Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI4Tech: AI Enabling Critical Technologies

MORL-CA: Dynamic Multi-Objective Reinforcement Learning for Chlor-Alkali Process Optimization Under Time-Varying Conditions

Derun Gan[1], **Renhao Yin**[1], **Guangzhi Qu**[2] and **Feng Zhang**[1]
[1]China University of Geosciences, Wuhan, China
[2]Oakland University, Rochester, MI, USA
{ganderun, yinrh}@cug.edu.cn, gqu@oakland.edu, jeff.f.zhang@gmail.com

Abstract

Chlor-alkali production is a large-scale industrial process whose operating conditions and equipment states evolve over time. Its process optimization requires ongoing trade-offs among conflicting objectives such as product yield, energy consumption, and equipment life. Existing optimization approaches are typically static and must be re-optimized after environmental changes, limiting their real-world applicability. In this work, we model the problem as a dynamic multi-objective sequential decision-making problem that continuously tracks a time-varying Pareto set under changing conditions. We propose MORL-CA, a multi-objective reinforcement learning framework that integrates offline pretraining on historical data with constrained online policy refinement. MORL-CA introduces a state-aware adaptive objective weighting mechanism within a multi-critic actor-critic architecture, enabling localized Pareto-improving policy updates while satisfying operational and safety constraints. Extensive experiments in an environment conducted from real chlor-alkali data demonstrate that MORL-CA achieves superior Pareto solution quality and smoother adaptation to dynamics compared with state-of-the-art multi-objective optimizers and MORL baselines.

1 Introduction

Artificial intelligence (AI) is increasingly reshaping chemical engineering by integrating data, domain knowledge, and automation into intelligent process systems [Chakraborty *et al.*, 2025]. Recent advances, including large language models and generative AI, have enabled data-driven process modeling, flowsheet design, and safety analysis, accelerating the transition toward adaptive and autonomous chemical manufacturing [Schweidtmann, 2024; Mann *et al.*, 2025].

Chlor-alkali production is a foundational chemical process that supplies essential chemicals such as caustic soda, chlorine, and hydrogen to numerous downstream industries [Kermeli and Worrell, 2025]. Figure 1, adapted from [Li *et al.*, 2021], illustrates the core working process of the electrolyzer, the key equipment for modern ion-exchange-membrane-based chlor-alkali production. Chlor-alkali production is highly energy-intensive, and each ton of alkali consumes about 2100-2600 kW·h of electricity. Therefore, energy conservation is a top priority regarding the economic benefits of chlor-alkali plants. To reserve energy, chlor-alkali production requires coordinated adjustment of multiple parameters (e.g., current, voltage, electrolyte flow, and temperature) to balance conflicting objectives: maximizing yield and current efficiency while minimizing energy consumption and extending equipment life, under strict safety and operational constraints [Ghanbarzadeh *et al.*, 2024].

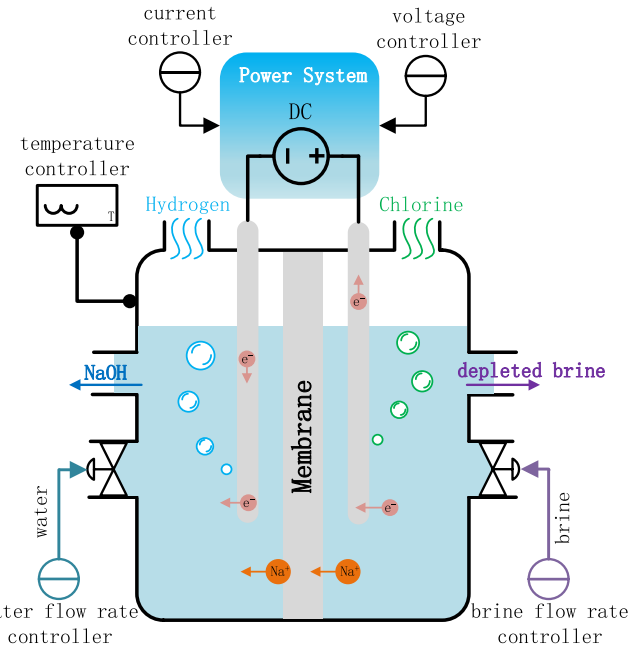


Figure 1: Working process of ion-exchange-membrane-based chlor-alkali production with electrolyzers [Li *et al.*, 2021].

Traditional optimization approaches, including gradient-based methods and evolutionary multi-objective algorithms, typically assume fixed operating conditions and require repeated re-optimization when the environment changes [Lee *et al.*, 2024; Silva *et al.*, 2024; Zhu *et al.*, 2024]. Such static formulations are poorly suited to the dynamic nature of modern chlor-alkali production, whose optimal operating strategies

must evolve over time due to objective environmental change factors like membrane aging, electrode degradation, and power supply fluctuations, as well as subjective environmental change factors such as total current limitations. In contrast, reinforcement learning (RL) learns with environmental changes via dynamic policy updates, exploration-exploitation balance, and lifelong learning, proving adaptive in process optimization across fields like chemicals. [Bloor *et al.*, 2025; Fortela *et al.*, 2025; Devarakonda *et al.*, 2025].

