process based on the non-dominated sorting and crowding distance metric can be described as follows: 1) If the ranks determined by non-dominated sorting of solutions are different, the solution with a smaller rank is selected. 2) If the ranks are the same, the solution with a larger crowding distance is selected. Third, it implements elitism, which retains the best solutions from the current and previous generations, ensuring that progress is not lost over generations.

5.3.3.2 Algorithm Implementation

The detailed steps of performing NSGA-II with the multi-time scale energy management strategy and load prediction models are summarized in Algorithm 5.1.

Algorithm 5.1 NSGA-II with the multi-time scale energy management strategy and load prediction models

```
1  Initialization: Load TN operation scenario data and ensemble load prediction
   models, set the hyperparameters for NSGA-II
2  Randomly generate an initial population $PA(0)$
3  $J^{\mathrm{LCC}}$ calculation: Predict the energy terms in (5.5)–(5.6) for all possible input
   conditions in $PA(0)$. Then, optimize the scheduling decisions and calculate the
   fitness function in (5.2) based on the prediction
4  $J^{\mathrm{EU}}$ calculation: Predict the energy terms in (5.8) based on $PA(0)$ and the
   optimal schedule. Then, calculate the fitness function in (5.7) based on the
   prediction
5  $J^{\mathrm{TT}}$ calculation: Calculate (5.10) directly from $PA(0)$
6  Implement fast NDS, crowding distance calculation, and elite selection,
   generate the parent population $PA(1)$
7  For generation $g_{\mathrm{iter}}=1, g_{\mathrm{iter,max}}$ do
8      Implement selection, crossover, and mutation operations, generate an off-
       spring population $OS(g_{\mathrm{iter}})$. Calculate the fitness function by step 2–4, simply
       substitute $PA(0)$ by $OS(g_{\mathrm{iter}})$
9      Merge the parent and off-spring population, obtain a new population with
       $2N^{\mathrm{pop}}$ solutions, $PA(g_{\mathrm{iter}})\cup OS(g_{\mathrm{iter}})$
10     Implement fast NDS, crowding distance calculation, and elite selection,
       generate the parent population $PA(g_{\mathrm{iter+1}})$
11 Output: Pareto solutions
```

5.4 Case Study

In this section, a detailed analysis of the aforementioned formulations and algorithms is conducted. First, the performance of the proposed ensemble load prediction models is illustrated. An optimal model selection process is conducted to verify the effectiveness of the proposed model compared with common load prediction models. The impacts of the training set size and the RL agent number on model prediction performance are discussed to demonstrate the trade-off effect between prediction accuracy and computational efficiency. Then, the optimal configuration of DHESS-integrated URT TNs is investigated. The configuration results consider only two objectives (economic and energy utilization indicators), and all three objectives (economic, energy utilization, and travel time indictors) are compared. Besides, the impact of battery degradation is analyzed with different ambient temperatures.

5.4.1 Setup

The search range of each decision variable is shown in Table 5.1, and the basis of the search range setting is illustrated as follows. For the DHESS sizing setting, the battery and supercapacitor specifications used in this case study are listed in Table 5.2 and Table 3.4, respectively. Due to the data availability, we use the LiFePO4 battery data reported in [190, 191] to replace the aforementioned LTO battery data. Referred to the analysis in subsection 4.4.2, the parallel number of supercapacitor modules $N_{\mathrm{SC},i}^{\mathrm{P}}$ is set to make the rated supercapacitor power cover 0–120% of the maximum power of its integrated traction substation. On the other hand, the parallel number of battery modules $N_{\mathrm{BT},i}^{\mathrm{P}}$ is set to make the rated battery energy cover 0–100% of the maximum energy of the PV in its integrated traction substation.

Table 5.1 Range of decision variables[1].

Parameter	Value	Parameter	Value	Parameter	Value
$N_{\mathrm{SC},i}^{\mathrm{P}}$	[0, 12]	A_i^{PV} (m^2)	[0, 1620]	$T_i^{\mathrm{DW,up}}$ (s)	[20, 50]
$N_{\mathrm{BT},i}^{\mathrm{P}}$	[0, 100]	U_1^{BR} (V)	[860, 1000]	$T_i^{\mathrm{DW,down}}$ (s)	[20, 50]

[1] $i \in \{1,2,\cdots,I\}$, $N_{\mathrm{SC},i}^{\mathrm{P}}$ is based on the maximum power of the integrated traction substation, $N_{\mathrm{BT},i}^{\mathrm{P}}$ is based on the maximum daily generation of PV in the integrated traction substation, A_i^{PV} is based on the maximum rooftop area, U_1^{BR} is based on the no-load and maximum operation voltages of the TN, $T_i^{\mathrm{DW,up}}$ and $T_i^{\mathrm{DW,down}}$ are based on possible dwell time ranges in [192].

Table 5.2 LiFePO4 battery parameters.

Item	Value	Item	Value
Nom. voltage	3.2 V	No. in series	210
Nom. capacity	10 Ah	No. in parallel	Table 5.1
Max. charge/discharge rate	1 C/2 C	$c_{\mathrm{BT}}^{\mathrm{OM}}$	1 \$/MWh
SoE range	0.2-0.8	$c_{\mathrm{BT}}^{\mathrm{INV}}$	11.58 \$/module

Table 5.3 NSGA-II parameters.

Parameter	Value	Parameter	Value
$g_{\mathrm{iter,max}}$	300	Mutation function	mutationpower
N^{pop}	150	Crossover function	crossoverlaplace
Selection function	selectiontournament	Crossover probability	0.8

Besides, with the change of battery type, considering the maximum PV-battery energy consumption and the computational burden, we set $\max\left(E^{\mathrm{ref}}\right)$=25 kWh and the increment of E^{ref} =2.5 kWh in this case study. As for the PV sizing setting, its occupation A_i^{PV} is limited by the total rooftop area of the station (2700 m^2, reported in [10]). It is worth noting that 30% of the total rooftop area is occupied by skylights. Besides, we assume 10% of the total rooftop area is reserved for laying gaps and maintenance aisles. Thus, the PV occupation can be 0–60% of the total rooftop area. In addition, the braking resistor threshold U_1^{BR} is bounded by the no-load and maximum operation voltages of the TN. The available dwell time ranges $T_i^{\mathrm{DW,up}}$ and $T_i^{\mathrm{DW,down}}$ for up and down train directions are reported in [192]. Furthermore, other economic parameters are the same as subsection 4.4.1, and a PV investment cost $c_{\mathrm{PV}}^{\mathrm{INV}}$ =462 $/kW is obtained from the statistics in China, 2025 [193] (taking the exchange rate 6.5 RMB/USD). The NSGA-II algorithm parameters are illustrated in Table 5.3. Since ensemble models are sensitive to hyperparameters, a fine-tuning model selection process is required for their optimal performance, and the optimal parameters are illustrated in the following subsection. XGBoost and the fitness calculation part of NSGA-II are performed with Python 3.9.13, and other parts of NSGA-II are conducted by Matlab 2022b. The operation scenario and related scheduling plan are simulated and optimized by Gurobi 10.0.0 with Python 3.9.13. All simulations are performed on the same device in subsection 2.4.1.

5.4.2 Analysis of Load Prediction Results

5.4.2.1 Model Selection Results

In this subsection, the ensemble model selection process, including key model parameter optimization and model prediction performance comparison, is illustrated. First, several key parameters are optimized, including: 1) the number of regression tree models M , 2) learning rate ξ_e, 3) maximum depth of regression trees, with a higher value for more complex models, 4) the minimum loss reduction “gamma” to divide further on the leaf nodes of the tree, which a higher value for more conservative models, and 5) the subsample rate.

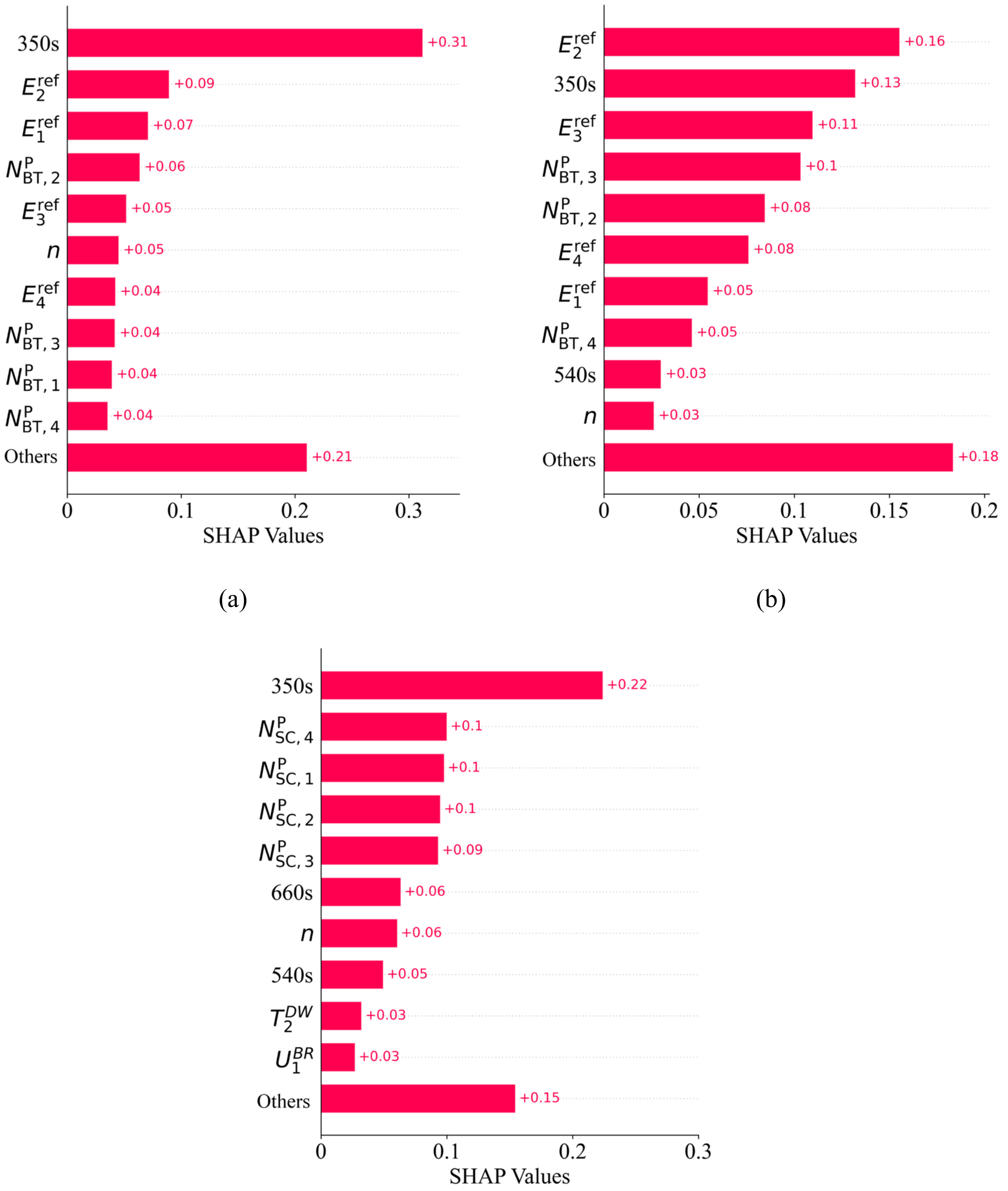
Table 5.4 XGBoost parameters.

Parameter	Searching range	Optimal value		
		Model #1	Model #2	Model #3
M	[100, 1000]	1000	1000	1000
ξ_e	[0.01, 0.1]	0.1	0.075	0.05
Maximum depth	[3, 10]	4	4	4
Gamma	[0, 0.5]	0	0	0.3
Subsample	[0.5, 0.9]	0.9	0.9	0.6

Table 5.5 Comparative performances under different prediction methods.

