第 49 卷 第 6 期：2264-2275 高电压技术 Vol.49, No.6: 2264-2275
2023 年 6 月 30 日 High Voltage Engineering June 30, 2023
DOI: 10.13336/j.1003-6520.hve.20230099

基于深度强化学习算法的风光互补可再生能源制氢系统调度方案

梁 涛[1]，孙博峰[1]，谭建鑫[2]，曹 欣[2]，孙鹤旭[3]
（1. 河北工业大学人工智能与数据科学学院，天津 300130；2. 河北建投能源投资股份有限公司，石家庄 050051；3. 河北科技大学电气工程学院，石家庄 050018）

摘 要：风光可再生能源制备“绿氢”是实现能源低碳化的重要途径，但风能、太阳能的波动性、间歇性等问题会使系统存在“弃风、弃光”现象。为解决该问题，构建了可再生能源并网制氢系统，针对传统 CPLEX 需要精准预测数据、基于状态控制法的监控策略控制效果不够理想的缺点，将协调控制转化为序列决策问题，采用深度强化学习连续近端策略优化算法进行解决。在发电量、负荷等多种因素变化的情况下，设计了适合解决可再生能源制氢系统调度问题的深度强化学习模型(renewable energy to hydrogen-proximal policy optimization，R2H-PPO)，经过足够的训练后能够实现在线决策控制，并与日前控制方案和基于状态控制法的监控策略进行了对比，证明所采用方法避免了传统方案的不足，并能有效处理不同时刻、天气、季节的场景。结果证明了所提出的 R2H-PPO 方法的可行性和有效性。

关键词：可再生能源；制氢系统；深度强化学习；近端策略优化；运行优化；R2H-PPO

Scheduling Scheme of Wind-solar Complementary Renewable Energy Hydrogen Production System Based on Deep Reinforcement Learning

LIANG Tao[1], SUN Bofeng[1], TAN Jianxin[2], CAO Xin[2], SUN Hexu[3]
(1. School of Artificial Intelligence, Hebei University of Technology, Tianjin 300130, China;
2. Hebei Construction & Investment Group Co., Ltd., Shijiazhuang 050051, China;
3. School of Electrical Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China)

Abstract：The preparation of “green hydrogen” from wind-solar renewable energy is an important way to achieve low carbon energy, but the volatility and intermittency of wind and solar energy will cause the phenomenon of “wind and solar abandonment” in the system. In order to solve this problem, a grid-connected hydrogen production system of renewable energy is constructed in this paper. In view of the shortcomings that traditional CPLEX requires accurate data prediction and the control effect of monitoring strategy based on state control method is not ideal, the coordinated control is transformed into a sequential decision problem, which is solved by deep reinforcement learning continuous proximal policy optimization algorithm. Under the variation of various factors such as electricity generation and load, a deep strengthening learning model (R2H-PPO) suitable for solving the scheduling problem of hydrogen production system from renewable energy sources is designed, thereby, online decision controls can be realized after sufficient training; moreover, the online decision controls are compared with day-ahead control scheme and monitoring strategy based on state control method. It is proved that the adopted method avoids the shortcomings of traditional scheme, and can effectively deal with different time, weather and season scenarios. The results show that the proposed R2H-PPO method is feasible and effective.

Key words：renewable energy; hydrogen production system; deep reinforcement learning；proximal policy optimization；operation optimization; R2H-PPO

0 引言

人类社会的进步与发展离不开能源，可随着诸如煤、石油、天然气等传统能源大量的开采及使用，碳排放量逐年上升，全球各个国家的能源储备量也急剧下降。近年来的“双碳”政策，纯绿色、无碳、无污染且高效方便的氢气得到了学者和专家的广泛

基金资助项目：河北省科技支撑计划 (19210108D；19214501D；20314501D；F2021202022)。
Project supported by Hebei Science and Technology Support Plan (19210108D, 19214501D, 20314501D, F2021202022).

关注[1]。能源低碳化转型是目前能源改革的必由之路，而氢能在能源去碳化进程中扮演着越来越重要的地位。风能、光能与氢能结合，构成风光互补可再生能源制氢系统，是能源低碳化的典型应用场景[2-3]。

微电网合理的调度分配可以使整个可再生能源系统实现自我控制、保护以及能量管理，以达到提高能源利用率、增加系统总收益的目的。CPLEX求解器作为能源系统日前调度的求解方案得到了广泛的应用。文献[4]提出了基于两级碳交易机制和需求响应的多微网分布式协调优化模型，调用CPLEX求解器求解日前调度问题，以促进碳排放配额和电能共享，降低总运行成本和总碳排放量，同时分布式求解能保证微网间交易策略公平合理。文献[5]为优化综合能源系统的运行总成本，提出了一种兼顾低碳性与经济性的优化调度模型，将原问题转化为混合整数线性问题，运用CPLEX进行求解。文献[6]采用区间数学的方法对微网中可再生能源出力及电负荷的不确定性进行描述，在此基础上建立了采用区间多目标线性规划方法的优化调度模型。应用CPLEX优化软件在减少经济成本的同时实现较好的环境效益，并有效处理微网中的不确定因素。文献[7]建立综合能源系统双层调度模型，上层为电网和气网系统，下层为能量枢纽系统，以运行成本最小为目标函数进行低碳经济调度。利用CPLEX进行优化求解合理平衡系统的经济性与低碳性。但是上述基于CPLEX的能源调度问题的解决方案在很大程度上依赖于未来不确定性较强的预测信息。由于获取准确的预测信息是困难而且不现实的，因此调度策略在很大程度上受到预测或模型准确性的限制，难以适应实际环境的动态变化。

但在实际现场往往难以获得精准的预测数据。所以使用基于状态控制法的监控策略设计实际的能源调度方案，所得到的结果往往较为保守，难以取得理想的控制效果。一些学者针对这个问题也进行了相应的研究。文献[8]针对最优功率分配问题，设计了一个模糊控制能量管理单元，以在考虑发电和消耗的实时和长期预测数据的情况下选择适当的运行模式。文献[9]提出了太阳能储氢微电网的滞环控制，根据能源效率、氢消耗、运行时间、电池和金属氢化物储氢罐的荷电状态对5种运行模式进行了比较，提出的滞环控制策略可以减少燃料电池系统的开关次数和运行成本。文献[10]提出了一种用于电–氢混合储能系统孤岛直流微电网的模糊功率分配策略和控制方法。通过分析直流微网的运行模式，设计模糊逻辑控制器，给出了不同运行模式下不同接口变流器的控制方法，以提高蓄电池的使用寿命和氢气的安全性，并减少控制模式切换的次数。但这类方法仍需设置较为繁琐的运行模式规则，这类模糊控制规则往往不易获得。这也增加了计算与设计的难度。此外，一旦能源系统的结构、规模和容量发生变化，这类方法由于缺乏对变化的适应性，需要相应地重新设计解决方案，使得这种方法在实践中不易维护。

随着人工智能技术的发展，强化学习受到了更多研究者的关注，并开始应用于众多电力系统领域。与上述调度方法相比，强化学习提供了一种新的思路用于解决动态决策问题，它可以保证实时性要求，并且对许多不确定性的干扰具有很强的适应性。文献[11]构建了一个孤岛复合能源系统，将复合储能系统的协调控制转化为序列决策问题，并采用DQN(deep Q network)方法加以解决。文献[12]以微能源网为研究对象，建立了基于能源总线模型的微能源网系统，提出利用DQN算法对微能源网进行能量管理与优化策略研究。文献[13]研究了一个能源管理系统的运行策略，以最小化年度运行成本为目标，提出了基于DQN算法的创新能量管理方法，以找到运行成本最低的系统运行策略。但DQN算法也具有其缺点，它只能用于处理小规模离散动作空间问题，处理大规模问题易发生“维数爆炸”，所以DQN的短板较为明显，约束了整体性能。上述方法虽然在不同场景取得了一定的成效，且提供了解决问题的方法和思路，但也存在一定的局限性。