Despite this potential, RL-based multi-objective optimization for chlor-alkali process remains largely underexplored. Existing studies [Mirzazadeh *et al.*, 2008; Kaveh *et al.*, 2008; Ghanbarzadeh *et al.*, 2024] primarily rely on heuristic algorithms or single-objective learning methods that fail to capture the high-dimensional, multi-objective, and time-varying characteristics of chlor-alkali production.

To address these challenges, this paper proposes MORL-CA, a multi-objective RL framework for dynamic process optimization in chlor-alkali production. MORL-CA facilitates continuous policy adaptation to dynamic environments via continual online learning.

The main contributions are summarized as follows:

- **Dynamic Pareto-Tracking Formulation for Industrial Process Optimization.** We formulate chlor-alkali process parameter optimization as a dynamic multi-objective sequential decision problem, aiming to continuously track a time-varying Pareto set rather than repeatedly solving static optimizations.
- **State-Aware Adaptive Objective Weighting for Dynamic MORL.** We introduce a state-conditioned adaptive objective weighting mechanism within a multi-critic actor-critic framework, enabling localized Pareto-improving policy updates under non-stationary conditions.
- **An Offline-to-Online MORL Framework for Low-Cost Industrial Adaptation.** We propose an industrial-oriented MORL framework that combines offline pretraining with constrained online refinement, achieving rapid adaptation to environmental changes with limited online interactions.

2 Related Work

2.1 Multi-Objective Reinforcement Learning

Multi-objective RL (MORL) extends classical RL to vector-valued rewards and aims to learn policies that balance conflicting objectives [Rodríguez Soto *et al.*, 2024]. Most MORL methods rely on scalarization or utility-based formulations to encode preferences, enabling either preference-conditioned policies or online-adjusted objective weights [Rodríguez Soto *et al.*, 2024; Shi *et al.*, 2025].

To support gradient-based optimization, actor-critic architectures with multiple critics or policy heads have been widely adopted, while closely related multi-agent formulations assign individual objectives to separate agents and coordinate their decisions [Lu *et al.*, 2020; Shen *et al.*, 2022; Geng *et al.*, 2025]. More recently, adaptive-weight (dynamic-reward) mechanisms have been proposed to adjust preferences online, improving robustness under time-varying operating conditions [Zhang and Ou, 2025].

In industrial and chemical engineering applications, MORL is often combined with surrogate modeling and hybrid offline-online training to leverage historical data while ensuring safe online adaptation [Lin *et al.*, 2024; Lu *et al.*, 2024].

Building on these ideas, MORL-CA integrates a multi-critic architecture with an adaptive-weight network to address continuous multi-objective chlor-alkali production process optimization.

2.2 Reinforcement Learning for Sequential Decision Optimization

RL has been increasingly used to reformulate optimization problems as sequential decision processes [Goldie *et al.*, 2024]. Existing approaches can be broadly categorized into constructive methods, which build solutions step by step [Liu *et al.*, 2024; Hottung *et al.*, 2025], and update-based methods, which learn iterative improvement operators or parameter update policies [Goldie *et al.*, 2024].

In chemical and process optimization, control-informed RL and offline-online deployment strategies have shown effectiveness in handling safety constraints and time-varying dynamics [Devarakonda *et al.*, 2025; Bloor *et al.*, 2025].

However, sequential decision formulations for chlor-alkali process parameter tuning remain largely unexplored. MORL-CA addresses this gap by modeling parameter adjustment as an update-based sequential decision problem, enabling continuous refinement toward the Pareto front.

2.3 Adaptation to Non-Stationary and Safety-Constrained Settings

In dynamic chemical production (e.g., chlor-alkali), operating conditions drift due to changing current limits, equipment aging, etc. Under such non-stationarity, heuristic methods like NSGA-II are static and cannot adapt online. Multi-objective Bayesian optimization (MOBO) [Rodemann and Augustin, 2024] assumes stationarity and requires costly GP retraining to track shifts. Thus, both are less suitable for long-running industrial processes with evolving conditions.

Adaptation under non-stationary dynamics has become a central topic in RL, with meta-RL, context-conditioned policies, and continual learning enabling rapid response to distribution shifts [Marin Moreno *et al.*, 2024; Xu and Zhu, 2024; Jackson *et al.*, 2024]. In safety-critical industrial settings, a hybrid paradigm combining offline pretraining with conservative online fine-tuning is commonly adopted to balance safety, data efficiency, and adaptability [Chen and Luo, 2025; Zheng *et al.*, 2024].

Motivated by these advances, MORL-CA adopts a hybrid offline-online learning framework, where historical data are used to initialize a stable policy and online interaction enables continuous adaptation to non-stationary multi-objective operating conditions.

3 The MORL-CA Framework

3.1 Problem Definition

We consider a multi-objective process optimization problem in an industrial chlor-alkali electrolyzer system. The primary goal is to find an optimal Pareto set that dynamically coordinates electrolyzers to achieve competing objectives, modeling it as a sequential decision-making problem for multi-objective RL.