Metrics	XGBoost-model			RF-model			LR-model		
	#1	#2	#3	#1	#2	#3	#1	#2	#3
MAE	0.239	0.008	2.183	0.296	0.011	2.523	0.430	0.016	3.585
RMSE	0.305	0.010	2.772	0.378	0.014	3.258	0.541	0.020	4.513
MAPE	5.347	7.448	5.728	6.676	9.166	6.502	9.520	15.985	9.540
R^2	0.898	0.920	0.916	0.843	0.861	0.883	0.678	0.712	0.776

The optimal values of these parameters are listed in Table 5.4. Then, the performance of different prediction methods is compared to verify the effectiveness of the proposed approach: *1) XGBoost:* the EL method utilized in this work. *2) Random Forest (RF):* a classical and widely used bagging-based EL method that builds decision trees and takes average predictions for regression. *3) Linear Regression (LR):* baseline method. The training and test samples are 20000 and 5000, respectively. The RL agent number is 500. The mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) are utilized as metrics to evaluate the prediction results. Table 5.5 summarizes the evaluation results, where XGBoost shows superior prediction performance on the test set for all metrics.



(a)

(b)

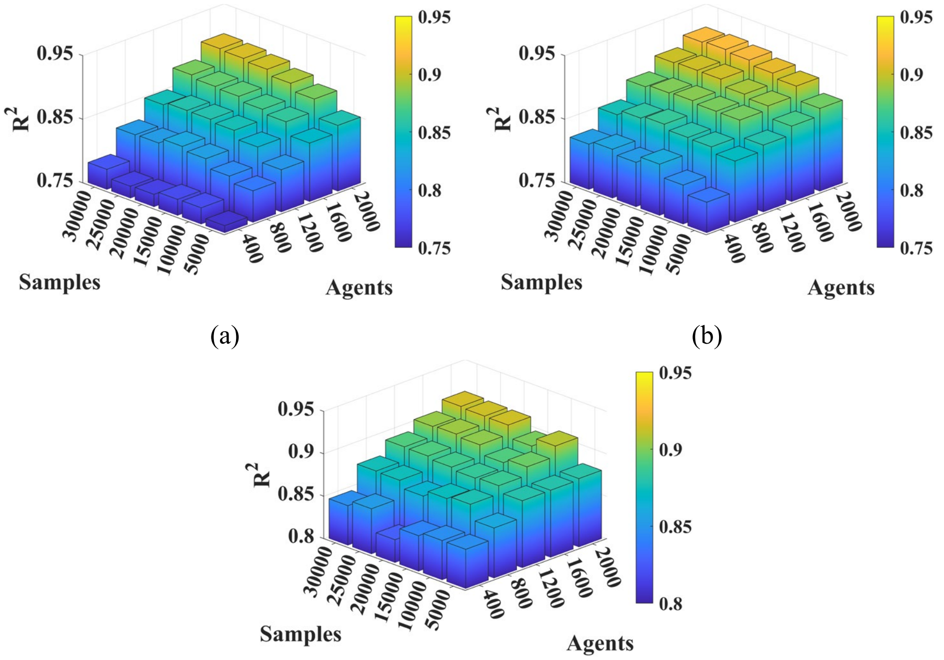
(c)

Fig. 5.2 Contribution ranking of each feature to prediction results for load prediction model (a) #1, (b) #2, and (c) #3.

Furthermore, in order to quantify the impact of input features on prediction results, the Shapley additive explanation (SHAP) [194], a unifying interpretability framework based on cooperative game theory, is introduced. SHAP uses the average marginal contribution of a feature to all feature coalitions with that feature to obtain explanations of machine learning models. Fig. 5.2 shows the final ranking diagram of the importance of all features affecting the load prediction models. It can be observed that, headway and time are two common major factors that affect the prediction results of all models. Besides, the referential PV-battery energy and battery module number in different stations are crucial factors for substation and supercapacitor energy demand predictions. In contrast, supercapacitor module number, dwell time, and braking resistor parameter substantially impact RB energy prediction.

5.4.2.2 Impact of Training Set Size and RL Agent Number

In this subsection, the impact of training set size and RL agent number on model prediction performance and the trade-off between prediction accuracy and computational efficiency is demonstrated. The test performance curves of each model with respect to training set size and RL agent number are depicted in Fig. 5.3. From the figure, with the increase of training set size and RL agent number, the model performance also increases gradually. Then, the total training time is calculated. Note that parallel computing has been implemented for RL agent training. The prediction models with the minimum total training time and satisfactory prediction performance (R^2>0.9) are selected, and their performances are listed in Table 5.6. From the table, most of the computational time is spent on RL agent training and data generation. Although the generation of one sample is less than 0.1 s, the large number of training samples can make the generation process long. However, this issue can be addressed by parallel computing. In general, compared with carrying out enumeration calculations in the entire sampling space to obtain accurate operation cost under different conditions, the proposed load prediction models significantly improve the computational efficiency while slightly reducing the accuracy of operation cost calculations.



(a)

(b)

(c)

Fig. 5.3 Test performance with respect to training set size and RL agent number for prediction model (a) #1, (b) #2, and (c) #3.

Table 5.6 Computational performance of prediction models 1–3[1,2].

Items	Model#1	Model #2	Model #3
EL training set size (sample)	20000	20000	10000
Number of RL agents	2000	1600	1600
Number of RL algorithm training times	500	400	400
Total RL training time (h)	25.77	20.61	20.61
Data generation time (s)	1506.574	1507.533	722.352
Data generation time per sample (s)	0.075	0.075	0.072
Total EL model training time (s)	0.411	0.369	0.0314
Execute time (s)	0.002	0.002	0.002

[1] Total training time includes RL training, data generation, EL model training, and execute times.

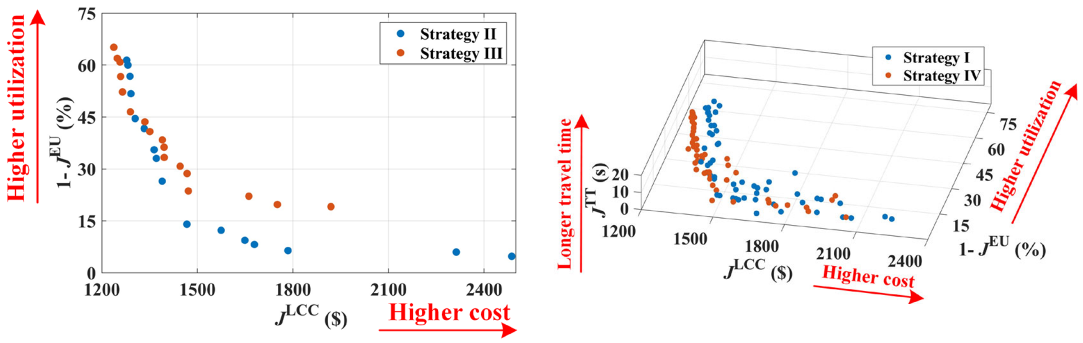
[2] Parallel computing was used, where 20 RL algorithms were trained simultaneously.

5.4.3 Analysis of Optimal Configuration Results

5.4.3.1 Comparison of Configuration Results

In this subsection, the optimal configuration results are discussed with the following configuration strategies: 1) *Strategy I (proposed):* the DHESS and train operation parameter configurations are coordinated to optimize all three objectives. The electrothermal-degradation (A.2)–(A.10) of batteries is considered. *2) Strategy II:* same as strategy I, except that only two objectives (economic and energy utilization) are considered. *3) Strategy III:* same as strategy II, except that only DHESS configurations are considered. For train operation parameters, we set U_1^{BR} =860 V, $T_i^{\mathrm{DW,up}} = T_i^{\mathrm{DW,down}}$ =50 s, $i \in \{1, 2 \cdots, I\}$. *4) Strategy IV:* same as strategy I, except that the multi-time scale energy management approach developed in chapter 4 is replaced by the framework DAIS in subsection 4.4.3.2. DAIS performs day-ahead and intraday scheduling without optimizing the charge/discharge thresholds and real-time PV-battery outputs of DHESSs. For strategies II-III, as only two objectives are optimized, the population N^{pop} is set as 50 to save computational resources. N^{pop}=150 for all other strategies.

The Pareto solutions of all strategies are shown in Fig. 5.4. From Fig. 5.4(a), it can be observed that the Pareto solutions of strategy II are generally better than those of strategy III. This is because the optimization of train operation parameters by strategy II increases RB energy generation and further releases its utilization flexibility. From Fig. 5.4(b), the Pareto front of strategy IV is located within $1200–2100 and is more concentrated than strategy I. Besides, the energy utilization of strategy I is generally higher than that of strategy IV when the LCC is higher than $1500.



(a)

(b)

Fig. 5.4 Pareto fronts of (a) strategies II-III and (b) strategies I and IV.

Table 5.7 Comparative train operation parameters of strategies I–IV.

Parameters	$T_i^{\text{DW,down}}$ (s)			$T_i^{\text{DW,up}}$ (s)			U_1^{BR} (V)
Station	1	2	3	4	3	2	
Current	30	30	30	30	30	30	900
Strategy I	24	44	26	32	26	38	910
Strategy II	32	24	26	28	38	44	910
Strategy III	50	50	50	50	50	50	860
Strategy IV	36	40	34	24	48	24	910

Then, a multi-criteria decision-making is implemented to select the optimal configuration solution from the Pareto [195]. After calculating the entropy term, the optimal solution is selected based on the weighted distance to the idea solution,

$$P^* = \max\left\{\frac{P_1^-}{P_1^- + P_1^+}, \frac{P_2^-}{P_2^- + P_2^+}, \cdots, \frac{P_W^-}{P_W^- + P_W^+}\right\}, \tag{5.18}$$

$$P_w^- = \sqrt{\sum_{o=1}^{O} \chi_o \left(F_{w,o} - \max\left(F_{w,o}\right)\right)^2}, P_w^+ = \sqrt{\sum_{o=1}^{O} \chi_o \left(F_{w,o} - \min\left(F_{w,o}\right)\right)^2}, \tag{5.19}$$

$$\chi_o = \left(1 - \Theta_o\right) / \sum_{o=1}^{O} \left(1 - \Theta_o\right), \tag{5.20}$$

where P^* is the optimal solution, P_w^- and P_w^+ are the positive and negative ideal distance of the wth Pareto solution, respectively, χ is the weight term, Θ is the entropy of information.

Table 5.8 Comparative PV sizes of strategies I–IV.

Parameters	PV size (m^2)			
Station	1	2	3	4
Strategy I	314.45	204.30	1204.18	526.29
Strategy II	1007.37	238.14	1288.44	239.22
Strategy III	677.97	293.22	620.46	714.69
Strategy IV	143.22	653.31	968.21	1097.12

Table 5.9 Comparative DHESS configuration results of strategies I–IV.

Power	HESS (MW)				Supercapacitor (MW)				Battery (MW)			
Station	1	2	3	4	1	2	3	4	1	2	3	4
I	1.18	1.44	1.01	0.65	0.70	0.87	0.61	0.35	0.48	0.56	0.40	0.30
II	0.25	1.56	1.67	0.18	0.09	0.61	0.61	0.09	0.16	0.97	1.06	0.09
III	1.51	0.23	1.00	1.39	1.05	0.17	0.96	0.79	0.46	0.05	0.04	0.60
IV	1.81	0.74	0.26	1.12	0.87	0.70	0.09	1.05	0.94	0.04	0.17	0.07
Energy	HESS (kWh)				Supercapacitor (kWh)				Battery (kWh)			
Station	1	2	3	4	1	2	3	4	1	2	3	4
I	325	386	274	189	83	103	72	41	242	282	202	148
II	91	556	603	57	10	72	72	10	81	484	531	47
III	353	48	134	396	124	21	114	93	228	27	20	302
IV	574	103	98	158	103	83	10	124	470	20	87	34

Table 5.10 Comparative performance of strategies I–IV.