所以本研究对位于河北省北部的可再生能源制氢系统以系统收益最大化并降低环境成本为目标，针对上述研究方法的缺陷，提出了一种解决可再生能源制氢系统调度问题的深度强化学习模型(renewable energy to hydrogen-proximal policy optimization，R2H-PPO)的方法，本方法利用复杂度低、性能良好的近端策略优化算法对风光氢能源系统进行能量管理与优化策略，并对实验与设计的基于状态控制法的监控策略及日前优化最优值进行对比，验证了本文所提出方法的可行性。

1　风光互补可再生能源制氢系统模型

本文所研究系统如图1所示，主要包括风电场、光伏电站、蓄电池、水电解制氢装置、氢储罐等。

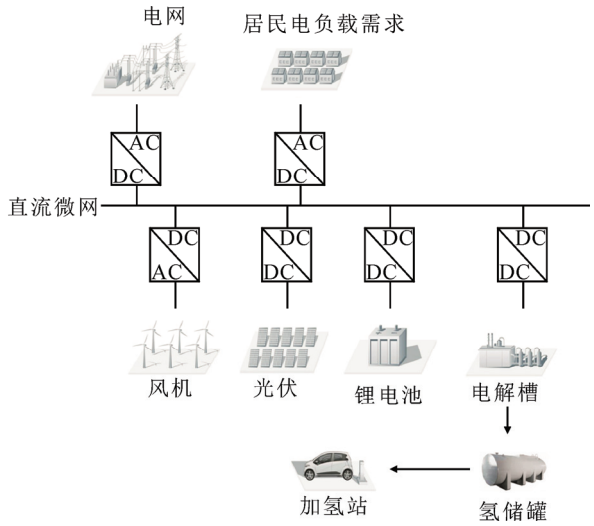


图 1 风光互补可再生能源制氢系统结构图

Fig.1 Wind-solar complementary renewable energy hydrogen production system

风电、光伏功率为整个系统提供可再生能源。系统需要满足波动的居民电负载需求，加氢站的氢能大巴提出波动的氢负载需求。风电、光伏及储氢罐内的氢气优先满足负载需求。当风电、光伏功率过剩时，可通过电解水制氢向储氢罐中存入氢气或向电网卖电；当风电、光伏功率降低时，可通过锂电池放电或从电网购电补充风电、光伏缺额功率，在每天结束时，储氢罐中氢气储量超过 50%的部分可以进行出售以换取经济收益。系统经济运行目标下，受制于运行成本、储氢容量及电价等因素影响。

1.1 能量供应模型

风力发电机组的基本原理是自然界的风流过风力发电机的风轮，使风轮转动，此过程将自然界的风能转化成了机械能，风轮经过增速后切割磁感线产生电流继而完成由机械能到电能的转化。风电场的功率输出与风速密切相关，其发电量[14]可表示为

$$P_{\mathrm{wt}}(v)=\begin{cases}P_{\mathrm{r}}\dfrac{v^2-v_{\mathrm{i}}^2}{v_{\mathrm{r}}^2-v_{\mathrm{i}}^2}, & v_{\mathrm{i}}<v<v_{\mathrm{r}}\\ P_{\mathrm{r}}, & v_{\mathrm{r}}\le v<v_{\mathrm{f}}\\ 0, & \text{其他}\end{cases} \tag{1}$$

式中：P_{wt} 为风电机组的额定功率；P_{r} 为风电机组的额定功率；v 为实时风速；v_{i} 为风机正常工作时的切入风速；v_{r} 为风机的额定风速；v_{f} 为风机正常工作的最大风速；t 为当前时刻。

光伏发电系统是依靠太阳能光伏电池进行发电的，通过吸收太阳光将这种能量储存到太阳能电池板中，其基本原理是依据光生伏特效应，通过电池板上的半导体材料吸收太阳能来转化成电能[15]。

光伏电池的发电输出功率为

$$P_{\mathrm{pv}}(t)=P_{\mathrm{STC}}\frac{L_{\mathrm{C}}(t)}{L_{\mathrm{STC}}}\left(1+\mu(T_{\mathrm{C}}-T_{\mathrm{STC}})\right) \tag{2}$$

式中：P_{pv} 表示太阳能光伏板实际发电功率；P_{STC} 表示太阳能光伏板额定功率；L_{C} 表示当前环境光照强度；L_{STC} 表示理想环境下的光照强度；μ 表示功率温度系数；T_{C} 表示工作时太阳能光伏板表面温度；T_{STC} 表示理想环境下太阳能光伏板表面温度。

1.2 电解水制氢模型

碱性电解槽以 KOH 水溶液作为电解液。在直流电的作用下，水被电解产生氢气和氧气。碱性电解槽于 20 世纪中期实现工业化，技术成熟，使用寿命可达 15 a 以上。碱性电解槽的结构特点是液体电解质和多孔隔板[16-17]。

碱性电解槽的工作电压模型表示为

$$U_{\mathrm{el}}=U_{\mathrm{rev}}+\frac{r_1+r_2T_{\mathrm{ele}}}{A}I_{\mathrm{el}}+s\ln\left(\frac{t_1+\dfrac{t_2}{T}+\dfrac{t_3}{T^2}}{A}I_{\mathrm{el}}+1\right) \tag{3}$$

式中：U_{el} 是碱性电解槽的工作电压；A 是电极面积；T_{ele} 是碱性电解槽的运行温度；I_{el} 是碱性电解槽的工作电流；r_1、r_2 是碱性电解槽的内阻；s、t_1、t_2、t_3 是电解槽中电极和电解液极化引起过电压的相关参数；U_{rev} 是可逆开路电压。

1.3 储能模型

锂电池具有能量密度大、安全性高、使用寿命长等优点，是风光互补可再生能源制氢系统中的重要组成单元。锂离子电池充放电时的荷电状态可以表示为

$$\begin{cases}B_{\mathrm{soc}}(t+1)=B_{\mathrm{soc}}(t)-\dfrac{P_{\mathrm{dis}}(t)}{Q_{\mathrm{r}}\eta_{\mathrm{dis}}}\\ B_{\mathrm{soc}}(t+1)=B_{\mathrm{soc}}(t)+\dfrac{P_{\mathrm{cha}}(t)\eta_{\mathrm{cha}}}{Q_{\mathrm{r}}}\end{cases} \tag{4}$$

式中：$B_{\mathrm{soc}}(t+1)$、$B_{\mathrm{soc}}(t)$ 分别表示锂离子电池在 $t+1$、t 时的荷电状态；$P_{\mathrm{cha}}(t)$、$P_{\mathrm{dis}}(t)$ 分别表示在单位时间内锂离子电池的充、放电功率；η_{dis} 表示锂离子电池的放电效率；η_{cha} 表示锂离子电池的充电效率；Q_{r} 是电池的额定容量。

2 基于深度学习的能源管理策略

2.1 PPO 算法原理

强化学习是指从环境状态到动作策略集的学习，以使得动作从环境中获得的累计奖励最大，学

习的过程如图 2 所示。本质上，智能体与环境的互动是一种马尔科夫决策过程(Markov decision process，MDP)。MDP 一般由一个四元组 (s,a,r,π) 定义[18]，其中：s 为所有环境状态的集合，$s_t \in s$ 表示智能体在 t 时段所处环境中感知到的状态；a 为智能体所有可执行动作的集合，a_t 表示智能体在 t 时刻采取的动作；r 为奖励函数，表示智能体在状态 s_t 执行动作 a_t 获得的即时奖励；π 为智能体的策略集，表示状态空间 s 到动作空间 a 的映射。