Decision Variables and Observation Space

The system consists of M parallel electrolytic cells operating under shared electrical and operational constraints. For each cell $m \in \{1, ..., M\}$, there are:

- K **controllable parameters**, including cell current, cathode outlet temperature, etc;
- L **uncontrollable but observable parameters**, including product caustic concentration and product caustic temperature.

Let $\mathbf{x}_m \in \mathbb{R}^K$, $\mathbf{z} \in \mathbb{R}^L$ denote the controllable parameters of cell m and the shared uncontrollable parameters, respectively. The full system observation vector is defined as

$$\mathbf{x} = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M, \mathbf{z}\right]^\top \in \mathcal{X} \subset \mathbb{R}^{MK+L}, \tag{1}$$

where $\mathcal{X}$ denotes the feasible observation space bounded by operational limits:

$$\mathcal{X} = [x_{min}, x_{max}]. \tag{2}$$

Optimization Objectives

Without loss of generality, we assume there are three objectives. The objectives are modeled as continuous functions of system states:

$$\mathbf{f}(\mathbf{x}) = \left[f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})\right], \tag{3}$$

where f_1 and f_2 are to be maximized, while f_3 is to be minimized. To be handled uniformly as a minimization problem, we define

$$\hat{f}_1 = -f_1(\mathbf{x}), \hat{f}_2 = -f_2(\mathbf{x}), \hat{f}_3 = f_3(\mathbf{x}). \tag{4}$$

Given these objectives, we derive a Pareto-optimal set $\mathcal{P}^*$ with robust convergence and diversity:

$$\mathcal{P}^* \subset \mathcal{X}, \text{ Pareto-nondominated w.r.t. } (\hat{f}_1, \hat{f}_2, \hat{f}_3). \tag{5}$$

Moreover, the solution must adapt iteratively: as the operating environment changes, the method should swiftly update $\mathcal{P}^*$ to reflect current constraints and conditions.

Constraints

The system is subject to a global electrical constraint on total operating current:

$$\sum_{m=1}^{M} I_m \leq I_{\max}, \tag{6}$$

where I_m is the operating current of cell m, and I_{max} the maximum total current allowed.

Equipment safety and operational feasibility impose additional constraints on the bounded domain $\mathcal{X}$.

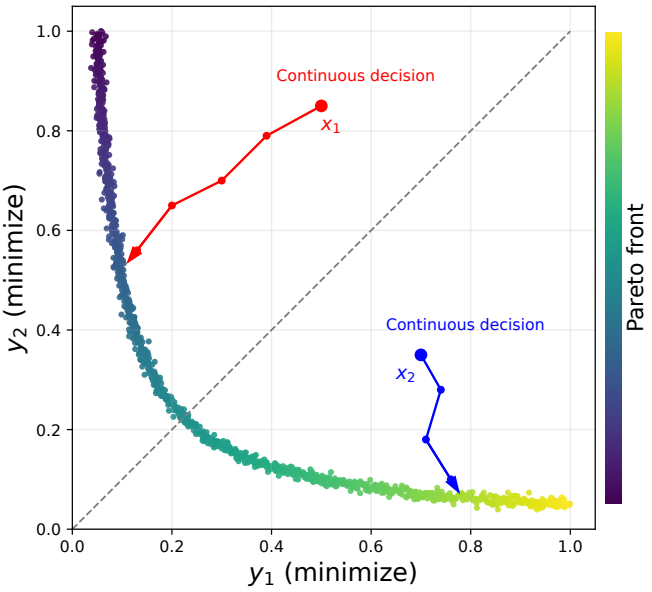


Figure 2: MOP is reformulated as a continuous sequential decision-making process.

Reformulating as a Sequential Decision Problem

To address the time-varying nature of chlor-alkali production, the multi-objective optimization is reformulated as a continuous sequential decision process.

Starting at $\mathbf{x}_0 \in \mathcal{X}$, the agent incrementally adjusts. At step t, it observes $\mathbf{x}_t$, takes action $\delta_t \in \mathbb{R}^{MK}$, and then:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \boldsymbol{\delta}_t, \tag{7}$$

subject to feasibility and operational constraints.

After a finite horizon T, the final operating point is:

$$\mathbf{x}_T = \mathbf{x}_0 + \sum_{t=0}^{T-1} \boldsymbol{\delta}_t, \tag{8}$$

with the objective of guiding $\mathbf{x}_T$ towards the Pareto-optimal set, as depicted in Figure 2. The adaptive objective weighting network influences $\mathbf{x}_t$'s change direction.

This method converts static multi-objective optimization into a dynamic continuous-control problem, ideally suited for RL, which excels in learning adaptive policies through sequential interaction, aligning well with continuous decision spaces and time-varying industrial settings.

3.2 Framework Overview

The MORL-CA framework consists of three modules: population initialization with offline policy pretraining, online RL, and inference and solution refinement.

First, offline pretraining leverages historical data to populate an experience buffer and establish a stable initial policy.