Item	Current	Strategy			
		I	II	III	IV
J^{LCC} (\$)	2141.44	1561.33	1784.15	1471.44	1487.20
J^{EU} (%)	24.38	81.53	93.58	76.36	78.66
J^{TT} (s)	10.00	11.67	12.00	30.00	14.33

Table 5.11 Comparative battery degradation of strategies I–IV.

Strategy	I			II		
Ambient temperature (°C)	25	35	45	25	35	45
J^{REP} (\$)	0.00	0.00	0.00	0.00	55.51	57.20
$L_{\mathrm{BT}}, i=1$ (year)	10.00	10.00	10.00	10.00	9.60	8.39
$L_{\mathrm{BT}}, i=2$ (year)	10.00	10.00	10.00	10.00	10.00	10.00
$L_{\mathrm{BT}}, i=3$ (year)	10.00	10.00	10.00	10.00	10.00	10.00
$L_{\mathrm{BT}}, i=4$ (year)	10.00	10.00	10.00	10.00	10.00	10.00
Strategy	III			IV		
Ambient temperature (°C)	25	35	45	25	35	45
J^{REP} (\$)	0.00	49.79	51.35	105.44	108.10	177.56
$L_{\mathrm{BT}}, i=1$ (year)	10.00	10.00	10.00	10.00	10.00	10.00
$L_{\mathrm{BT}}, i=2$ (year)	10.00	10.00	10.00	7.93	6.92	6.02
$L_{\mathrm{BT}}, i=3$ (year)	10.00	9.79	8.54	10.00	10.00	8.81
$L_{\mathrm{BT}}, i=4$ (year)	10.00	10.00	10.00	8.00	6.99	6.10

Table 5.7–Table 5.9 shows the configuration results of the optimal solution of each strategy, and the current operation situation with no installation of PVs and DHESSs is also compared (U_1^{BR} =900 V, $T_i^{\mathrm{DW,up}} = T_i^{\mathrm{DW,down}}$ =30 s, $i \in \{1,2\cdots,I\}$). From Table 5.7, the optimized dwell time of each station in strategies I–IV is generally higher than the current dwell time schedule to increase the economic benefits and energy saving. The optimized braking resistor start-up voltage threshold of strategies I, II, and IV is higher than the current threshold, promoting more RB energy generation.

From Table 5.8 and Table 5.9, due to the change in the spatial-temporal distribution of the traction load and RB energy under the optimized dwell timetable and braking resistor parameter, the PV and DHESS sizes of each station vary. The average HESS power and energy per station of strategies I-IV reach 0.92–1.07 MW and 232.75–326.75 kWh, respectively. Besides, all strategies utilize 20–27% of the total rooftop area on average (or 33–45% of the available area) for PV arrays in each station. This indicates that the current rooftop area is enough for the PV installation, considering the economic PV energy use for traction demands.

Table 5.10 summarizes the performance of the optimal solution of each strategy. Compared with strategies II-III, which ignore the travel time indicator, the passenger travel time saving of strategy I is improved by at least 2.75%. Compared with strategy IV, strategy I using the proposed multi-time scale energy management approach improves the PV–RB energy utilization and the passenger travel time saving by 3.65% and 18.56%, respectively, while decreasing the LCC by 4.98%. Moreover, compared with strategy III, strategy II optimizes the train operation parameters, which promotes PV–RB energy utilization and passenger travel time-saving. In general, different strategies improve the LCC by 16.68–31.29% and the PV–RB energy utilization by 213.21–283.84% while decreasing the travel time saving by 16.7–200%. The proposed strategy I is superior to other strategies on at least two objectives, which verifies its effectiveness.

5.4.3.2 Impact of Battery Degradation

To analyze the impact of battery degradation, the battery lifetime and replacement cost under different ambient temperatures are investigated. Noted that the ambient temperature will affect the battery degradation calculation by (A.3), (A.7)–(A.8), and (A.10) in appendix A. Table 5.11 illustrates the battery degradation of strategies I–IV under different ambient temperatures. With the increase in temperature, the battery lifetimes are generally shortened, and the replacement cost grows. Strategy IV is more sensitive to temperature change than other strategies, where its average battery lifetime can be shortened by 13.92%, and its total replacement cost can be increased by 68.40%. In contrast, strategy I reduces battery usage through a reasonable scheduling plan, intelligent control of RL agents, and joint configuration of DHESSs and train operation parameters, limiting the replacement times under different ambient temperatures.

5.5 Summary

In this chapter, an MTMARL–DDMOCO approach is proposed for promoting an optimal synergy between the economic and energy efficiencies of the DHESS-integrated TN operation and the travel time of the passengers. The research mainly includes the following aspects.

A multi-objective configuration optimization model considering the electrothermal-degradation relationship of batteries is formulated for balancing economy, energy efficiency, and passenger demands based on the developed MTMARL–MTSEM approach in Chapter 4. The NSGA-II algorithm is incorporated with ensemble load prediction models to solve the multi-objective configuration optimization model in a data-driven manner.

The key findings of the designated case study are summarized as follows: 1) By quantitatively analyzing the impacts of EL methods and training parameters on load prediction, the XGBoost method and critical parameters achieving $R^2>0.9$ are selected to obtain optimal substation, supercapacitor, and RB energy output predictions under minimum total training time. 2) The MTMARL–DDMOCO improves the LCC and PV–RB energy utilization by 27.09% and 234.41%, respectively, compared with the current situation. 3) The MTMARL–DDMOCO is also superior to other configuration strategies on at least two objectives and reduces battery usage by leveraging a reasonable scheduling plan, intelligent control of RL agents, and joint configuration of DHESSs and train operation parameters.

Chapter 6: Conclusions and Future Perspectives

6.1 Conclusions

The integration of URT TNs with HESSs has become a technologically and socioeconomically crucial pathway to enable highly efficient and convenient mass public transportation within urban areas while promoting energy consumption reduction and carbon-neutral transformation of URTs. In order to address the spatial-temporal uncertainties and complexities arising from passenger demand, urban traffic congestion, widespread distribution, operational disturbances, etc., this thesis reports using RL as a machine learning base technique to develop three different levels of energy management and configuration strategies for HESS-integrated URT TNs in a model-free and data-driven manner. The specific works and conclusions are summarized as follows.

An SRL–EETTO approach is developed for automatic train operation at the 1st (train) level. The real-time train operation model under uncertain disturbances is formulated as an MDP, and an S-TD3 algorithm with improved effectiveness is developed to solve it and generate optimal train trajectories online. Satisfactory performances on reduced traction energy use of 18.5% and evaluation indices of safety, punctuality, and ride comfort are verified compared to the practical driving data for the proposed SRL–EETTO approach. Besides, its maximum trip time error under real-time uncertain disturbances is decreased by 11.6% compared to state-of-the-art RL-based EETTO algorithms, while its adaptability to uncertain train masses and resistance conditions is demonstrated. A suggestion for the train trajectory configuration based on the proposed SRL–EETTO approach is given.

An MTRL–SCO approach is proposed for HESS-integrated traction substation operation at the 2nd (substation) level. The configuration-specific HESS control problem under various spatial-temporal traction load distributions is formulated as an MTMDP based on a DTM, and an iterative sizing optimization approach considering daily service patterns is devised to minimize the HESS LCC. Then, a KT-D3QN algorithm is presented to learn a generalized HESS control policy adapting to multiple train service patterns by leveraging a shareable cross-task experience for solving the MTMDP. With the joint optimization of voltage thresholds and power allocations to effectively adjust SoEs, the operation cost can be reduced by 5.89% compared with conventional rule-based strategies using fixed thresholds and power allocations. Under multi-task learning and knowledge transfer, the operation cost can be further decreased by at most 13.06% compared with state-of-the-art HESS control optimization methods. Considering the spatial-temporal traction load characteristics, the HESS LCC is reduced by 2.65% while the battery life is extended by 86.22% compared with conventional sizing optimization methods.

An MTMARL–MTSEM is presented for DHESS-integrated TN operation at the 3rd (network) level. A two-stage stochastic scheduling is performed on a long-time scale to minimize daily operation and carbon trading costs at the upper level and correct day-ahead scheduling deviations against multi-source uncertainties at the middle level. An MTMARL–RTEMA is established to optimize PV–RB power flow and promote utilization through decentralized coordination of DHESSs at the lower level. By leveraging the synergetic consideration of energy management of the TN with multiple time scales, the overall daily operation cost is reduced by 11.98%, and the PV–RB energy utilization is improved by 13.94% compared with the conventional long-time-scale scheduling approach. Similar to the conclusion of the proposed MTRL–SCO approach, the joint optimization of voltage thresholds and PV–RB power allocations increases the spatial-temporal energy complementation of PV–RB by 10.31% compared to the uncoordinated control scheme. Finally, the adaptability of the proposed MTMARL–MTSEM approach under various train service patterns and network uncertainties is verified.

An MTMARL–DDMOCO approach is established for furthering the DHESS-integrated TN operation at the 3rd (network) level. Based on the developed MTMARL–MTSEM approach, a multi-objective configuration optimization model considering the electrothermal aging of batteries is formulated to optimize DHESS capacities and train operation parameters simultaneously. Then, a non-dominated sorting genetic algorithm is incorporated with data-driven EL-based load prediction models to solve the multi-objective configuration optimization model. The optimal load prediction models (R^2>0.9) under minimum total training time are obtained with the XGBoost method and critical training parameters. By leveraging a reasonable scheduling plan, intelligent control of RL agents, and joint configuration of DHESSs and train operation parameters, the proposed MTMARL–DDMOCO improves the LCC and PV–RB energy utilization by 27.09% and 234.41%, respectively, compared with the current situation, and is superior to other configuration strategies on at least two objectives.

6.2 Future Perspectives

Considering the diverse operation objectives, geographically dispersed infrastructure and equipment, frequent train services, and highly complicated, dynamic, and uncertain TN energy flows possessed by URTs, the research of this thesis can be extended from the following aspects.

6.2.1 Further Consideration of Timetabling and Circulation Planning

In chapters 3-4, the impact of the real-time rescheduled timetable is considered in the substation-level and network-level energy management strategies, and in chapter 5, the dwell time is incorporated into the configuration strategy. However, for energy management, the coordinated optimization of RTTR and TN power flow to fully release the regulation flexibility of traction load has not been explored further. Thus, the traction load flexibility regulation methods (e.g., train driving and real-time rescheduling of timetables and rolling stocks) and the HESS control optimization methods can be comprehensively utilized to investigate the demand response strategy of traction loads and the coordinated operation strategy of URTs. On the other hand, the impact of delay propagation should be taken into account for train operation simulation, and the adaptability of the proposed approach under delay propagation needs to be further verified. Similarly, for the configuration problem, a more thorough consideration of timetable and rolling stock scheduling, including dwell time, running time, headway, etc., is required. For passenger travel time, a refined formulation such as that of [156] can be integrated into the multi-objective configuration optimization model to improve its model performance. Furthermore, under certain situations (first/last train), the trains may skip a few stations to have a higher commercial speed. This skip-stopping strategy can also be considered in future works.

6.2.2 Implementation of Distributed Computational Architecture

In this thesis, RL has been demonstrated as a crucial base and effective tool for energy management and configuration of HESS-integrated TNs. However, it is necessary to consider the distributed deployment issue of in trains and traction substations (1st and 2nd levels) and the decentralized operation issue of DHESS-integrated TNs (3rd level) of the proposed RL-based energy management strategies. In addition, with the expansion of the URT system, the communication and data processing capability of the centralized control center can be insufficient to deal with huge training data and commands. Furthermore, the NSGA-II algorithm in Chapter 5 can have the curse of dimensionality when dealing with a large number of decision variables.