在强化学习方法中，与基于值的方法相比，基于策略的方法更适合于连续状态和动作空间问题。对于可再生能源制氢系统调度优化问题规模的增加带来的状态空间和行动空间的高维灾难，有必要使用基于策略的方法。作为一种基于策略的方法，近端策略优化算法(proximal policy optimization，PPO)算法[19]可以通过限制策略更新来降低参数设置的敏感性，从而获得良好的性能。其目的是将策略更新限制在一个小范围内，以确保更新后的策略不会偏离收集培训数据的旧策略太远。PPO 算法训练时，从经验回放库中抽取一个小批量经验样本供网络参数更新。Critic 网络通过时序差分误差 ς 的学习方法更新网络参数 ϕ。Actor 网络通过改进置信域策略优化算法(trust region policy optimization，TRPO)[20]的目标函数更新网络参数 θ。PPO 算法不仅解决了一般策略梯度方法数据样本利用率低和鲁棒性差的问题，而且通过对 TRPO 算法的目标函数进行转换以提高学习效率。此外，PPO 算法为进一步简化计算过程，加入裁剪机制得到新的目标函数，实现了一种与随机梯度下降兼容的置信域修正方法，并通过消除 KL 损失以简化算法及降低适应性修正的需求。

2.2　算法应用流程

PPO 算法在本文能源调度问题的应用流程分为离线训练和在线部署。

(1) 离线训练。在离线训练的过程中，首先对 Actor 和 Critic 的网络参数 θ 和 ϕ 进行初始化操作，获得训练集中随机一天的初始状态 s_1。然后在每一个时间步长内，智能体都会根据当前的策略 $\pi_\theta(a|s)$ 及当前时刻根据环境所获取的状态信息 s_t 选择合适的动作 a_t，通过执行动作 a_t 得到改变后的状态及奖励值 r_t，并将经验存储到缓冲区；接下来，从缓冲区随机抽取 L 个样本，将其放入 Actor 网络，并计算新旧策略的比值 ξ_l。当对 Critic 网络进行训练时，抽取样本作为输入值，计算差分误差 ς，并更新网络参数 ϕ，然后计算所得到的优势函数。最后通过裁剪得到目标函数 L^{clip}，通过更新网络参数 θ 获取新策略。接下来反复执行上述指令，直到迭代次数到达程序设置的上限。

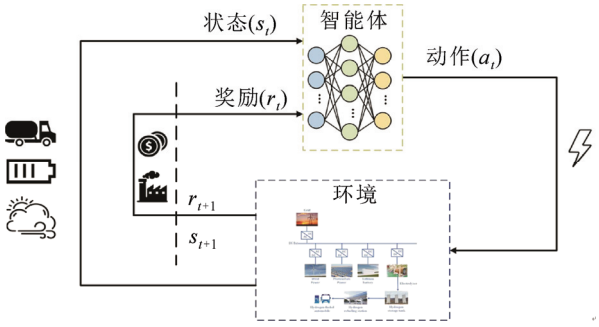


图 2　R2H-PPO 系统强化学习示意图

Fig.2　Schematic diagram of reinforcement learning in R2H-PPO system

(2) 在线部署。智能体进行训练后，需要将训练好所得到的策略在环境中执行，这时仅需要使用已经训练结束的 Actor 网络即可。首先获取 Actor 网络训练所最优参数 $\theta*$，然后获取实际需要测试的典型日初始状态，再在每一个时间步长根据所得到的策略 $\pi_{\theta*}$ 获取应当执行的动作 a_t，最后与环境交互直到迭代停止即可获得训练学习所得到的智能体最优调度方案。

2.3　约束条件

1）功率平衡约束

$$P_{\text{wt}}(t)+P_{\text{pv}}(t)+P_{\text{bat}}(t)+P_{\text{grid}}(t)=P_{\text{load}}(t)+P_{\text{ele}}(t) \tag{5}$$

式中：P_{wt} 代表风机出力；P_{pv} 代表光伏出力；P_{bat} 代表锂电池当前交换功率；P_{grid} 代表交流电网交换功率；P_{load} 代表波动电负载需求；P_{ele} 代表电解槽的工作功率。

2）锂电池单位时间交换功率约束

$$\begin{cases} 0<P_{\text{cha}}(t)<P_{\text{cha}}^{\max}, P_{\text{dis}}(t)=0 \\ 0<P_{\text{dis}}(t)<P_{\text{dis}}^{\max}, P_{\text{cha}}(t)=0 \end{cases} \tag{6}$$

式中：$P_{\text{cha}}(t)$ 和 $P_{\text{dis}}(t)$ 分别代表锂电池的实时充放电功率；$P_{\text{cha}}^{\max}$ 和 $P_{\text{dis}}^{\max}$ 分别代表锂电池实时充放电功率上限。

3）电解槽工作功率上下限

碱性电解槽作为一种特殊的用电设备，具有其特殊的工作性质。由于电解槽内部材料的特性，电解槽的运行功率不能低于某一限值，否则存在氢气氧气混和发生超过爆炸极限的风险，故需设置其安全工作功率的下限。

电解槽制氢量可以表示为

$$\begin{cases} P_{\text{elework}}^{\min} < P_{\text{ele}}(t) < P_{\text{elework}}^{\max}, P_{\text{H}}(t) = \eta_{\text{ele}} P_{\text{ele}}(t) \\ P_{\text{ele}}(t) = 0, P_{\text{H}} = 0 \end{cases} \tag{7}$$

式中：$P_{\text{ele}}(t)$ 是电解槽实时消耗功率；$P_{\text{elework}}^{\max}$ 和 $P_{\text{elework}}^{\min}$ 分别是电解槽工作功率的上、下限；P_{H} 是实时氢气产量；η_{ele} 是制氢效率，根据文献[21]取0.0192。

4）储氢罐 SOC 上下限

$$H_{\text{soc}}^{\min} \le H_{\text{soc}}(t) \le H_{\text{soc}}^{\max} \tag{8}$$

式中：$H_{\text{soc}}(t)$ 是储氢罐实时储氢状态；$H_{\text{soc}}^{\max}$ 是储氢罐储氢上限；$H_{\text{soc}}^{\min}$ 是储氢罐储氢下限。

5）锂电池 SOC 上下限

$$B_{\text{soc}}^{\min} \le B_{\text{soc}}(t) \le B_{\text{soc}}^{\max} \tag{9}$$

式中：$B_{\text{soc}}(t)$ 是锂电池实时状态；$B_{\text{soc}}^{\max}$ 是锂电池储电上限；$B_{\text{soc}}^{\min}$ 是锂电池储电下限。

6）SOC 复位约束

$$D_{\text{SOC}} = \min\left|\sum_{t=1}^{T} P_{\text{bat}}(t)\right| \tag{10}$$

式中：D_{SOC} 代表后文中所使用到的储能设备复位惩罚；T 是系统调度周期内的最大周期数。为使 R2H 系统储能元件在一天结束时恢复初始储能状态，需要设置储能元件复位约束，即最小化单个运行周期的锂电池出力之和。

7）SOCH 复位约束

为使氢储罐在一天结束时恢复初始储能状态，需要设置储能元件复位约束，即

$$D_{\text{SOCH}} = \min\left|\sum_{t=1}^{T} (H_{\text{in}}(t) - H_{\text{out}}(t))\right| \tag{11}$$

式中：D_{SOCH} 代表后文中所使用到的储能设备复位惩罚；$H_{\text{in}}(t)$、$H_{\text{out}}(t)$ 分别代表储氢罐实时流入、流出氢气量。

8）环境成本

环境成本主要来源于风光氢系统从交流电网购电，由于中国主要电力来源为火力发电，燃煤会产生大量的二氧化碳造成温室效应，所以系统需要最小化系统全天环境成本[22]，即

$$D_{\text{C}} = \min\sum_{t=1}^{T} C_{\text{C}}(t) \tag{12}$$

式中：$C_{\text{C}}(t)$ 代表系统购电导致的二氧化碳排放量；D_{C} 表示后文中提及的环境成本惩罚。

2.4 状态空间及动作空间

对于微能源网模型，环境提供给智能体的信息一般为可再生能源出力、波动电负载、氢能大巴负载、分时电价、负荷和电储能的荷电状态。因此微能源网模型的状态空间定义为

$$\left[P_{\text{wt}}, P_{\text{pv}}, P_{\text{load}}, H_{\text{load}}, C_{\text{buy}}, C_{\text{sell}}, B_{\text{soc}}, H_{\text{soc}}\right] \tag{13}$$