Second, online training initializes the environment with a Latin Hypercube Sampling (LHS)-generated population. The RL agent interacts continuously, selecting actions, storing transitions, and updating its policy. A training episode covers the entire population, allowing progressive policy adaptation.

Finally, in inference, NSGA-II generates initial candidates, which the trained RL policy refines through fixed decision steps toward the Pareto front. Adjustment trajectories are aggregated, pruned using SPEA2-based density estimation, and sorted to yield the final solution set.

Figure 3 illustrates the modules and their interactions.

3.3 Population Initialization and Offline Learning

Population initialization is customized for training and inference to balance exploration and convergence. During online training, we sample the initial population with LHS across the feasible parameter domain; LHS's space-filling property ensures extensive, uniform coverage and yields diverse experiences for policy learning.

Prior to online interaction, an offline pretraining stage leverages historical production data to populate an experience buffer and initialize an initial action-selection policy. This pretraining step stabilizes early learning and facilitates online adaptation.

During inference, initialization favors convergence: when a Pareto set from the immediately prior environment is available, it serves as the seed for the initial population; only at system startup, when no such prior set is available, do we generate the initial population with NSGA-II. This strategy reduces required refinement steps and accelerates convergence under incremental environment changes.

3.4 Reinforcement-Learning Agent Architecture

We utilize a weighted adaptive actor-critic framework, adapted from Soft Actor-Critic (SAC), and extend it to incorporate multiple critics for multi-objective optimization. Given n objectives, the environment yields a reward vector at time t:

$$\mathbf{r}_t = [r_{t,1}, r_{t,2}, \ldots, r_{t,n}]^\top \in \mathbb{R}^n. \tag{9}$$

We employ n critics $Q_i^\theta(s,a)$ (with parameters θ) and their corresponding target critics $Q_i^{\bar{\theta}}$. The stochastic policy $\pi_\phi(a|s)$ includes an entropy term regulated by $\alpha > 0$ (where $\alpha = exp(\eta)$ and η is optimized).

Environment Creation

Since direct interaction with the production plant is impractical, a data-driven surrogate environment is established for training, offering state transitions and multi-objective rewards.

State, Action and Reward

Our definitions of state, action, and reward in RL are as follows:

- State s_t: Measured production parameters (e.g., current density, cell voltage, temperature, flow rates).
- Action a_t: Incremental adjustments to controllable parameters.
- Reward: Vector $\mathbf{r}_t$ (see Equation (9)), with each component representing an optimization objective (e.g., yield, energy, current efficiency).

Multi-Critic Target and Critic Loss

For a sampled transition $(s_t, a_t, \mathbf{r}_t, s_{t+1})$, the action $a_{t+1} \sim \pi_\phi(\cdot|s_{t+1})$ is sampled and its log-probability $\log \pi_\phi(a_{t+1}|s_{t+1})$ is evaluated. Let $\alpha = exp(\eta)$ denote the entropy temperature, we define the soft Bellman backup for the i-th objective as

$$\mathcal{B}_i(s_{t+1}) \triangleq Q_i^{\bar{\theta}}(s_{t+1}, a_{t+1}) - \alpha \log \pi_\phi(a_{t+1} \mid s_{t+1}). \tag{10}$$

The corresponding TD-target is then given by

$$\hat{Q}_{t,i} = r_{t,i} + \gamma(1 - d_t)\mathcal{B}_i(s_{t+1}), \quad i = 1, \ldots, n. \tag{11}$$

Stacking these targets yields $\hat{Q}_t \in \mathbb{R}^n$. Each critic Q_i^θ is trained to regress $\hat{Q}_{t,i}$ with mean squared error:

$$\mathcal{L}_{\text{critic},i}(\theta) = \mathbb{E}\Big[\big(Q_i^\theta(s_t, a_t) - \hat{Q}_{t,i}\big)^2\Big]. \tag{12}$$

State-Aware Adaptive Objective Weighting

To adapt RL to multi-objective optimization, we equip the actor with a state-aware adaptive objective weighting (AOW) network $h_\psi : \mathcal{S} \to \mathbb{R}^n$, producing weights:

$$\mathbf{w}(s) = \mathrm{softmax}\big(h_\psi(s)\big), \ \sum_{i=1}^{n} w_i(s) = 1, \tag{13}$$

where $w_i(s)$ reflects the emphasis on objective i at state s. This enables each state to "exploit its strengths": states excelling in certain objectives receive higher corresponding weights, fostering focused decisions and faster convergence to the nearest Pareto region.

The AOW is learned incrementally via supervised regression. For each sampled state s, we compute normalized objectives $\hat{\mathbf{f}}(s)$, derive goodness scores $g(s) = 1 - \hat{\mathbf{f}}(s)$, and establish target weights $\mathbf{w}^*(s) = \mathrm{softmax}\big(g(s)\big)$, training h_ψ to regress $\mathbf{w}^*(s)$.