In this regard, distributed computing architecture has great potential for future applications of RL-based energy management and configuration strategies. For instance, edge computing [196], which provides computational capabilities close to the dispersed end users with internet-of-things (IoT) devices, has been regarded as a practically viable solution. By implementing edge computing, not only can the computational burden be offloaded to edge servers/devices, but the communication issue from various end users to the centralized control center can be mitigated. To further reduce the computational burden of configuration strategies, a selection process of decision variables can be conducted before the optimization, where experiments containing different combinations of decision variables can be carried out by parallel computing to speed up the selection process. Moreover, it is worth noting that the combination of RL-based strategies and IoT devices via edge computing enables artificial intelligence of things (AIoT) [197], which contributes to intelligent, efficient, carbon-neutral, and convenient URT services.

6.2.3 Investigation on Green Artificial Intelligence

With the increased adoption of AI, a more powerful AI model generally requires more energy consumption with respect to computing hardware manufacturing, model training, and model execution. In this regard, the concept of green AI [198] has been proposed to address the research and application concerns of AI-related energy consumption and environmental issues. Although the existing studies were limited, there have been various directions for achieving green AI, such as optimizing the workload during deployment, improving computational efficiency, assessing the carbon footprint, etc. In order to leverage AI for more sustainable URTs, it is necessary to incorporate green AI technologies into the proposed RL-based energy management and configuration strategies as an extension of this thesis.

Appendix A: Degradation Estimation of Hybrid Energy Storage Systems

A.1 Degradation Estimation Based on Railflow Counting

The rainflow method analyzes the cyclic loading history of a material or structure and converts the loading history into a series of closed-loop cycles, which are then added up as lifetime losses. In this work, as the supercapacitor generally has a much longer cycle life (up to 10^6) than the battery, its replacement during the system lifetime is ignored, and its rainflow counting is not implemented. For the battery case, generally, a survey which is provided by the cell manufacturer relates the number of counted cycles to the end of the battery lifetime as a function of the DoD. Thus, the battery degradation estimation process based on the survey and rainflow counting can be summarized as follows. The battery data is obtained from a LTO battery [166].

Step 1: Identify all local maximums and minimums of the SoE profile.

Step 2: Count the discharge semi-cycles from each local maximum to the minimum, as indicated by the red line in Fig. A.1.

Step 3: Similar to step 2, count the charge semi-cycles from each local minimum to the maximum.

Step 4: Match the discharge semi-cycles with the charge semi-cycles to form complete cycles. Group the cycles according to their DoDs. The relationship between DoD and battery cycle life is shown in Table A.1.

Step 5: Calculate the lifetime loss in each group and then add up to estimate the lifetime of the battery,

$$N^{\mathrm{REP}} = \left\lceil \frac{L}{L_{\mathrm{BT}}} - 1 \right\rceil, L_{\mathrm{BT}} = \min\left(L, \sum\nolimits_{i=1}^{I} \frac{C_{\mathrm{BT},i}^{\mathrm{Norm}}}{365 \cdot C_{\mathrm{BT},i}} \right), \tag{A.1}$$

where i is the index of DoD ranges, $C_{\mathrm{BT}}^{\mathrm{Norm}}$ is the amount of available life cycles, C_{BT} is the number of cycles counted per day, L_{BT} is the estimated battery lifetime, L is the system lifetime.

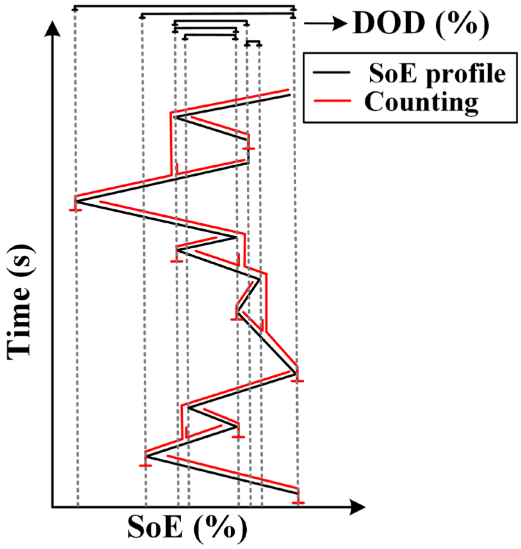


Fig. A.1 Illustration of the cycle counting process.

Table A.1 DoD ranges and battery (LTO 20Ah) cycle life.

Item	Value								
DoD (%)	<15	15-25	25-35	35-45	45-55	55-65	65-75	75-85	>85
$C_{\mathrm{BT}}^{\mathrm{Norm}}$	70000	31000	18100	11800	8100	5800	4300	3300	2500

A.2 Degradation Estimation Based on Electrothermal Coupling

The degradation estimation consists of two steps: first, the electrothermal model of the battery is developed; then, a dynamic capacity degradation model is utilized to quantify the impact of battery power, accumulated electric charge, and temperature on battery life. Due to the data availability, we use the LiFePO4 battery data reported in

[190], instead of the aforementioned LTO battery data, to model the electrothermal characteristics. Namely, the LTO battery data is used for case studies in chapters 3–4, and LiFePO4 battery data is used for case studies in chapter 5.

For the first step, considering that the HESS contains many battery modules in series and parallel, we assume that the operating state of cells is the same, and the space between cells is large enough. In order to analyze the thermal behavior of the battery, the simple equivalent circuit model in Fig. 3.2 is replaced by a first-order equivalent circuit model (Fig. A.2), where a constant RC network is added to capture the battery relaxation process.

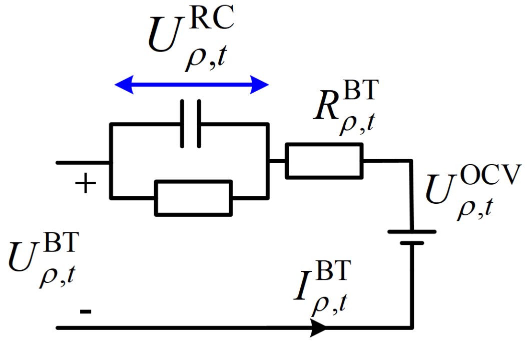

$U_{\rho,t}^{\mathrm{RC}}$
$R_{\rho,t}^{\mathrm{BT}}$
+
$U_{\rho,t}^{\mathrm{OCV}}$
$U_{\rho,t}^{\mathrm{BT}}$
$I_{\rho,t}^{\mathrm{BT}}$
-

Fig. A.2 Battery first-order equivalent circuit model.

According to [190], the battery OCV $U_{\rho,t}^{\mathrm{OCV}}$ is insensitive to internal temperature $T_{\rho,t}^{\mathrm{BT,in}}$. Besides, as the impact of SoE $\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}$ on internal resistance $R_{\rho,t}^{\mathrm{BT}}$ is relatively minor, we only model the relationship between temperature and internal resistance. Thus, the SoE-OCV and temperature-resistance relationship of the LiFePO4 battery can be fitted by

$$\begin{aligned} U_{\rho,t}^{\mathrm{OCV}} = {} & 1.769\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^4 - 2.593\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^3 \\ & +0.874\left(\mathrm{SoE}_{\rho,t}^{\mathrm{BT}}\right)^2 +0.222\mathrm{SoE}_{\rho,t}^{\mathrm{BT}} + 3.174, \end{aligned} \tag{A.2}$$

$$
\begin{aligned}
R_{\rho,t}^{\mathrm{BT}} &= 8.684\times 10^{-11}\left(T_{\rho,t}^{\mathrm{BT,in}}\right)^{6} - 1.683\times 10^{-8}\left(T_{\rho,t}^{\mathrm{BT,in}}\right)^{5} \\
&+1.289\times 10^{-6}\left(T_{\rho,t}^{\mathrm{BT,in}}\right)^{4} - 4.920\times 10^{-5}\left(T_{\rho,t}^{\mathrm{BT,in}}\right)^{3} \\
&+9.746\times 10^{-4}\left(T_{\rho,t}^{\mathrm{BT,in}}\right)^{2} - 9.519\times 10^{-3} T_{\rho,t}^{\mathrm{BT,in}} + 0.052.
\end{aligned}
\tag{A.3}
$$

The rest of the model is formulated by

$$U_{\rho,t}^{\mathrm{RC}} = 0.982 U_{\rho,t-1}^{\mathrm{RC}} + 2.1\times 10^{-4} I_{\rho,t-1}^{\mathrm{BT}}, \tag{A.4}$$

$$U_{\rho,t}^{\mathrm{BT}} = U_{\rho,t}^{\mathrm{OCV}} - I_{\rho,t}^{\mathrm{BT}} R_{\rho,t}^{\mathrm{BT}} - U_{\rho,t}^{\mathrm{RC}}, \tag{A.5}$$

$$P_{\rho,t}^{\mathrm{BT}} = I_{\rho,t}^{\mathrm{BT}}\left(U_{\rho,t}^{\mathrm{OCV}} - I_{\rho,t}^{\mathrm{BT}} R_{\rho,t}^{\mathrm{BT}} - U_{\rho,t}^{\mathrm{RC}}\right), \tag{A.6}$$

where $P_{\rho,t}^{\mathrm{BT}}$ is the charge or discharge power, $U_{\rho,t}^{\mathrm{RC}}$ is the voltage of the RC network. The coefficients in (A.4) are obtained by field tests.

The thermal model of the battery can be referred to [190], and we also provide it below

$$264.7 T_{\rho,t}^{\mathrm{BT,in}} = \left(I_{\rho,t}^{\mathrm{BT}}\right)^{2} R_{\rho,t}^{\mathrm{BT}} - 1.286\left(T_{\rho,t-1}^{\mathrm{BT,in}} - T_{\rho,t-1}^{\mathrm{BT,sh}}\right), \tag{A.7}$$

$$30.7 T_{\rho,t}^{\mathrm{BT,sh}} = 1.286\left(T_{\rho,t-1}^{\mathrm{BT,in}} - T_{\rho,t-1}^{\mathrm{BT,sh}}\right) - 0.3009\left(T_{\rho,t-1}^{\mathrm{BT,sh}} - T^{\mathrm{am}}\right), \tag{A.8}$$

where $T_{\rho,t}^{\mathrm{BT,sh}}$ is the shell temperature, T^{am} is the ambient temperature. The coefficients in (A.5)–(A.6) are obtained by field tests.

It is also worth noting that, as the studied URT stations have HVAC systems, we assume that the ambient temperature is constant during operation. Besides, manufacturers have fully tested the thermal stability of batteries and supercapacitors. Therefore, the thermal management issue and the thermal constraint of DHESSs are out of the scope, and this thesis only considers the impact of temperature rise on the degradation.

For the second step, a dynamic capacity degradation model [167] is utilized, as formulated below

$$N^{\mathrm{REP}} = \left\lceil \frac{L}{L_{\mathrm{BT}}} - 1 \right\rceil, L_{\mathrm{BT}} = \min\left(L, \frac{20}{365 \cdot \Delta Q^{\mathrm{BT}}} \right), \tag{A.9}$$

$$\Delta Q^{\mathrm{BT}} = \sum_{t=1}^{86400} \left[\left(\frac{1}{3600} \left| I_t^{\mathrm{BT}} \right| \right)^{0.824} \times 0.0032 \exp\left(\frac{-15162 + 1516C\text{-rate}}{0.824 R^{\mathrm{G}} T_t^{\mathrm{BT,in}}} \right) \right], \tag{A.10}$$

where ΔQ^{BT} is the percentage capacity loss. Generally, 20% capacity loss is regarded as the end of battery lifetime. ΔQ^{BT} is calculated at a daily basis. R^{G} is the ideal gas constant, $R^{\mathrm{G}} = 8.31$ J/(mol·K). C-rate is the charge or discharge rate. Other coefficients in (A.10) are obtained by field tests.