式中：H_{load} 表示氢负载需求；C_{buy} 和 C_{sell} 分别代表当前时刻购售电价。

智能体观测到环境的状态信息后，根据自身策略集 π 在动作空间 a 中选择一个动作。微能源网模型中的设备较为复杂，可以分为能量转换设备和储能设备。可再生能源制氢系统中能量转换设备较多，每台设备的出力共同构成了调度问题的动作空间，即

$$\left[P_{\text{bat}}, P_{\text{ele}}, A_{\text{SOCH}}\right] \tag{14}$$

式中 A_{SOCH} 表示储氢罐动作。

若动作空间 a 较复杂，则会增加强化学习算法的学习难度和学习时间，所以需要对其进行分析来简化变量以减轻计算负荷。由于储氢罐只与电解槽和外部氢负载相连，电解槽制备的氢气输入储氢罐，储氢罐中储存的氢气用于满足波动氢负载需求，所以储氢罐的动作空间无需通过算法决策，进而简化后的动作空间 a 可以表示为

$$\left[P_{\text{bat}}, P_{\text{ele}}\right] \tag{15}$$

2.5 优化目标及奖励函数

在本文所提出系统满足电、氢负载的前提下，通过将电能售卖给交流电网可以获取经济收益，再计算与各组件的运行折旧成本总和的差值。在一天结束时，如果储氢罐中有超过初始状态的氢气，将储氢罐复位至初始状态，并售卖超过初始状态的氢气计算总收益，使收益最大化，同时减少购电量减少环境成本。系统运行收益如下：

$$\max f(t) = \sum_{t=1}^{T}\begin{bmatrix} I_{\text{grid}}(t) + I_{\text{hydrogen}}(t) \\ -C_{\text{equipments}}(t) \end{bmatrix} \tag{16}$$

$$\begin{cases} I_{\text{grid}} = \sum_{t=1}^{T} (C_{\text{sell}}(t) P_{\text{sell}}(t) - C_{\text{buy}}(t) P_{\text{buy}}(t)) \\ I_{\text{hydrogen}} = C_{\text{hydrogen}} H_{\text{sell}} \\ C_{\text{equipments}} = \sum_{t=1}^{T} C_{\text{battery}} P_{\text{bat}}(t) + C_{\text{ele}} P_{\text{ele}}(t) \end{cases} \tag{17}$$

系统运行收益计算如式(16)所示，主要与系统运行过程中设备维护成本、电网交换功率收益支出及氢气售卖收益相关。其中 I_{grid} 是电网收益/支出；I_{hydrogen} 是氢气售卖收益；$C_{\text{equipments}}$ 是设备维护费用；

$P_{\text{sell}}(t)$ 和 $P_{\text{buy}}(t)$ 分别是系统向交流电网售、购电功率；C_{hydrogen} 是氢气售价；H_{sell} 是氢气周期结算销售量；C_{battery} 和 C_{ele} 分别是锂电池和电解槽的维护成本。

因此，强化学习中的奖励函数可以设置为系统收益和各组件惩罚及环境成本惩罚的总和，表示为

$$R = R_{\text{sys}} + R_{\text{Hsell}} - D_{\text{SOC}} - D_{\text{SOCH}} - D_{\text{bat}} - D_{\text{H}} - D_{\text{grid}} - D_{\text{C}} \tag{18}$$

式中：R_{sys} 表示系统收益奖励；R_{Hsell} 表示氢气售卖收益奖励；D_{bat}、D_{H} 分别表示选择动作时储能设备超限及缺氢弃氢惩罚；D_{grid} 表示系统弃电缺电惩罚。

3　实验

为验证上文所提出的方法的有效性，本文从中国河北崇礼某地春夏秋冬 4 季各选取一典型日，以图 1 所示风光互补可再生能源制氢系统为算例模型。系统中供电单元为风力发电机组、光伏发电模组及交流电网；储能元件是锂电池；系统使用碱性电解槽电解氢氧化钾溶液制备氢气。本文研究建立在山区风光资源充足的条件，在保证供电可靠性的前提之下，充分利用可再生能源，通过合理的优化调度，提高整个风光互补可再生能源制氢系统的运行收益，同时降低系统从交流电网购电、减少碳排放以响应能源低碳化的号召，最后给出各个子系统设备的出力计划。

为针对该系统解决上述优化调度问题，将全天能源优化问题以小时分辨率划分为 24 个子问题。在能源价格方面，本研究使用分时电价，设置每日 23:00—次日 07:00 为谷段，每日 08:00—11:00 和 15:00—18:00 为平段，每日 12:00—14:00 及 19:00—22:00 为峰段，其中谷段电价为 0.4 元/kWh，上网电价为 0.2 元/kWh；平段电价为 0.75 元/kWh，上网电价为 0.35 元/kWh；峰段电价为 1.2 元/kWh，上网电价为 0.5 元/kWh。碱性电解槽运行成本取 0.022 元/kWh，锂电池运行成本取 0.005 元/kWh，售氢价格为 35 元/kg。图 3、4 是本文所使用典型日的风光资源。

3.1　基于状态控制法的监控策略

基于状态控制法的监控系统流程图如附录 A 图 A1 所示，通过净功率 P_{net} 作为标准决定实际现场功率分配调度情况，在实际现场的状态控制法中，需要优先保证系统的安全平稳运行，经济效益考虑情况不足。

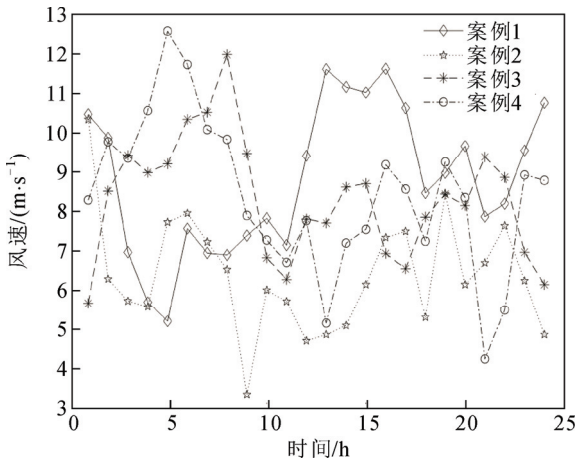


图 3　风速

Fig.3　Wind speed

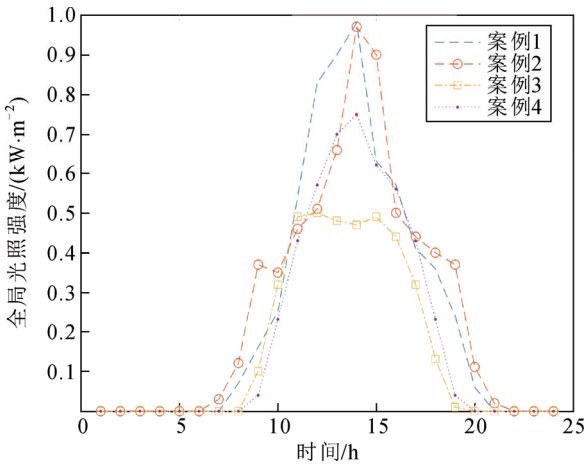


图 4　光照强度

Fig.4　Light intensity

1）风光功率不足情况：判断储氢罐及锂电池储能状态，当储氢罐未处在最大储氢状态时，按锂电池、交流电网购电的顺序，补足风光功率缺额，并使电解槽工作在当前时刻最大工作功率。

2）风光功率过剩情况：通过判断储氢罐及锂电池储能状态，优先给电解槽供电制氢以保证满足氢负载需求。在电解槽制氢后如果仍有剩余，再给蓄电池充电以做储备能源。最后将多余电量出售给交流电网以获取经济收益。