Actor Update

Given per-objective critics $Q_i(s,a)$, we define an aggregated state-action value by linearly combining the critics with the state-dependent weights from the AOW:

$$Q_{\text{agg}}(s,a) = \sum_{i=1}^{n} w_i(s) \times Q_i(s,a), \tag{14}$$

where $w_i(s)$ come from $h_\psi(s)$. The actor maximizes this aggregated value with an entropy regularizer. Formally, the actor loss is:

$$\mathcal{L}_{\text{actor}}(\phi) = \mathbb{E}\Big[- Q_{\text{agg}}(s,a) + \alpha \log \pi_\phi(a \mid s)\Big], \tag{15}$$

and the entropy coefficient $\alpha = exp(\eta)$ is adapted by minimizing

$$\mathcal{L}(\eta) = \mathbb{E}\Big[- \eta\big(\log \pi_\phi(a_t|s_t) + \bar{\mathcal{H}}\big)\Big], \tag{16}$$

where $\bar{\mathcal{H}}$ is a target entropy. This formulation follows the standard maximum entropy reinforcement learning framework, adapted here to the multi-objective setting with state-aware preferences.

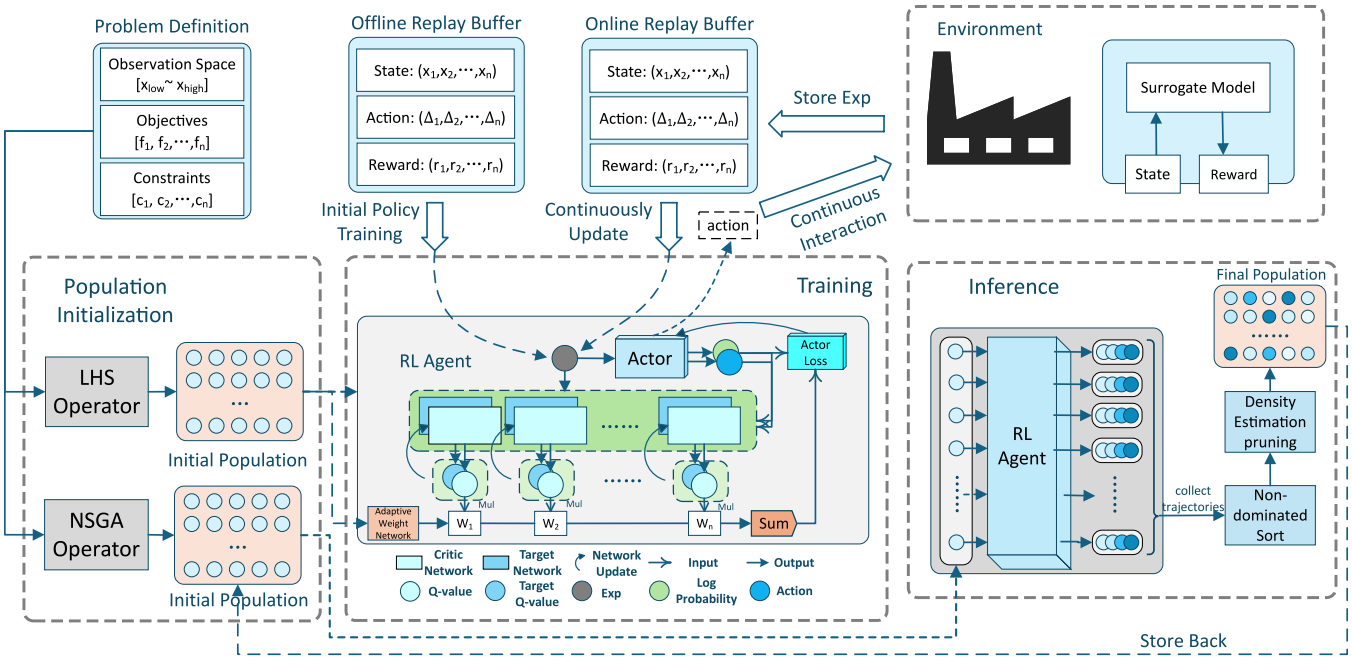


Figure 3: MORL-CA's population initialization, offline and online learning, and inference modules.

3.5 Online Learning and Inference

Following offline pretraining on historical data, the agent proceeds online, interacting with the live plant (or its surrogate) and accumulates transitions (s, a, r, s') into an online replay buffer. Periodically, the agent refines its policy via actor-critic gradient updates on mini-batches randomly sampled from the buffer, maintaining stability inherited from offline initialization while adapting to gradual or sudden environmental changes.

Inference combines population initialization with agent-guided refinement. An initial population of $\mathcal{N}$ candidates is refined by executing the learned policy for a predetermined, consistent number of sequential actions. For a candidate x_i, the refinement trajectory proceeds as follows:

$$\mathcal{T}_i = \{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(steps)}\}, \tag{17}$$

and the pooled candidate set comprises a total of $S = \mathcal{N} \times steps$ solutions. The set is purified via: (1) non-dominated sorting to rank solutions based on their optimality and (2) SPEA2-style density pruning that favors solutions in sparse regions [Zaferani *et al.*, 2024]. The resulted solutions constitute the final Pareto front approximation, which is then presented to the decision-makers.