Appendix B: Calculation of Passenger Demands

A passenger demand calculation method is provided in this appendix and the aim is to find the average number of onboard passengers during each time interval *n*. According to the predetermined arrival rates and OD table, the arrival passengers are distributed to each station and then to each in-service train. Assume only one train direction (e.g., up or down) is considered,

$$N_{k,i,j,\rho}^{\mathrm{W}} = H_k \alpha_{i,j} \beta_{i,n}, \tag{B.1}$$

$$N_{k,i,\rho}^{\mathrm{W}} = \begin{cases} \sum_{j=i+1}^{I} N_{k,i,j,\rho}^{\mathrm{W}}, & k=1, \\ N_{k-1,i,\rho}^{\mathrm{W}} - N_{k-1,i,\rho}^{\mathrm{ON}} + \sum_{j=i+1}^{I} N_{k,i,j,\rho}^{\mathrm{W}}, & k \neq 1, \end{cases} \tag{B.2}$$

$$N_{k,i,\rho}^{\mathrm{ON}} = \begin{cases} \min\left(N_{k,i,\rho}^{\mathrm{W}}, N_{\max}^{\mathrm{B}}\right), & i=1, \\ \min\left(N_{k,i,\rho}^{\mathrm{W}}, N_{\max}^{\mathrm{B}} - \sum_{j=1}^{i-1} N_{k,j,\rho}^{\mathrm{ON}} + \sum_{j=2}^{i} N_{k,j,\rho}^{\mathrm{OFF}}\right), & i \neq 1, \end{cases} \tag{B.3}$$

$$N_{k,j,\rho}^{\mathrm{OFF}} = \sum_{i=1}^{j-1} N_{k,i,j,\rho}^{\mathrm{ON}} = \sum_{i=1}^{j-1} \frac{N_{k,i,j,\rho}^{\mathrm{W}}}{N_{k,i,\rho}^{\mathrm{W}}} N_{k,i,\rho}^{\mathrm{ON}}, \tag{B.4}$$

$$N_{k,i,\rho}^{\mathrm{B}} = \begin{cases} N_{k,i,\rho}^{\mathrm{ON}}, & i=1, \\ N_{k,i-1,\rho}^{\mathrm{B}} + N_{k,i,\rho}^{\mathrm{ON}} - N_{k,i,\rho}^{\mathrm{OFF}}, & i \neq 1, \end{cases} \tag{B.5}$$

$$N_{i,\rho,n}^{\mathrm{B}} = \frac{1}{K_n^{\mathrm{N}}} \sum_{k=1}^{K_n} N_{k,i,\rho}^{\mathrm{B}}, \tag{B.6}$$

where the total waiting passengers $N_{k,i,\rho}^{\mathrm{W}}$ at station *i* is calculated by (B.1)–(B.2), H_k is the headway of train *k*, $\alpha_{i,j}$ is the OD element, $\beta_{i,n}$ is the arrival rate, $N_{k,i,j,\rho}^{\mathrm{W}}$ is the proportion of $N_{k,i,\rho}^{\mathrm{W}}$ who travel from station *i* to *j*. Then, the passengers getting on $N_{k,i,\rho}^{\mathrm{ON}}$ and off $N_{k,i,\rho}^{\mathrm{OFF}}$ can be calculated by (B.3)–(B.4), $N_{\max}^{\mathrm{B}}$ is the maximum train capacity. The onboard passengers $N_{k,i,\rho}^{\mathrm{B}}$ on each train can be obtained by (B.5). Finally, the average onboard passengers $N_{i,\rho,n}^{\mathrm{B}}$ is obtained by (B.6), K_{n}^{N} is the number of trains running at interval *n*.

List of References

[1] *Regulation for Operation Technology of Urban Rail Transit*, GB/T 38707-2020, State Administration for Market Regulation (China), 2020.

[2] S. Verkehr. "Global urban rail passenger traffic from 2005 to 2025 (in billion passenger-kilometers)." Statista. https://www.statista.com/statistics/739801/urban-rail-passenger-transport-performance/ (accessed 30 Nov. 2024).

[3] A. Feng, "Data statistics and development analysis of China urban rail transit in 2021 (in Chinese)," *Tunnel Construction,* vol. 42, no. 2, p. 12, 2022.

[4] L. S. Chan, "Transition from fossil fuel propelled transport to electrified mass transit railway system—Experience from Hong Kong," *Energy Policy,* vol. 173, p. 113372, 2023.

[5] Q. Tan, D. He, Z. Sun, Z. Yao, J. Zhou, and T. Chen, "A deep reinforcement learning based metro train operation control optimization considering energy conservation and passenger comfort," *Engineering Research Express,* 2025.

[6] W. Li, L. Bao, Y. Li, H. Si, and Y. Li, "Assessing the transition to low-carbon urban transport: A global comparison," *Resources, Conservation and Recycling,* vol. 180, p. 106179, 2022.

[7] M. Khodaparastan, A. A. Mohamed, and W. Brandauer, "Recuperation of regenerative braking energy in electric rail transit systems," *IEEE Transactions on Intelligent Transportation Systems,* vol. 20, no. 8, pp. 2831–2847, 2019.

[8] H. Douglas, C. Roberts, S. Hillmansen, and F. Schmid, "An assessment of available measures to reduce traction energy use in railway networks," *Energy Conversion and Management,* vol. 106, pp. 1149–1165, 2015.

[9] A. González-Gil, R. Palacin, P. Batty, and J. P. Powell, "A systems approach to reduce urban rail energy consumption," *Energy Conversion and Management,* vol. 80, pp. 509–524, 2014.

[10] B. Guan, H. Yang, H. Li, H. Gao, T. Zhang, and X. Liu, "Energy consumption characteristics and rooftop photovoltaic potential assessment of elevated metro station," *Sustainable Cities and Society,* vol. 99, p. 104928, 2023.

[11] B. Guan, H. Yang, T. Zhang, X. Liu, and X. Wang, "Technoeconomic analysis of rooftop PV system in elevated metro station for cost-effective operation and clean electrification," *Renewable Energy,* vol. 226, p. 120305, 2024.

[12] S. A. Al-Janahi, O. Ellabban, and S. G. Al-Ghamdi, "Technoeconomic feasibility study of grid-connected building-integrated photovoltaics system for clean electrification: A case study of Doha metro," *Energy Reports,* vol. 6, pp. 407–414, 2020.

[13] H. Yang, B. Guan, J. Zhang, T. Zhang, X. Liu, and X. Wang, "Application potential of rooftop

photovoltaics (PV) in elevated metro station for a low-carbon future: Characteristic analysis and strategies for supply-demand mismatch," *Renewable Energy,* vol. 238, p. 121983, 2025.

[14] C. Wu, B. Han, S. Lu, F. Xue, and F. Zhong, "Carbon-reducing train rescheduling method for urban railway systems considering the grid with wind power supply," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, Macau, China, 2022, pp. 164–169.

[15] A. Ruvio *et al.*, "Integrated procedure to design optimal hybrid renewable power plant for railways' traction power substation," *Sustainable Energy, Grids and Networks,* vol. 39, p. 101446, 2024.

[16] H. Yu, Y. Wang, and Z. Chen, "A novel renewable microgrid-enabled metro traction power system—Concepts, framework, and operation strategy," *IEEE Transactions on Transportation Electrification,* vol. 7, no. 3, pp. 1733–1749, 2021.

[17] M. S. Simoiu, I. Fagarasan, S. Ploix, and V. Calofir, "Optimising the self-consumption and self-sufficiency: A novel approach for adequately sizing a photovoltaic plant with application to a metropolitan station," *Journal of Cleaner Production,* vol. 327, p. 129399, 2021.

[18] G. Li and S. W. Or, "DRL-based adaptive energy management for hybrid electric storage systems under dynamic spatial-temporal traffic in urban rail transits," in *2023 IEEE International Conference on Energy Technologies for Future Grids (ETFG)*, Wollongong, Austalia, 2023, pp. 1–6.

[19] G. M. S. Kumar and S. Cao, "Leveraging energy flexibilities for enhancing the cost-effectiveness and grid-responsiveness of net-zero-energy metro railway and station systems," *Applied Energy,* vol. 333, p. 120632, 2023.

[20] X. Shen, H. Wei, and L. Wei, "Study of trackside photovoltaic power integration into the traction power system of suburban elevated urban rail transit line," *Applied Energy,* vol. 260, p. 114177, 2020.

[21] D. Feng, H. Zhu, F. Wang, X. Sun, S. Lin, and Z. He, "Evaluation of voltage quality and energy saving benefits of urban rail transit power supply systems considering the access of photovoltaics," *CSEE Journal of Power and Energy Systems,* vol. 9, no. 6, pp. 2309–2320, 2023.

[22] L. Jian and C. Min, "Application of solar PV grid-connected power generation system in Shanghai rail transit," in *2018 China International Conference on Electricity Distribution (CICED)*, Tianjin, China, 2018, pp. 110–113.

[23] A. González-Gil, R. Palacin, and P. Batty, "Sustainable urban rail systems: Strategies and technologies for optimal management of regenerative braking energy," *Energy Conversion and Management,* vol. 75, pp. 374–388, 2013.

[24] D. Cornic, "Efficient recovery of braking energy through a reversible dc substation," in *Electrical Systems for Aircraft, Railway, and Ship Propulsion*, 2010: IEEE, pp. 1–9.

[25] N. Boizumeau Jr and P. Leguay, "Overview of braking energy recovery technologies in the public transport field," in *Workshop on Braking Energy Recovery Systems-Ticket to Kyoto Project*, 2011, no. March.

[26] H. Ibaiondo and A. Romo, "Kinetic energy recovery on railway systems with feedback to the grid," in *Proceedings of 14th International Power Electronics and Motion Control Conference EPE-PEMC 2010*, Ohrid, Macedonia, 2010, pp. T9-94–T9-97.

[27] X. Yang, X. Li, B. Ning, and T. Tang, "A survey on energy-efficient train operation for urban rail transit," *IEEE Transactions on Intelligent Transportation Systems,* vol. 17, no. 1, pp. 2–13, 2016.

[28] Z. Zhang, H. Zhao, X. Yao, Z. Xing, and X. Liu, "Metro timetable optimization for improving regenerative braking energy utilization efficiency," *Journal of Cleaner Production,* vol. 450, p. 141970, 2024.

[29] J. Wang, D. Huang, Y. Hu, N. Qin, and Y. Zhu, "Modeling and SOC estimation of on-board energy storage device for trains under emergency traction," *Journal of Energy Storage,* vol. 100, p. 113414, 2024.

[30] A. Çiçek *et al.*, "Integrated rail system and EV parking lot operation with regenerative braking energy, energy storage system and PV availability," *IEEE Transactions on Smart Grid,* vol. 13, no. 4, pp. 3049–3058, 2022.

[31] A. Bettinelli, A. Santini, and D. Vigo, "A real-time conflict solution algorithm for the train rescheduling problem," *Transportation Research Part B: Methodological,* vol. 106, pp. 237–265, 2017.

[32] J. Yin, D. Chen, L. Yang, T. Tang, and B. Ran, "Efficient real-time train operation algorithms with uncertain passenger demands," *IEEE Transactions on Intelligent Transportation Systems,* vol. 17, no. 9, pp. 2600–2612, 2015.

[33] P. G. Tzouras, H. Farah, E. Papadimitriou, N. van Oort, and M. Hagenzieker, "Tram drivers perceived safety and driving stress evaluation. A stated preference experiment," *Transportation Research Interdisciplinary Perspectives,* vol. 7, p. 100205, 2020.

[34] S. M. M. Gazafrudi, A. T. Langerudy, E. F. Fuchs, and K. Al-Haddad, "Power quality issues in railway electrification: A comprehensive perspective," *IEEE Transactions on Industrial Electronics,* vol. 62, no. 5, pp. 3081–3090, 2015.

[35] A. K. Shakya, G. Pillai, and S. Chakrabarty, "Reinforcement learning algorithms: A brief survey," *Expert Systems with Applications,* vol. 231, p. 120495, 2023.

[36] A. Australia. "ABB recycles spare energy in Melbourne’s rail network." Electric & hybrid

rail technology. https://www.electricandhybridrail.com/content/opinion/abb-recycles-spare-energy-in-melbourne-s-rail-network/ (accessed Jan 31, 2025).