3.2　基于 CPLEX 的日前调度方法

CPLEX 求解器作为日前调度的典型求解方法，由于可以对单目标问题求得最优解，得到了广泛的应用。选取典型日风光资源为例，设置系统收益最大化作为目标函数，可以求得最优解。但由于日前调度方法需要精准的预测数据，风光资源数据具有

较强的不确定性，预测结果往往存在偏差，因此该方法无法得到准确的实时调度方案，不满足调度的实时性。因此在本文中，将不同典型场景下的CPLEX 求解混合整数规划问题，得到最优解，并将PPO-Continuous 的方案与对应样本的最优方案进行比较，以对比本文所提出方法的实验效果。

3.3 强化学习实验

3.3.1 训练过程

为了证明所研究方法的收敛性能，图 5 显示了训练过程中的奖励曲线。可以观察到，由于智能体在开始时不熟悉环境，执行调度决策后收到的奖励很小。随着训练过程的持续，智能体不断与环境交互，积累经验，奖励的总体趋势逐渐增加，最终趋于稳定。由于训练数据，如风光资源、电氢负荷曲线在每次迭代中都在变化，因此奖励在训练过程中会振荡。为了展示智能体在训练过程中对调度策略的进步，分别选取 500、2000、5000 代这几个点的奖励及具有代表性的夏季典型日调度图进行展示，如图 6 所示，其中储能曲线中，SOC 代表锂电池储能状态，SOCH 代表储氢罐储能状态。

由于选取的夏季典型日风光资源较为紧缺且负载需求也相应增大，可以发现，系统很难在这一天满足系统电氢需求的同时减少环境成本，通过展示不同代的调度方案可以发现，R2H-PPO 智能体随着迭代次数的增加，取得了如下效果：首先，智能体在刚刚开始学习到 500 代时，系统会在几个时刻大幅购电以满足电氢负载需求，而且在一天结束时，锂电池的 SOC 降至最小值，与当天初始 SOC 有一定的差值，不利于系统第 2 天的调度，储氢罐的氢气储能降到了最小值，并在最后几个时刻出现了缺氢状况，未满足氢负载需求，这显然不符合系统调度的原则。随着迭代次数的增加，到了 2000 代时，系统通过多次小幅购电，满足了电氢负载需求，并未出现缺氢的状况，一天结束后，锂电池和储氢罐的储能也在最小值之上。当学习到 5000 代时，系统通过对各组件出力的合理调度分配，仅从电网小幅购电一次，满足了电氢负载的同时，在一天结束后，锂电池 SOC 状态与当天初始值接近，有利于第 2 天的调度分配，储氢罐的氢气含量也比 2000 代时有了提高。这足以证明智能体在不断的迭代学习过程中取得了不错的效果。

3.3.2 对比实验

为针对之前学者所做研究的不足进行改进，首先使用广泛应用的强化学习算法 DQN 设置相同的环境与连续空间 PPO 即 PPO-Continuous 算法的奖励曲线进行比较，如图 7 所示。可以发现，作为基于值的 DQN 算法，由于设置的动作空间均为离散动作，进而导致系统收敛效果变差，震荡较大，最终难以得到满意的调度方案及奖励曲线。而相比于针对确定性策略、动作是离散型随机变量的基于值的强化学习方法，基于策略的强化学习方法 PPO-Continuous 在解决可再生能源制氢问题上能够取得更好的效果。从迭代图中可以看出，PPO-Continuous 算法在最终取得稳定的奖励后，不光能取得更高的奖励值，且震荡幅度更小。这也证明本文所提出的将 PPO-Continuous 应用于并网可再生能源制氢系统调度的 R2H-PPO 方法，更适用于本文所提出问题的调度优化。

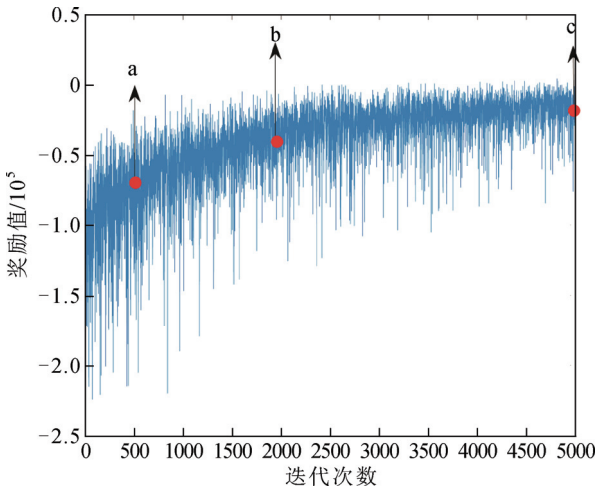


图 5 R2H-PPO 系统训练奖励迭代曲线

Fig.5 Training reward iteration curve of R2H-PPO system

下文将对 4 个案例分别使用基于状态控制的方法、CPLEX 日前调度及 R2H-PPO 方法进行实验，4 组实验 SOC 及 SOCH 状态展示如图 8 所示，调度对比图如图 9 所示。对日前调度方法进行观察，可以发现，由于提前得到了风光预测数据，站在全局视角的日前调度方案，可以在电价低时进行电能的提前储备，在电价高时进行售卖并避免电解水制氢设备的无用消耗。这种经济性调度方案是接近完美的，而正如前文所述，由于风光资源的不确定性，获取精准的风光预测数据较为困难，所以这种方法虽然能够取得不错的调度结果，但往往难以应用于工业现场。