This online-inference cycle facilitates ongoing policy improvement by leveraging deployment data, simultaneously generating a concise yet diverse Pareto set. This is achieved without the necessity for exhaustive re-optimization following every alteration in the environment.

4 Experiments and Results

4.1 Experimental Setup and Dynamic Protocol

We build a surrogate environment using a real industrial dataset from 8 electrolytic cells, each with 7 controllable inputs (current, cathode outlet temperature, acid dosing flow, anolyte flow, catholyte flow, brine feed flow, and catholyte header level) and 2 non-controllable observables (product caustic concentration and product caustic temperature), totaling $8 \times 7 + 2 = 58$ inputs. An MLP regression model maps these inputs to three objectives, serving as the interactive surrogate for RL training and inference.

The three competing objectives are: maximize daily caustic soda production ($f_1(\mathbf{x})$), maximize caustic current efficiency ($f_2(\mathbf{x})$), and minimize specific energy use ($f_3(\mathbf{x})$).

To simulate dynamics, we vary the total-current limit I_{max} from 86 kA to 119 kA. This reflects industrial chlor-alkali practice, where current density is adjusted to electricity prices and grid constraints—a structured, operationally meaningful shift rather than random noise.

For each new current limit, the policy is adapted with 3 online episodes. The Pareto set is generated by NSGA-II (86 kA cold-start) or seeded from the previous Pareto set, then refined and pruned to at most 1,000 solutions.

This protocol embodies a practical workflow: swift, incremental online adjustments followed by efficient inference, avoiding full re-optimization after each change.

4.2 Correlation Analysis of State-Aware AOW with the Objective

Before comparative experiments, we verify the AOW network's correctness and efficacy to ensure it accurately reflects local objective dominance. We perform a targeted correlation analysis by randomly sampling 1,000 states from each of three distinct objective-space regions (region_1, region_2, region_3), where $\hat{f}_1(\mathbf{x})$, $\hat{f}_2(\mathbf{x})$, and $\hat{f}_3(\mathbf{x})$ are locally dominant, respectively. Each sampled state $\mathbf{x}$ is processed by the trained AOW network to yield $\mathbf{w}(\mathbf{x}) = [w_1, w_2, w_3]^\top$, from

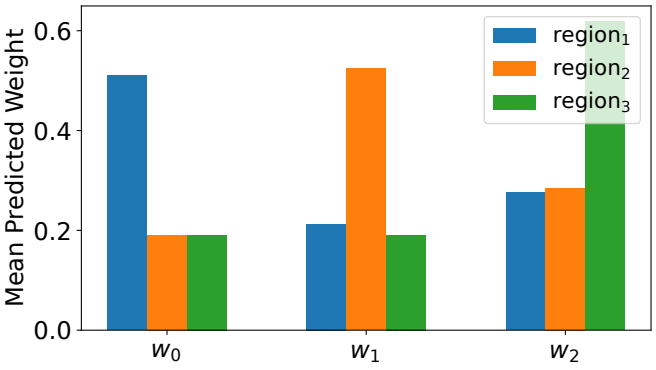


Figure 4: Average adaptive objective weights in different objective-dominant regions.

which per-region averages $\overline{w}_i$ are computed.

Figure 4 presents the mean weight vectors per region, revealing that states from the $\hat{f}_1$-dominant region exhibit significantly higher w_1 values (similarly for w_2 and w_3). This indicates that the AOW network prioritizes the locally dominant objective, fostering swift convergence toward the nearest Pareto region. Such empirical evidence underscores the AOW network's state-aware weighting mechanism.

4.3 Baselines

We compare MORL-CA against representative state-of-the-art evolutionary and RL baselines multi-objective problems: C3M [Sun *et al.*, 2022], c-DPEA [Ming *et al.*, 2021], CMDEIPCM [Liang *et al.*, 2022], IDBEA [Sun *et al.*, 2020], ToP [Liu and Wang, 2019], MADDPG [Wang *et al.*, 2022], and SAC [Haarnoja *et al.*, 2018].

All heuristic baselines are run on PlateEMO [Tian *et al.*, 2017], with a fixed population size of 1,000 and default hyperparameters for all algorithms.

For RL baselines like MADDPG and SAC, we employ the same online adaptation as MORL-CA: agents interact with the updated environment for three episodes to refine their policies. Notably, SAC, being single-objective, uses a scalarized reward with equal weights for all objectives.

4.4 Evaluation

Metrics

We evaluate the quality of Pareto set using two domain-standard metrics: Hypervolume (HV) [Lu *et al.*, 2024] and Inverted Generational Distance (IGD) [Wang *et al.*, 2022].

To qualify performance stability across environmental changes, we introduce a smoothness metric for the time series of evaluation metrics. Let $y = [y_1, ..., y_T]$ be the metric sequence (HV or IGD) across T constraint scenarios. We define the smoothness metric, mean absolute first difference (MAFD), as:

$$MAFD(y) = \frac{1}{T-1}\sum_{t=1}^{T-1}|y_{t+1} - y_t|. \tag{18}$$

A lower MAFD indicates a smoother metric curve, indicating superior dynamic adaptation. We also present the average ranking across scenarios for HV and IGD, where a lower rank is preferable.