[37] Y. Liu, Z. Yang, X. Wu, D. Sha, F. Lin, and X. Fang, "An adaptive energy management strategy of stationary hybrid energy storage system," *IEEE Transactions on Transportation Electrification,* vol. 8, no. 2, pp. 2261–2272, 2022.

[38] T. Ratniyomchai, S. Hillmansen, and P. Tricoli, "Recent developments and applications of energy storage devices in electrified railways," *IET Electrical Systems in Transportation,* vol. 4, no. 1, pp. 9–20, 2014.

[39] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *Journal of Machine Learning Research,* vol. 21, no. 1, pp. 7234–7284, 2020.

[40] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control*, Cham, Switzerland: Springer, Cham, 2021, pp. 321–384.

[41] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning,* vol. 8, pp. 279–292, 1992.

[42] V. Mnih, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602,* 2013.

[43] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, Phoenix, AZ, USA, 2016, vol. 30, no. 1, pp. 2094–2100.

[44] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952,* 2015.

[45] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning*, New York, NY, USA , 2016, pp. 1995–2003.

[46] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," *arXiv preprint arXiv:1507.06527,* 2015.

[47] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems,* vol. 12, 1999.

[48] V. Mnih, "Asynchronous methods for deep reinforcement learning," *arXiv preprint arXiv:1602.01783,* 2016.

[49] J. Schulman, "Trust region policy optimization," *arXiv preprint arXiv:1502.05477,* 2015.

[50] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347,* 2017.

[51] N. Casas, "Deep deterministic policy gradient for urban traffic light control," *arXiv preprint arXiv:1703.09035,* 2017.

[52] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholmsmässan, Sweden, 2018, pp. 1587–1596.

[53] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems," *The Knowledge Engineering Review,* vol. 27, no. 1, pp. 1–31, 2012.

[54] S. Sukhbaatar and R. Fergus, "Learning multiagent communication with backpropagation," *Advances in Neural Information Processing Systems,* vol. 29, 2016.

[55] P. Peng *et al.*, "Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games," *arXiv preprint arXiv:1703.10069,* 2017.

[56] P. Sunehag *et al.*, "Value-decomposition networks for cooperative multi-agent learning," *arXiv preprint arXiv:1706.05296,* 2017.

[57] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018, vol. 32, no. 1, pp. 2974–2982.

[58] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems,* vol. 30, 2017.

[59] J. Ackermann, V. Gabler, T. Osa, and M. Sugiyama, "Reducing overestimation bias in multi-agent domains using double centralized critics," *arXiv preprint arXiv:1910.01465,* 2019.

[60] F. Fathinezhad, V. Derhami, and M. Rezaeian, "Supervised fuzzy reinforcement learning for robot navigation," *Applied Soft Computing,* vol. 40, pp. 33–41, 2016.

[61] Y. Lin *et al.*, "A survey on reinforcement learning for recommender systems," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 35, no. 10, pp. 13164–13184, 2024.

[62] F. Liu, R. Tang, H. Guo, X. Li, Y. Ye, and X. He, "Top-aware reinforcement learning based recommendation," *Neurocomputing,* vol. 417, pp. 255–269, 2020.

[63] C. Shiranthika, K. W. Chen, C. Y. Wang, C. Y. Yang, B. H. Sudantha, and W. F. Li, "Supervised optimal chemotherapy regimen based on offline reinforcement learning," *IEEE Journal of Biomedical and Health Informatics,* vol. 26, no. 9, pp. 4763–4772, 2022.

[64] L. Wang, W. Zhang, X. He, and H. Zha, "Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation," in *Proceedings of the 24th ACM*

SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 2018, pp. 2447–2456.

[65] J. Wang and F. Zhu, "ExSelfRL: An exploration-inspired self-supervised reinforcement learning approach to molecular generation," *Expert Systems with Applications,* vol. 260, p. 125410, 2025.

[66] C. Qi *et al.*, "Self-supervised reinforcement learning-based energy management for a hybrid electric vehicle," *Journal of Power Sources,* vol. 514, p. 230584, 2021.

[67] L. Espeholt *et al.*, "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholmsmässan, Sweden, 2018, pp. 1407–1416.

[68] M. Hessel, H. Soyer, L. Espeholt, W. Czarnecki, S. Schmitt, and H. Van Hasselt, "Multi-task deep reinforcement learning with popart," in *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 2019, vol. 33, no. 01, pp. 3796–3803.

[69] R. Yang, H. Xu, Y. Wu, and X. Wang, "Multi-task reinforcement learning with soft modularization," *Advances in Neural Information Processing Systems,* vol. 33, pp. 4767–4777, 2020.

[70] Y. Teh *et al.*, "Distral: Robust multitask reinforcement learning," *Advances in Neural Information Processing Systems,* vol. 30, 2017.

[71] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems,* vol. 33, pp. 5824–5836, 2020.

[72] B. Liu, Y. Feng, P. Stone, and Q. Liu, "Famo: Fast adaptive multitask optimization," *Advances in Neural Information Processing Systems,* vol. 36, 2024.

[73] A. Jha, A. Kumar, B. Banerjee, and S. Chaudhuri, "Adamt-net: An adaptive weight learning based multi-task learning model for scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, 2020, pp. 706–707.

[74] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone, "Curriculum learning for reinforcement learning domains: A framework and survey," *Journal of Machine Learning Research*, vol. 21, no. 181, pp. 1–50, 2020.

[75] Z. Bai, P. Hao, W. ShangGuan, B. Cai, and M. J. Barth, "Hybrid reinforcement learning-based eco-driving strategy for connected and automated vehicles at signalized Intersections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15850–15863, 2022.

[76] S. Khaitan and J. M. Dolan, "State dropout-based curriculum reinforcement learning for self-

driving at unsignalized intersections," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, 2022, pp. 12219‒12224.

[77] X. Shi, D. Jiang, H. Liu, and X. Hu, "Research on energy management optimization of hybrid electric vehicles based on improved curriculum learning," *Energy*, vol. 324, p. 136061, 2025.

[78] A. R. Albrecht, P. G. Howlett, P. J. Pudney, and X. Vu, "Energy-efficient train control: From local convexity to global optimization and uniqueness," *Automatica,* vol. 49, no. 10, pp. 3072–3078, 2013.

[79] J. Yin, T. Tang, L. Yang, J. Xun, Y. Huang, and Z. Gao, "Research and development of automatic train operation for railway transportation systems: A survey," *Transportation Research Part C: Emerging Technologies,* vol. 85, pp. 548–572, 2017.

[80] E. Khmelnitsky, "On an optimal control problem of train operation," *IEEE Transactions on Automatic Control,* vol. 45, no. 7, pp. 1257–1266, 2000.

[81] C. Jiaxin and P. Howlett, "A note on the calculation of optimal strategies for the minimization of fuel consumption in the control of trains," *IEEE Transactions on Automatic Control,* vol. 38, no. 11, pp. 1730–1734, 1993.

[80] P. G. Howlett, P. J. Pudney, and X. Vu, "Local energy minimization in optimal train control," *Automatica,* vol. 45, no. 11, pp. 2692–2698, 2009.

[83] A. Albrecht, P. Howlett, P. Pudney, X. Vu, and P. Zhou, "The key principles of optimal train control—Part 1: Formulation of the model, strategies of optimal type, evolutionary lines, location of optimal switching points," *Transportation Research Part B: Methodological,* vol. 94, pp. 482–508, 2016.

[84] A. Albrecht, P. Howlett, P. Pudney, X. Vu, and P. Zhou, "The key principles of optimal train control—Part 2: Existence of an optimal strategy, the local energy minimization principle, uniqueness, computational techniques," *Transportation Research Part B: Methodological,* vol. 94, pp. 509–538, 2016.

[85] K. Ichikawa, "Application of optimization theory for bounded state variable problems to the operation of train," *Bulletin of JSME,* vol. 11, no. 47, pp. 857–865, 1968.

[86] X. Luan, Y. Wang, B. De Schutter, L. Meng, G. Lodewijks, and F. Corman, "Integration of real-time traffic management and train control for rail networks-part 1: Optimization problems and solution approaches," *Transportation Research Part B: Methodological,* vol. 115, pp. 41–71, 2018.

[87] X. Luan, Y. Wang, B. De Schutter, L. Meng, G. Lodewijks, and F. Corman, "Integration of real-time traffic management and train control for rail networks-Part 2: Extensions towards energy-efficient train operations," *Transportation Research Part B: Methodological,* vol. 115, pp. 72–94, 2018.

[88] H. Ko, T. Koseki, and M. Miyatake, "Application of dynamic programming to the optimization of the running profile of a train," *WIT Transactions on The Built Environment,* vol. 74, 2004.

[89] J. T. Haahr, D. Pisinger, and M. Sabbaghian, "A dynamic programming approach for optimizing train speed profiles with speed restrictions and passage points," *Transportation Research Part B: Methodological,* vol. 99, pp. 167–182, 2017.

[90] P. Wang, A. Trivella, R. M. Goverde, and F. Corman, "Train trajectory optimization for improved on-time arrival under parametric uncertainty," *Transportation Research Part C: Emerging Technologies,* vol. 119, p. 102680, 2020.

[91] P. Wang and R. M. Goverde, "Multiple-phase train trajectory optimization with signalling and operational constraints," *Transportation Research Part C: Emerging Technologies,* vol. 69, pp. 255–275, 2016.

[92] C. Chang and S. Sim, "Optimising train movements through coast control using genetic algorithms," *IEE Proceedings-Electric Power Applications,* vol. 144, no. 1, pp. 65–73, 1997.

[93] K. Wong and T. K. Ho, "Dynamic coast control of train movement with genetic algorithm," *International Journal of Systems Science,* vol. 35, no. 13–14, pp. 835–846, 2004.

[94] C. Sicre, A. Cucala, and A. Fernández-Cardador, "Real time regulation of efficient driving of high speed trains based on a genetic algorithm and a fuzzy model of manual driving," *Engineering Applications of Artificial Intelligence,* vol. 29, pp. 79–92, 2014.

[95] X. Yang, A. Chen, B. Ning, and T. Tang, "A stochastic model for the integrated optimization on metro timetable and speed profile with uncertain train mass," *Transportation Research Part B: Methodological,* vol. 91, pp. 424–445, 2016.

[96] S. Lu, S. Hillmansen, T. K. Ho, and C. Roberts, "Single-train trajectory optimization," *IEEE Transactions on Intelligent Transportation Systems,* vol. 14, no. 2, pp. 743–750, 2013.

[97] J. Yin, D. Chen, and L. Li, "Intelligent train operation algorithms for subway by expert system and reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems,* vol. 15, no. 6, pp. 2561–2571, 2014.

[98] Y. Wang, B. Ning, T. Van den Boom, and B. De Schutter, “Background: train operations and scheduling”, in *Optimal Trajectory Planning and Train Scheduling for Urban Rail Transit Systems*. Cham, Switzerland: Springer, Cham, 2016, pp. 7–21.

[99] J. Huang, E. Zhang, J. Zhang, S. Huang, and Z. Zhong, "Deep reinforcement learning based train driving optimization," in *2019 Chinese Automation Congress (CAC)*, Hangzhou, China, 2019, pp. 2375–2381.

[100] K. Zhou, S. Song, A. Xue, K. You, and H. Wu, "Smart train operation algorithms based on expert knowledge and reinforcement learning," *IEEE Transactions on Systems, Man, and*

Cybernetics: Systems, vol. 52, no. 2, pp. 716–727, 2022.

[101] M. Shang, Y. Zhou, and H. Fujita, "Deep reinforcement learning with reference system to handle constraints for energy-efficient train control," *Information Sciences,* vol. 570, pp. 708–721, 2021.