本文以日前调度方法取得的调度方案作为参照，可以得到传统的监控策略方法和强化学习方法

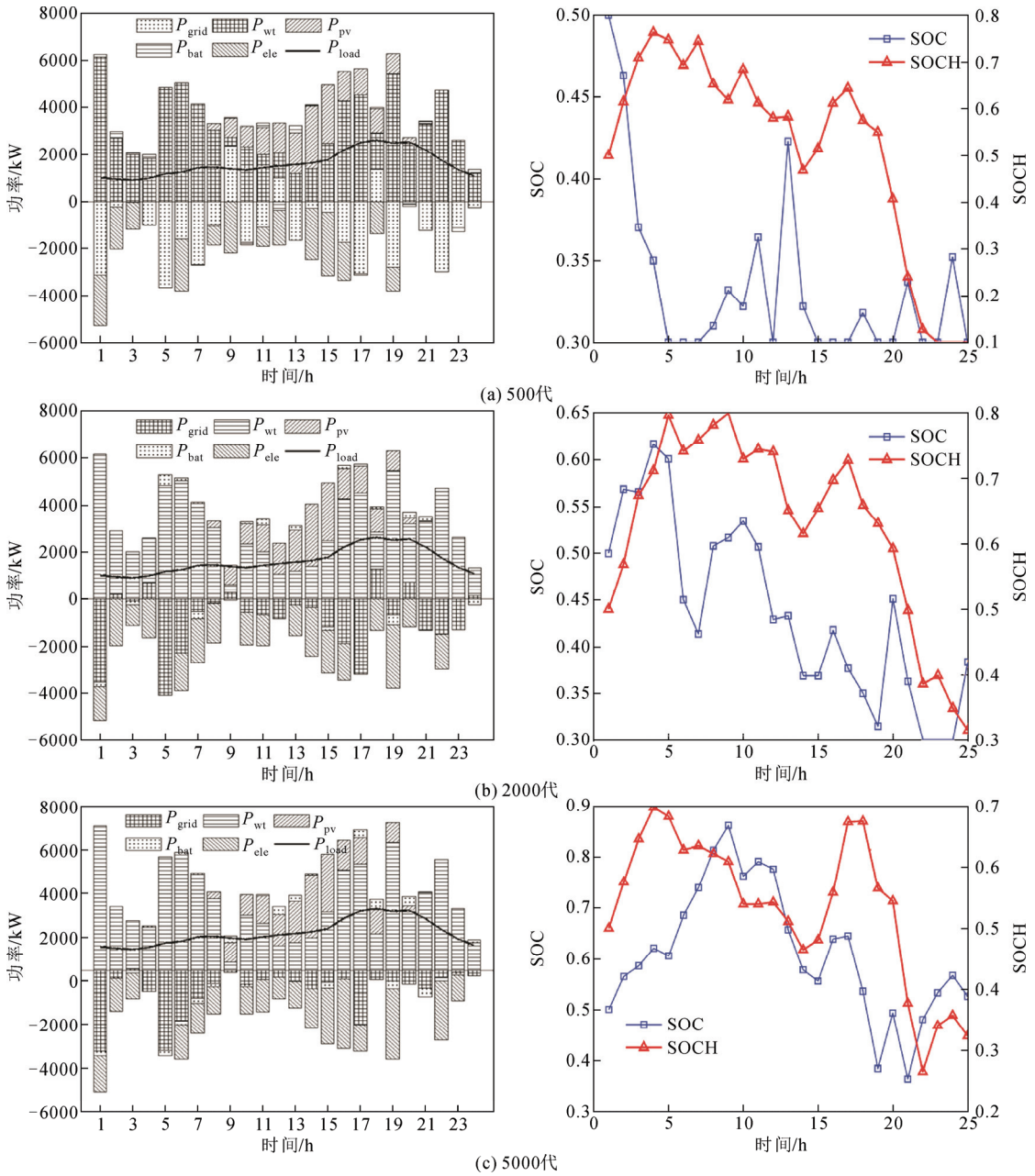


图 6　智能体迭代训练效果展示

Fig.6　Display of agent iterative training effect

的结果如图 8、图 9 及表 1 所示，可以得到如下结论：

1）作为基准的 CPLEX 求解器得出的日前调度方案可以归结为：在白天可再生能源充足时将能源存储起来供晚上实时电价比较高的用电高峰时刻利用，对储能设备的合理调度分配以达到节约购电成本目的的同时，通过合理的购售电以达到最佳经济收益。

2）基于监控策略的方案，系统为满足电氢负载的需求，避免出现缺电缺氢的情况，尽可能在任意时刻都保证锂电池和储氢罐有较高的储量，分配给电解槽大量的功率，减少了向电网卖电的功率，增加了设备的运行成本，减少了卖电收益。但这种

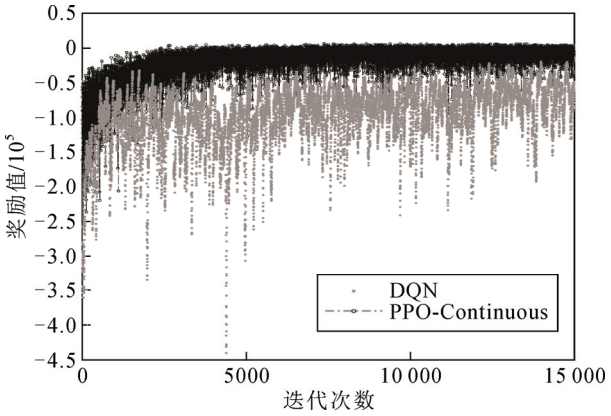
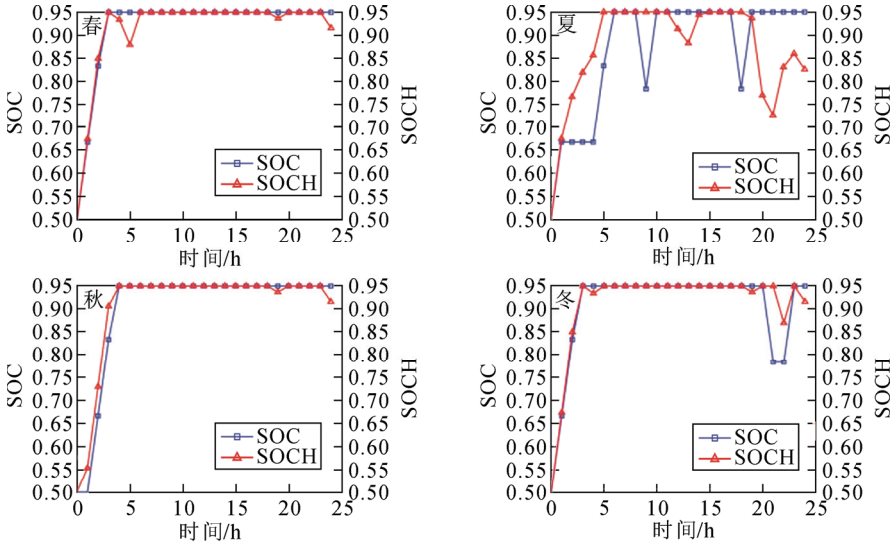
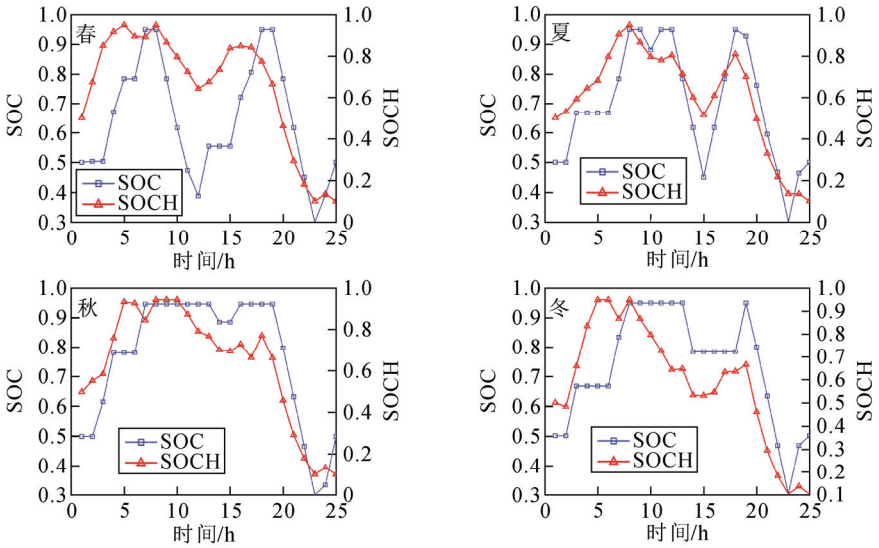


图 7　R2H-PPO 与 DQN 算法迭代曲线对比图

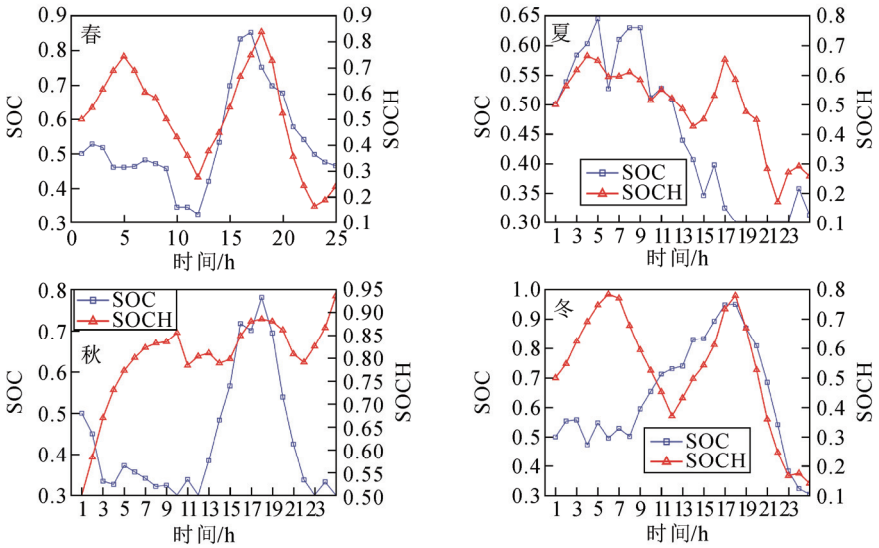
Fig.7　Comparison of iteration curves between R2H-PPO and DQN algorithms