Reference Pareto front

When optimizing parameters in the chlor-alkali process, the Pareto front lacks an analytical solution. For each scenario k, we construct an empirical reference front by aggregating solutions from all methodologies and subsequently retaining only the non-dominated points:

$$\mathcal{R}_k = NDS(\bigcup_m \mathcal{S}_k^{(m)}), \tag{19}$$

where $\mathcal{S}_k^{(m)}$ is the solution set of algorithm m under scenario k, and $NDS(\cdot)$ is the non-dominated sort operator. IGD and HV are computed with respect to $\mathcal{R}_k$.

4.5 Results Overview

Figures 5a and 5b show HV and IGD respectively for all algorithms across varying current limits. Table 1 summarizes each method's average HV and IGD ranks, along with their MAFD values.

MORL-CA leads with the best average rankings in both IGD and HV across tests. It also records the lowest MAFD for both metrics, indicating smaller performance fluctuations as the total-current limit changes. These findings demonstrate MORL-CA's superior Pareto quality and its stable performance amid environmental shifts.

Algorithms	HV Rank	IGD Rank	HV MAFD	IGD MAFD
C3M	2.88	3.71	1.08e+5	2.91e-1
c-DPEA	2.35	2.06	8.78e+4	2.65e-1
CMDEIPCM	6.38	6.12	6.57e+4	2.83e-1
ToP	7.97	7.97	9.20e+4	2.31e-1
IDBEA	4.71	4.94	1.45e+5	2.97e-1
MADDPG	6.00	6.76	4.82e+5	2.80e-1
SAC	4.65	4.94	8.52e+4	2.88e-1
MORL-CA	**1.06**	**1.00**	**6.42e+4**	**1.81e-1**

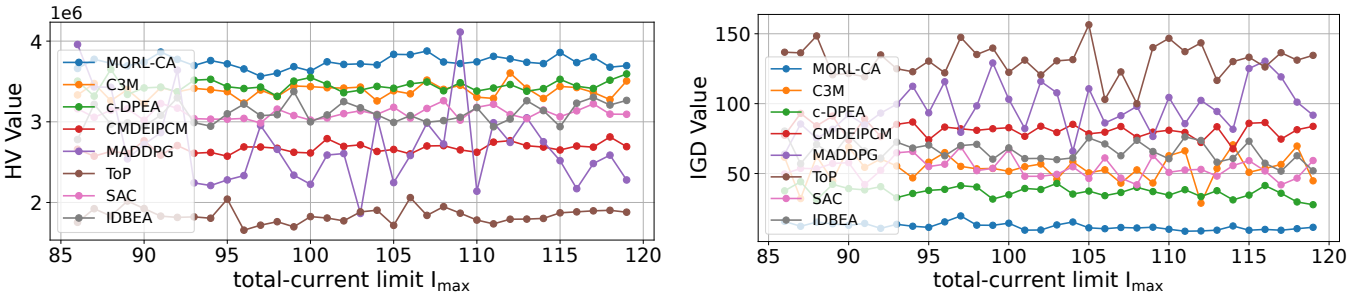
Table 1: The average HV and IGD rankings of each algorithm, and the MAFD indices for HV and IGD. Rank values are presented with two decimal places.

4.6 Ablation Study

We conduct two ablation studies to quantify the impacts of (1) the online learning module and (2) the density-estimation pruning used in inference. Results are summarized in Figure 6 and Table 2.

Online Learning Ablation

We disable online updates, relying solely on a pretrained policy for inference. Comparing HV and IGD against full MORL-CA under current-limit sequences, Figure 6 reveals that the algorithm with online learning (W) outperforms its counterpart without (W/o), showing smaller indicator fluctuations across environments. This confirms that brief online adaptation significantly improves both Pareto quality and stability under changing constraints.



(a) HV performance of each algorithm.

(b) IGD performance of each algorithm.

Figure 5: A comparative analysis of algorithm performance in terms of HV and IGD under different total current constraints.

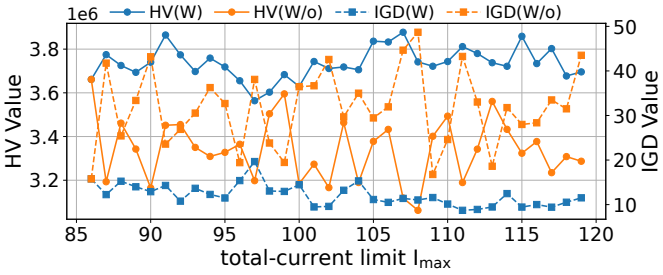


Figure 6: Ablation experiment results of online learning module.