[102] Ş. İ *et al.*, "Energy management of a smart railway station considering regenerative braking and stochastic behaviour of ESS and PV generation," *IEEE Transactions on Sustainable Energy,* vol. 9, no. 3, pp. 1041–1050, 2018.

[103] Y. Liu, M. Chen, Z. Cheng, Y. Chen, and Q. Li, "Robust energy management of high-speed railway co-phase traction substation with uncertain PV generation and traction load," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 6, pp. 5079–5091, 2022.

[104] H. Novak, V. Lešić, and M. Vašak, "Hierarchical model predictive control for coordinated electric railway traction system energy management," *IEEE Transactions on Intelligent Transportation Systems,* vol. 20, no. 7, pp. 2715–2727, 2019.

[105] A. Zahedmanesh, K. M. Muttaqi, and D. Sutanto, "A sequential decision-making process for optimal technoeconomic operation of a grid-connected electrical traction substation integrated with solar PV and BESS," *IEEE Transactions on Industrial Electronics,* vol. 68, no. 2, pp. 1353–1364, 2021.

[106] J. Tian *et al.*, "Energy-saving optimal scheduling under multi-mode “source-network-load-storage” combined system in metro station based on modified Gray Wolf Algorithm," *Archives of Electrical Engineering,* vol. 73, no. 1, pp. 121–143, 2024.

[107] P. J. Grbovic, P. Delarue, P. Le Moigne, and P. Bartholomeus, "Modeling and control of the ultracapacitor-based regenerative controlled electric drives," *IEEE Transactions on Industrial Electronics,* vol. 58, no. 8, pp. 3471–3484, 2010.

[108] P. J. Grbović, P. Delarue, P. L. Moigne, and P. Bartholomeus, "The ultracapacitor-based controlled electric drives with braking and ride-through capability: Overview and analysis," *IEEE Transactions on Industrial Electronics,* vol. 58, no. 3, pp. 925–936, 2011.

[109] F. Ciccarelli, D. Iannuzzi, K. Kondo, and L. Fratelli, "Line-voltage control based on wayside energy storage systems for tramway networks," *IEEE Transactions on Power Electronics,* vol. 31, no. 1, pp. 884–899, 2016.

[110] D. Ramsey, T. Letrouve, A. Bouscayrol, and P. Delarue, "Comparison of energy recovery solutions on a suburban DC railway system," *IEEE Transactions on Transportation Electrification,* vol. 7, no. 3, pp. 1849–1857, 2021.

[111] F. Zhu, Z. Yang, F. Lin, and Y. Xin, "Dynamic threshold adjustment strategy of supercapacitor energy storage system based on no-load voltage identification in urban rail transit," in *2019 IEEE Transportation Electrification Conference and Expo, Asia-Pacific (ITEC Asia-Pacific)*,

Seogwipo, South Korea, 2019, pp. 1–6.

[112] H. H. Alnuman, D. T. Gladwin, M. P. Foster, and E. M. Ahmed, "Enhancing energy management of a stationary energy storage system in a DC electric railway using fuzzy logic control," *International Journal of Electrical Power & Energy Systems,* vol. 142, p. 108345, 2022.

[113] Y. Liu *et al.*, "Adaptive threshold adjustment strategy based on fuzzy logic control for ground energy storage system in urban rail transit," *IEEE Transactions on Vehicular Technology,* vol. 70, no. 10, pp. 9945–9956, 2021.

[114] D. Iannuzzi, D. Lauria, and P. Tricoli, "Optimal design of stationary supercapacitors storage devices for light electrical transportation systems," *Optimization and Engineering,* vol. 13, pp. 689–704, 2012.

[115] T. Ratniyomchai, S. Hillmansen, and P. Tricoli, "Energy loss minimisation by optimal design of stationary supercapacitors for light railways," in *2015 international conference on clean electrical power (ICCEP)*, Taormina, Italy, 2015, pp. 511–517.

[116] E. Bilbao, P. Barrade, I. Etxeberria-Otadui, A. Rufer, S. Luri, and I. Gil, "Optimal energy management strategy of an improved elevator with energy storage capacity based on dynamic programming," *IEEE Transactions on Industry Applications,* vol. 50, no. 2, pp. 1233–1244, 2013.

[117] F. Zhu, Z. Yang, H. Xia, and F. Lin, "Hierarchical control and full-range dynamic performance optimization of the supercapacitor energy storage system in urban railway," *IEEE Transactions on Industrial Electronics,* vol. 65, no. 8, pp. 6646–6656, 2018.

[118] S. Lu *et al.*, "Energy-efficient train control considering energy storage devices and traction power network using a model predictive control framework," *IEEE Transactions on Transportation Electrification,* vol. 10, no. 4, pp. 10451–10467, 2024.

[119] H. Dong, Z. Tian, J. W. Spencer, D. Fletcher, and S. Hajiabady, "Bi-level optimization of sizing and control strategy of hybrid energy storage system in urban rail transit considering substation operation stability," *IEEE Transactions on Transportation Electrification,* vol. 10, no. 4, pp. 10102–10114, 2024.

[120] Z. Yang, F. Zhu, and F. Lin, "Deep-reinforcement-learning-based energy management strategy for supercapacitor energy storage systems in urban rail transit," *IEEE Transactions on Intelligent Transportation Systems,* vol. 22, no. 2, pp. 1150–1160, 2020.

[121] X. Wang, Y. Luo, B. Qin, and L. Guo, "Power allocation strategy for urban rail HESS based on deep reinforcement learning sequential decision optimization," *IEEE Transactions on Transportation Electrification,* vol. 9, no. 2, pp. 2693–2710, 2022.

[122] J. Luo, S. Gao, X. Wei, Z. Tian, and X. Wang, "Parallel-reinforcement-learning-based online

energy management strategy for energy storage traction substations in electrified railroad," *IEEE Transactions on Transportation Electrification,* vol. 10, no. 1, pp. 2112–2123, 2023.

[123] J. A. Aguado, A. J. S. Racero, and S. d. l. Torre, "Optimal operation of electric railways with renewable energy and electric storage systems," *IEEE Transactions on Smart Grid,* vol. 9, no. 2, pp. 993–1001, 2018.

[124] C. F. Calvillo, A. Sánchez-Miralles, J. Villar, and F. Martín, "Impact of EV penetration in the interconnected urban environment of a smart city," *Energy,* vol. 141, pp. 2218–2233, 2017.

[125] S. Liu, C. Lu, and G. He, "Distributed electric bicycle batteries for subway station energy management as a virtual power plant," *Applied Energy,* vol. 370, p. 123094, 2024.

[126] M. Roustaei *et al.*, "Enhancing smart city operation management: Integrating energy systems with a subway synergism hub," *Sustainable Cities and Society,* vol. 107, p. 105446, 2024.

[127] H. Xia, H. Chen, Z. Yang, F. Lin, and B. Wang, "Optimal energy management, location and size for stationary energy storage system in a metro line based on genetic algorithm," *Energies,* vol. 8, no. 10, pp. 11618–11640, 2015.

[128] Q. Qin, T. Guo, F. Lin, and Z. Yang, "Energy transfer strategy for urban rail transit battery energy storage system to reduce peak power of traction substation," *IEEE Transactions on Vehicular Technology,* vol. 68, no. 12, pp. 11714–11724, 2019.

[129] X. Wang, Y. Luo, Y. Zhou, Y. Qin, and B. Qin, "Hybrid energy management strategy based on dynamic setting and coordinated control for urban rail train with PMSM," *IET Renewable Power Generation,* vol. 15, no. 12, pp. 2740–2752, 2021.

[130] Y. Zhao, Z. Zhong, F. Lin, and Z. Yang, "Multi time scale management and coordination strategy for stationary super capacitor energy storage in urban rail transit power supply system," *Electric Power Systems Research,* vol. 228, p. 110046, 2024.

[131] F. Zhu, Z. Yang, F. Lin, and Y. Xin, "Decentralized cooperative control of multiple energy storage systems in urban railway based on multiagent deep reinforcement learning," *IEEE Transactions on Power Electronics,* vol. 35, no. 9, pp. 9368–9379, 2020.

[132] H. Dong, Z. Tian, J. W. Spencer, D. Fletcher, and S. Hajiabady, "Coordinated control strategy of railway multisource traction system with energy storage and renewable energy," *IEEE Transactions on Intelligent Transportation Systems,* pp. 1–12, 2023.

[133] S. Khayyam, N. Berr, L. Razik, M. Fleck, F. Ponci, and A. Monti, "Railway system energy management optimization demonstrated at offline and online case studies," *IEEE Transactions on Intelligent Transportation Systems,* vol. 19, no. 11, pp. 3570–3583, 2018.

[134] L. Razik, N. Berr, S. Khayyam, F. Ponci, and A. Monti, "REM-S–Railway Energy Management in Real Rail Operation," *IEEE Transactions on Vehicular Technology,* vol. 68, no. 2, pp. 1266–1277, 2019.

[135] J. Chen *et al.*, "Integrated regenerative braking energy utilization system for multi-substations in electrified railways," *IEEE Transactions on Industrial Electronics,* vol. 70, no. 1, pp. 298–310, 2023.

[136] Y. Ge, H. Hu, J. Chen, K. Wang, and Z. He, "Hierarchical energy management of networked flexible traction substations for efficient RBE and PV energy utilization within ERs," *IEEE Transactions on Sustainable Energy,* vol. 14, no. 3, pp. 1397–1410, 2023.

[137] D. Feng, H. Zhu, X. Sun, and S. Lin, "Evaluation of power supply capability and quality for traction power supply system considering the access of distributed generations," *IET Renewable Power Generation,* vol. 14, no. 18, pp. 3644–3652, 2020.

[138] K. Deng *et al.*, "Deep reinforcement learning based energy management strategy of fuel cell hybrid railway vehicles considering fuel cell aging," *Energy Conversion and Management,* vol. 251, p. 115030, 2022.

[139] S. Su, T. Tang, J. Xun, F. Cao, and Y. Wang, "Design of running grades for energy-efficient train regulation: A case study for Beijing Yizhuang line," *IEEE Intelligent Transportation Systems Magazine,* vol. 13, no. 2, pp. 189–200, 2021.

[140] Y. Wang, S. Zhu, A. D'Ariano, J. Yin, J. Miao, and L. Meng, "Energy-efficient timetabling and rolling stock circulation planning based on automatic train operation levels for metro lines," *Transportation Research Part C: Emerging Technologies,* vol. 129, p. 103209, 2021.

[141] P. Wang and R. M. P. Goverde, "Multi-train trajectory optimization for energy efficiency and delay recovery on single-track railway lines," *Transportation Research Part B: Methodological,* vol. 105, pp. 340–361, 2017.

[142] A. Fernandez-Rodriguez, A. Fernández-Cardador, A. P. Cucala, M. Domínguez, and T. Gonsalves, "Design of robust and energy-efficient ATO speed profiles of metropolitan lines considering train load variations and delays," *IEEE Transactions on Intelligent Transportation Systems,* vol. 16, no. 4, pp. 2061–2071, 2015.

[143] M. Domínguez, A. Fernández-Cardador, A. P. Cucala, T. Gonsalves, and A. Fernández, "Multi objective particle swarm optimization algorithm for the design of efficient ATO speed profiles in metro lines," *Engineering Applications of Artificial Intelligence,* vol. 29, pp. 43–53, 2014.

[144] L. Zhang, D. He, Y. He, B. Liu, Y. Chen, and S. Shan, "Real-time energy saving optimization method for urban rail transit train timetable under delay condition," *Energy,* vol. 258, p. 124853, 2022.

[145] X. Chen, K. Li, L. Zhang, and Z. Tian, "Robust optimization of energy-saving train trajectories under passenger load uncertainty based on p-NSGA-II," *IEEE Transactions on Transportation Electrification,* vol. 9, no. 1, pp. 1826–1844, 2023.