(a) 监控策略



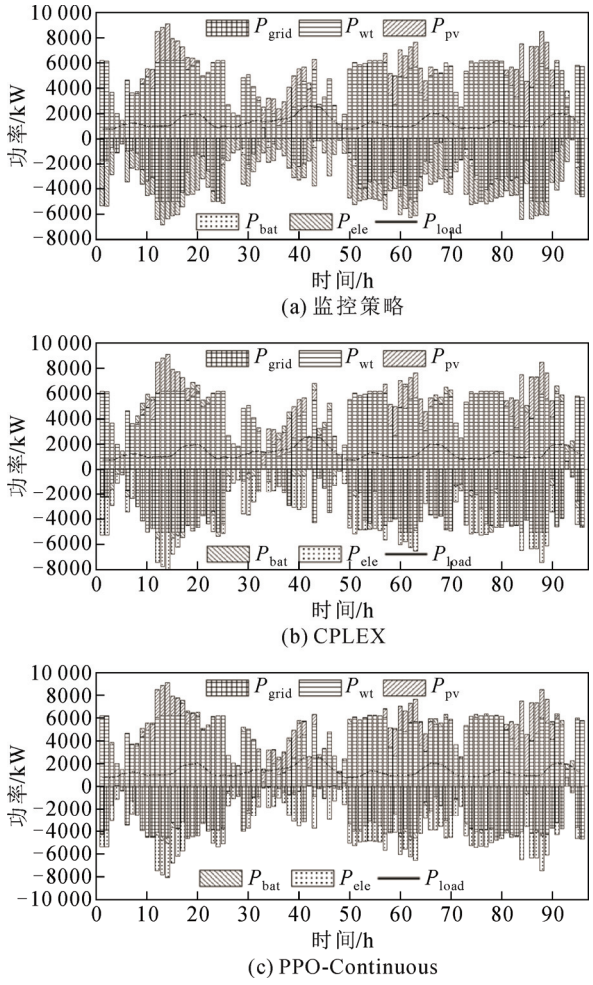
(b) CPLEX



(c) PPO-Continuous

图 8 R2H-PPO 算法典型日储能曲线对比

Fig.8 Comparison of typical daily energy storage curves of R2H-PPO algorithm



(a) 监控策略

(b) CPLEX

(c) PPO-Continuous

图 9　RH2-PPO 系统典型日调度曲线图

Fig.9　Typical daily scheduling curve of RH2-PPO system

表 1　R2H-PPO 系统典型案例训练结果

Table 1　Training results of R2H-PPO system typical cases

典型案例	系统收益/元		
	CPLEX	监控策略	PPO-Continuous
案例 1	31 718.9	27 269.1	29 407.5
案例 2	9523.7	4885.4	5914.8
案例 3	29 676.9	26 426.7	27 231.9
案例 4	27 777.1	23 168.6	25 572.9

方法较为保守，只能保证系统的稳定运行，无法满足系统总体收益较高的目标。

3）结合表 1 中数据、调度曲线及储能状态曲线可以发现，R2H-PPO 方法得到的系统调度策略可以充分利用电氢储能，在满足电氢负载的前提下，在电价较低的时段减小售电量，并增加制氢功率存储在储氢罐中；在电网用电高峰期，电价升高，此时增加电网售电量，以获得较高的售电收益，并减少此时分配给电解槽的功率，以降低运行成本，从而提高系统总体收益。同时通过对环境的学习，合理售卖氢气及电网售电，使 SOC 和 SOCH 在合理的范围内变化，而不是一直处于最高限。

综上，在迭代相同的次数后，本文提出的 R2H-PPO 算法相较于离散动作空间的 DQN 算法，取得的奖励值更高、更收敛。通过对春夏秋冬 4 个典型日案例的分析可知，R2H-PPO 方法在每个案例中取得的收益比监控策略的收益要高，更能满足系统的目标。R2H-PPO 算法的控制效果也更加接近于 CPLEX 日前调度方法的控制效果，这也证明了本文所提出方法的合理性。此外，由于 R2H-PPO 算法通过离线学习的方式，相较于 CPLEX 日前调度方法减少了对实时风光资源预测的难度及需求，这也就使得本文所提出方法更适合应用于实际现场。

4　结论

本文提出并网可再生能源制氢系统 R2H-PPO 智能调度方法，以经济性为主目标，完成了微能源网的能量管理与优化。在此基础上，本文通过不同复杂场景的仿真实验完成了迭代过程和优化结果的比较，验证了算法的可行性和有效性。本文做出的贡献及主要贡献如下：

1）本文建立了并网可再生能源制氢系统经济性调度模型，将深度强化学习应用于可再生能源制氢系统中，以达到节能减排、实现双碳愿景的目的。

2）针对日前调度方案需要获取精确预测数据的难点，以及传统状态监控方法智能性不足和基于值的算法的缺点，将动态能量转换问题表述为马尔可夫决策过程，提出了一种 R2H-PPO 的动态能量分配策略。

3）设计了一个基于状态控制法的监控策略，并将其与日前 CPLEX 调度方案和 R2H-PPO 方法进行了对比分析，以证明所提出的能源管理策略的有效性。

由此证明，强化学习算法可以通过人工智能手段的学习，在不需要风光日前预测数据的前提下进行调度策略且取得了较好的实验结果，对复杂的能源能量管理与优化方法提供了有益参考，践行了双碳政策。

附录见本刊网络版(http://hve.epri.sgcc.com.cn)。

参考文献　References

[1] 俞红梅，邵志刚，侯　明，等. 电解水制氢技术研究进展与发展建议[J]. 中国工程科学，2021，23(2)：146-152.
YU Hongmei, SHAO Zhigang, HOU Ming, et al. Hydrogen produc-

tion by water electrolysis: progress and suggestions[J]. Strategic Study of CAE, 2021, 23(2): 146-152.

[2] 袁铁江，曹继雷. 计及风电−负荷不确定性的风氢低碳能源系统容量优化配置[J]. 高电压技术，2022，48(6)：2037-2044.
YUAN Tiejiang, CAO Jilei. Capacity optimization allocation of wind hydrogen low-carbon energy system considering wind power-load uncertainty[J]. High Voltage Engineering, 2022, 48(6): 2037-2044.

[3] 韩 莹，于三川，李荦一，等. 计及阶梯式碳交易的风光氢储微电网低碳经济配置方法[J]. 高电压技术，2022，48(7)：2523-2533.
HAN Ying, YU Sanchuan, LI Luoyi, et al. Low-carbon and economic configuration method for solar hydrogen storage microgrid including stepped carbon trading[J]. High Voltage Engineering, 2022, 48(7): 2523-2533.

[4] 乔学博，杨志祥，李 勇，等. 计及两级碳交易和需求响应的多微网合作运行优化策略[J]. 高电压技术，2022，48(7)：2573-2583.
QIAO Xuebo, YANG Zhixiang, LI Yong, et al. Optimization strategy for cooperative operation of multi-microgrids considering two-level carbon trading and demand response[J]. High Voltage Engineering, 2022, 48(7): 2573-2583.

[5] 陈锦鹏，胡志坚，陈嘉滨，等. 考虑阶梯式碳交易与供需灵活双响应的综合能源系统优化调度[J]. 高电压技术，2021，47(9)：3094-3104.
CHEN Jinpeng, HU Zhijian, CHEN Jiabin, et al. Optimal dispatch of integrated energy system considering ladder-type carbon trading and flexible double response of supply and demand[J]. High Voltage Engineering, 2021, 47(9): 3094-3104.

[6] 朱晓荣，谢婉莹，鹿国微. 采用区间多目标线性规划法的热电联供型微网日前调度[J]. 高电压技术，2021，47(8)：2668-2677.
ZHU Xiaorong, XIE Wanying, LU Guowei. Day-ahead scheduling of combined heating and power microgrid with the interval multi-objective linear programming[J]. High Voltage Engineering, 2021, 47(8): 2668-2677.

[7] 刘哲远，邢海军，程浩忠，等. 考虑碳排放流及需求响应的综合能源系统双层优化调度[J]. 高电压技术，2023，49(1)：169-178.
LIU Zheyuan, XING Haijun, CHENG Haozhong, et al. Bi-level optimal scheduling of integrated energy system considering carbon emission flow and demand response[J]. High Voltage Engineering, 2023, 49(1): 169-178.

[8] JAFARI M, MALEKJAMSHIDI Z, LU D D C, et al. Development of a fuzzy-logic-based energy management system for a multiport multi-operation mode residential smart microgrid[J]. IEEE Transactions on Power Electronics, 2019, 34(4): 3283-3301.

[9] KUMAR K, ALAM M, VERMA S, et al. Effect of hysteresis band control strategy on energy efficiency and durability of solar-hydrogen storage based microgrid in partial cloudy condition[J]. Journal of Energy Storage, 2020, 32: 101936.

[10] ZHANG X, PEI W, MEI C X, et al. Transform from gasoline stations to electric-hydrogen hybrid refueling stations: an islanding DC microgrid with electric-hydrogen hybrid energy storage system and its control strategy[J]. International Journal of Electrical Power & Energy Systems, 2022, 136: 107684.

[11] 张自东，邱才明，张东霞，等. 基于深度强化学习的微电网复合储能协调控制方法[J]. 电网技术，2019，43(6)：1914-1921.
ZHANG Zidong, QIU Caiming, ZHANG Dongxia, et al. A coordinated control method for hybrid energy storage system in microgrid based on deep reinforcement learning[J]. Power System Technology, 2019, 43(6): 1914-1921.

[12] 刘俊峰，陈剑龙，王晓生，等. 基于深度强化学习的微能源网能量管理与优化策略研究[J]. 电网技术，2020，44(10)：3794-3803.
LIU Junfeng, CHEN Jianlong, WANG Xiaosheng, et al. Energy management and optimization of multi-energy grid based on deep reinforcement learning[J]. Power System Technology, 2020, 44(10): 3794-3803.

[13] SANAYE S, SARRAFI A. A novel energy management method based on deep Q network algorithm for low operating cost of an integrated hybrid system[J]. Energy Reports, 2021, 7: 2647-2663.

[14] LI N, ZHAO X W, SHI X P, et al. Integrated energy systems with CCHP and hydrogen supply: a new outlet for curtailed wind power[J]. Applied Energy, 2021, 303: 117619.

[15] ZAHEDI A. Maximizing solar PV energy penetration using energy storage technology[J]. Renewable and Sustainable Energy Reviews, 2011, 15(1): 866-870.

[16] HAUG P, KREITZ B, KOJ M, et al. Process modelling of an alkaline water electrolyzer[J]. International Journal of Hydrogen Energy, 2017, 42(24): 15689-15707.

[17] FANG R M, LIANG Y. Control strategy of electrolyzer in a wind-hydrogen system considering the constraints of switching times[J]. International Journal of Hydrogen Energy, 2019, 44(46): 25104-25111.

[18] SIGAUD O, BUFFET O. Markov decision processes in artificial intelligence[M]. Hoboken, USA: John Wiley & Sons, Inc., 2013.

[19] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL].[2021-07-05]. https://arxiv. org/pdf/1707.06347. pdf.

[20] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR. org, 2015: 1889-1897.

[21] 马腾飞，裴 玮，肖 浩，等. 基于纳什谈判理论的风−光−氢多主体能源系统合作运行方法[J]. 中国电机工程学报，2021，41(1)：25-39.
MA Tengfei, PEI Wei, XIAO Hao, et al. Cooperative operation method for wind-solar-hydrogen multi-agent energy system based on Nash bargaining theory[J]. Proceedings of the CSEE, 2021, 41(1): 25-39.

[22] FANG R M. Multi-objective optimized operation of integrated energy system with hydrogen storage[J]. International Journal of Hydrogen Energy, 2019, 44(56): 29409-29417.



梁 涛(通信作者)
1975—，男，博士，教授，硕导
主要从事可再生能源制氢、新能源大数据挖掘的研究
E-mail: 54008214@qq.com

LIANG Tao
Ph.D., Professor
Corresponding author

收稿日期 2023-01-20 修回日期 2023-04-13 编辑 程子丰

附录 A

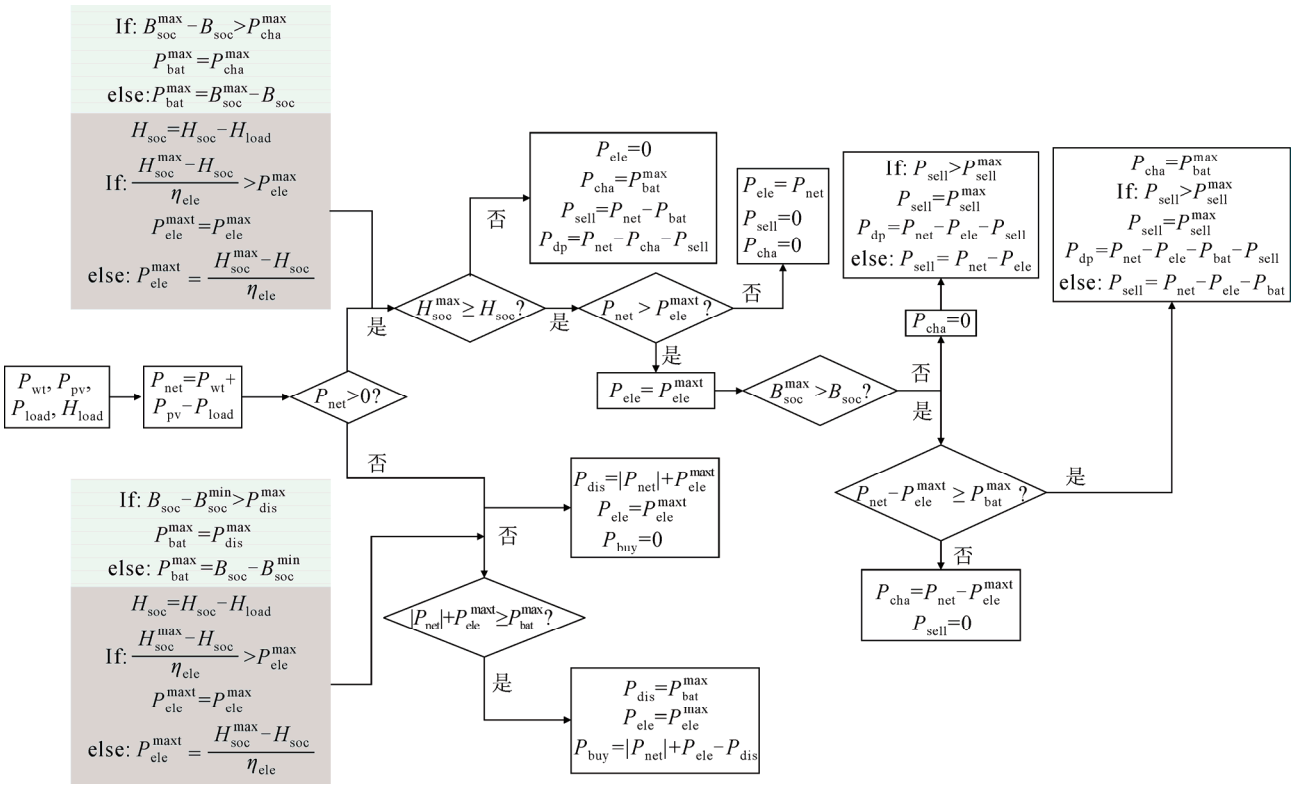


图 A1　基于状态控制法的监控策略

Fig.A1　Monitoring strategy based on state control method