[146] R. Teymourfar, B. Asaei, H. Iman-Eini, and R. Nejati fard, "Stationary super-capacitor energy storage system to save regenerative braking energy in a metro line," *Energy Conversion and Management,* vol. 56, pp. 206–214, 2012.

[147] G. Leoutsakos *et al.*, "Metro traction power measurements sizing a hybrid energy storage system utilizing trains regenerative braking," *Journal of Energy Storage,* vol. 57, p. 106115, 2023.

[148] D. Roch-Dupré, T. Gonsalves, A. P. Cucala, R. R. Pecharromán, Á. J. López-López, and A. Fernández-Cardador, "Determining the optimum installation of energy storage systems in railway electrical infrastructures by means of swarm and evolutionary optimization algorithms," *International Journal of Electrical Power & Energy Systems,* vol. 124, p. 106295, 2021.

[149] S. Wei *et al.*, "Optimisation of a catenary-free tramline equipped with stationary energy storage systems," *IEEE Transactions on Vehicular Technology,* vol. 69, no. 3, pp. 2449–2462, 2020.

[150] S. Wei, N. Murgovski, J. Jiang, X. Hu, W. Zhang, and C. Zhang, "Stochastic optimization of a stationary energy storage system for a catenary-free tramline," *Applied Energy,* vol. 280, p. 115711, 2020.

[151] X. Wang, P. Sun, Q. Wang, J. Ding, and X. Feng, "Joint optimization combining the capacity of subway on‐board energy storage device and timetable," *IET Intelligent Transport Systems,* vol. 17, no. 1, pp. 193–210, 2023.

[152] P. Liu, L. Yang, Z. Gao, Y. Huang, S. Li, and Y. Gao, "Energy-efficient train timetable optimization in the subway system with energy storage devices," *IEEE Transactions on Intelligent Transportation Systems,* vol. 19, no. 12, pp. 3947–3963, 2018.

[153] S. Yang, J. Xiong, D. Cao, and J. Wu, "A coordination optimization for train operation and energy infrastructure control in a metro system," *IEEE Transactions on Intelligent Transportation Systems,* vol. 25, no. 3, pp. 2656–2668, 2024.

[154] F. Zhu, Z. Yang, Z. Zhao, and F. Lin, "Two-stage synthetic optimization of supercapacitor-based energy storage systems, traction power parameters and train operation in urban rail transit," *IEEE Transactions on Vehicular Technology,* vol. 70, no. 9, pp. 8590–8605, 2021.

[155] P. M. Fernández, I. V. Sanchís, V. Yepes, and R. I. Franco, "A review of modelling and optimisation methods applied to railways energy consumption," *Journal of Cleaner Production,* vol. 222, pp. 153–162, 2019.

[156] S. Yang, Y. Chen, Z. Dong, and J. Wu, "A collaborative operation mode of energy storage system and train operation system in power supply network," *Energy,* vol. 276, p. 127617, 2023.

[157] D. Roch-Dupré, A. P. Cucala, R. R. Pecharromán, Á. J. López-López, and A. Fernández-Cardador, "Evaluation of the impact that the traffic model used in railway electrical simulation has on the assessment of the installation of a Reversible Substation," *International Journal of Electrical Power & Energy Systems,* vol. 102, pp. 201–210, 2018.

[158] A. Nash and D. Huerlimann, "Railroad simulation using OpenTrack," *WIT Transactions on The Built Environment,* vol. 74, 2004.

[159] R. Zhou and S. Song, "Optimal automatic train operation via deep reinforcement learning," in *2018 10th International Conference on Advanced Computational Intelligence (ICACI)*, Xiamen, China, 2018, pp. 103–108.

[160] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Reward function and initial values: better choices for accelerated goal-directed reinforcement learning," in *Artificial Neural Networks – ICANN 2006*, Berlin, Germany: Springer, Berlin Heidelberg, 2006, pp. 840–849.

[161] J. P. Powell and R. Palacín, "Passenger stability within moving railway vehicles: limits on maximum longitudinal acceleration," *Urban Rail Transit,* vol. 1, pp. 95–103, 2015.

[162] J. Yin, D. Chen, and Y. Li, "Smart train operation algorithms based on expert knowledge and ensemble CART for the electric locomotive," *Knowledge-Based Systems,* vol. 92, pp. 78–91, 2016.

[163] S. Su, X. Li, T. Tang, and Z. Gao, "A subway train timetable optimization approach based on energy-efficient operation strategy," *IEEE Transactions on Intelligent Transportation Systems,* vol. 14, no. 2, pp. 883–893, 2013.

[164] X. Yang, X. Li, B. Ning, and T. Tang, "An optimisation method for train scheduling with minimum energy consumption and travel time in metro rail systems," *Transportmetrica B: Transport Dynamics,* vol. 3, no. 2, pp. 79–98, 2015.

[165] J. Yuan, “Stochastic modelling of train delays and delay propagation in stations,” Ph.D. dissertation, Department of Transportation and Planning, Delft University of Technology, Delft, 2006.

[166] V. I. Herrera, H. Gaztañaga, A. Milo, A. Saez-de-Ibarra, I. Etxeberria-Otadui, and T. Nieva, "Optimal energy management and sizing of a battery--supercapacitor-based light rail vehicle with a multiobjective approach," *IEEE Transactions on Industry Applications,* vol. 52, no. 4, pp. 3367–3377, 2016.

[167] X. Zhang, H. Peng, H. Wang, and M. Ouyang, "Hybrid lithium iron phosphate battery and lithium titanate battery systems for electric buses," *IEEE Transactions on Vehicular Technology,* vol. 67, no. 2, pp. 956–965, 2018.

[168] G. Li, S. W. Or, and K. W. Chan, "Intelligent energy-efficient train trajectory optimization approach based on supervised reinforcement learning for urban rail transits," *IEEE Access,*

vol. 11, pp. 31508–31521, 2023.

[169] S. Su, T. Tang, and Y. Wang, "Evaluation of strategies to reducing traction energy consumption of metro systems using an optimal train control simulation model," *Energies,* vol. 9, no. 2, p. 105, 2016.

[170] Q. Zhang, H. Wang, Y. Zhang, and M. Chai, "An adaptive safety control approach for virtual coupling system with model parametric uncertainties," *Transportation Research Part C: Emerging Technologies,* vol. 154, p. 104235, 2023.

[171] G. Li and S. W. Or, "Multi-agent deep reinforcement learning-based multi-time scale energy management of urban rail traction networks with distributed photovoltaic–regenerative braking hybrid energy storage systems," *Journal of Cleaner Production,* vol. 466, p. 142842, 2024.

[172] M. Z. Kamh and R. Iravani, "Steady-state model and power-flow analysis of single-phase electronically coupled distributed energy resources," *IEEE Transactions on Power Delivery,* vol. 27, no. 1, pp. 131–139, 2012.

[173] B. Zhang, W. Hu, X. Xu, Z. Zhang, and Z. Chen, "Hybrid data-driven method for low-carbon economic energy management strategy in electricity-gas coupled energy systems based on transformer network and deep reinforcement learning," *Energy,* vol. 273, p. 127183, 2023.

[174] J. Hare, "Dealing with sparse rewards in reinforcement learning," *arXiv preprint arXiv:1910.09281,* 2019.

[175] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney, "Recurrent experience replay in distributed reinforcement learning," in *7th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[176] H. Dong, Y. Fu, Q. Jia, T. Zhang, and D. Meng, "Low carbon optimization of integrated energy microgrid based on life cycle analysis method and multi time scale energy storage," *Renewable Energy,* vol. 206, pp. 60–71, 2023.

[177] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (NSRDB)," *Renewable and sustainable energy reviews,* vol. 89, pp. 51–60, 2018.

[178] H. Wu, Y. Yuan, X. Zhang, A. Miao, and J. Zhu, "Robust comprehensive PV hosting capacity assessment model for active distribution networks with spatiotemporal correlation," *Applied Energy,* vol. 323, p. 119558, 2022.

[179] Z. Xin-gang and W. Zhen, "Technology, cost, economic performance of distributed photovoltaic industry in China," *Renewable and Sustainable Energy Reviews,* vol. 110, pp. 53-64, 2019.

[180] İ. Şengör, H. C. Kılıçkıran, H. Akdemir, B. Kekezoğlu, O. Erdinc, and J. P. Catalao, "Energy

management of a smart railway station considering regenerative braking and stochastic behaviour of ESS and PV generation," *IEEE Transactions on Sustainable Energy,* vol. 9, no. 3, pp. 1041–1050, 2017.

[181] S. Theocharides, G. Makrides, A. Livera, M. Theristis, P. Kaimakis, and G. E. Georghiou, "Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing," *Applied Energy,* vol. 268, p. 115023, 2020.

[182] J. Zhang, F. Chen, Z. Cui, Y. Guo, and Y. Zhu, "Deep learning architecture for short-term passenger flow forecasting in urban rail transit," *IEEE Transactions on Intelligent Transportation Systems,* vol. 22, no. 11, pp. 7004–7014, 2021.

[183] P. Arévalo, D. Benavides, J. Lata-García, and F. Jurado, "Energy control and size optimization of a hybrid system (photovoltaic-hidrokinetic) using various storage technologies," *Sustainable Cities and Society,* vol. 52, p. 101773, 2020.

[184] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence,* vol. 115, p. 105151, 2022.

[185] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Systems with Applications,* vol. 244, p. 122778, 2024.

[186] H. Yan, K. Yan, and G. Ji, "Optimization and prediction in the early design stage of office buildings using genetic and XGBoost algorithms," *Building and Environment,* vol. 218, p. 109081, 2022.

[187] S. Forrest, "Genetic algorithms," *ACM Computing Surveys (CSUR),* vol. 28, no. 1, pp. 77–80, 1996.

[188] Y. Tian *et al.*, "Non-dominated sorting artificial rabbit multi-objective sizing optimization for a conceptual powertrain of a 6× 4 battery electric tractor truck," *Energy,* p. 132009, 2024.

[189] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II," in *Parallel Problem Solving from Nature PPSN VI*, Berlin, Germany: Springer, Berlin Heidelberg, 2000, pp. 849–858.

[190] C. Zhang, K. Li, and J. Deng, "Real-time estimation of battery internal temperature based on a simplified thermoelectric model," *Journal of Power Sources,* vol. 302, pp. 146–154, 2016.

[191] K. Mongird *et al.*, "Energy storage technology and cost characterization report," Pacific Northwest National Laboratory (PNNL), Richland, WA, USA, 2019.

[192] X. Yang, A. Chen, X. Li, B. Ning, and T. Tang, "An energy-efficient scheduling approach to improve the utilization of regenerative energy for metro systems," *Transportation Research Part C: Emerging Technologies,* vol. 57, pp. 13–29, 2015.

[193] Z. Ren, Y. Chen, C. Song, M. Liu, A. Xu, and Q. Zhang, "Economic analysis of rooftop photovoltaics system under different shadowing conditions for 20 cities in China," in *Building Simulation*, vol. 17, no. 2, Beijing, China: Tsinghua University Press, 2024, pp. 235–252.

[194] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874,* 2017.

[195] Y.-K. Lin, P.-C. Chang, L. C.-L. Yeng, and S.-F. Huang, "Bi-objective optimization for a multistate job-shop production network using NSGA-II and TOPSIS," *Journal of Manufacturing Systems,* vol. 52, pp. 43–54, 2019.

[196] T. Wang, Y. Liang, X. Shen, X. Zheng, A. Mahmood, and Q. Z. Sheng, "Edge computing and sensor-cloud: Overview, solutions, and directions," *ACM Computing Surveys,* vol. 55, no. 13s, pp. 1–37, 2023.

[197] J. Zhang and D. Tao, "Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet of Things Journal,* vol. 8, no. 10, pp. 7789–7817, 2020.

[198] Y. I. Alzoubi and A. Mishra, "Green artificial intelligence initiatives: Potentials and challenges," *Journal of Cleaner Production,* vol. 468, p. 143090, 2024.