Density-Estimation Pruning Ablation

We evaluate MORL-CA with and without SPEA2-style density-estimation pruning (W vs. W/o DE). With DE, we retain 1,000 solutions per scenario; without DE, all candidates are kept. For each scenario, we monitor average HV, IGD, and solution-set size, then compute the relative change rate for each metric after enabling DE using the following formula:

$$\text{Chang}(\%) = \frac{\text{Metric}_{\text{W}} - \text{Metric}_{\text{W/o}}}{\text{Metric}_{\text{W/o}}} \tag{20}$$

Table 2 shows that density-estimation pruning greatly reduces the solution-set size, which simplifies decision-making and speeds up metric computation. However, it also degrades HV and IGD, indicating a clear trade-off. In practice, the pruning strength should be adjusted according to specific requirements.

Metric	W/o DE	W DE	Change (%)
HV	3.38e+6	3.08e+6	-8.85%
IGD	3.54e+1	5.32e+1	49.98%
Set Size	2.24e+3	1.00e+3	-55.35%

Table 2: Ablation results of density estimation (DE) pruning in the inference stage.

4.7 Analysis and Discussion

The experimental results can be summarized as follows:

- **Convergence and Diversity**: MORL-CA's multi-critic SAC framework with adaptive weight aggregation produces Pareto sets with superior convergence (lower IGD) and better diversity (higher HV) compared to classical MOEAs and baseline RL methods.
- **Dynamic Adaptation**: Upon environmental changes (new I_{max}), MORL-CA quickly updates with few online episodes and lightweight inference, recovering high-quality Pareto sets with smaller metric fluctuations.

In real chlor-alkali plants, environmental changes are moderate but frequent. MORL-CA's workflow—online policy adaptation plus fast re-inference from prior Pareto sets—requires far fewer environment evaluations than restarting a MOEA each time, leading to a much lower wall-clock cost for adapting to new constraints. This makes it ideal for near-real-time production adaptation.

5 Conclusion and Future Work

This paper presents MORL-CA, a dynamic multi-objective RL framework for process optimization in ion-exchange membrane-based chlor-alkali production. By integrating state-aware adaptive objective weighting with offline-to-online learning, MORL-CA continuously adapts to time-varying conditions and facilitates localized Pareto-improving updates without repeated optimization. Experiments conducted in an environment constructed from real industrial data demonstrate its effectiveness in tracking evolving Pareto trade-offs and achieving superior solutions under dynamic environments. The framework code is publicly available at https://github.com/hoshinojyunn/morl-ca-framework.

While all validation is performed in a realistic data-driven surrogate environment due to industrial safety constraints, the proposed MORL-CA framework is designed for practical deployment. Future work will implement conservative, constraint-guaranteed online fine-tuning for physical electrolyzers in closed-loop operation within physical chlor-alkali plants. In addition, although this study focuses on chlor-alkali production, the framework can be generalized to a broader class of dynamic multi-objective industrial control tasks. Future research will also investigate theoretical guarantees and extensions to other industrial domains.

Acknowledgements

This work was supported in part by the CUG Enterprise Cooperation Project (No. 20251961419).

References

[Bloor *et al.*, 2025] Maximilian Bloor, Akhil Ahmed, Niki Kotecha, Mehmet Mercangoz, Calvin Tsay, and Ehecatl Antonio del Rio-Chanona. Control-informed reinforcement learning for chemical processes. *Industrial & Engineering Chemistry Research*, 64:4966–4978, 2025.

[Chakraborty *et al.*, 2025] Arijit Chakraborty, Naz Pinar Taskiran, Rishab Kottooru, Vipul Mann, and Venkat Venkatasubramanian. Building hybrid AI models in chemical engineering: A tutorial review. *Computers Chemical Engineering*, 201:109236, 2025.

[Chen and Luo, 2025] Jiyang Chen and Na Luo. An offline-to-online reinforcement learning framework with trajectory-guided exploration for industrial process control. *Journal of Process Control*, 154:103535, 2025.

[Devarakonda *et al.*, 2025] Venkata Srikar Devarakonda, Wei Sun, Xun Tang, and Yuhe Tian. Recent advances in reinforcement learning for chemical process control. *Processes*, 13(6):1791, 2025.

[Fortela *et al.*, 2025] Dhan Lord B Fortela, Holden Broussard, Renee Ward, Carly Broussard, Ashley P Mikolajczyk, Magdy A Bayoumi, and Mark E Zappi. Soft actor-critic reinforcement learning improves distillation column internals design optimization. *ChemEngineering*, 9(2):34, 2025.

[Geng *et al.*, 2025] Minghong Geng, Shubham Pateria, Budhitama Subagdja, Lin Li, Xin Zhao, and Ah-Hwee Tan. L2m2: A hierarchical framework integrating large language model and multi-agent reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2025.

[Ghanbarzadeh *et al.*, 2024] Samira Ghanbarzadeh, Sergei Nikolaevich Mironov, Tzu-Chia Chen, Ayad F. Alkaim, Surendar Aravindhan, and Lakshmi Thangavelu. Determination of cell voltage and current efficiency in a chlor-alkali membrane cell based on machine learning approach. *Petroleum Science and Technology*, 42(15):1898–1910, 2024.

[Goldie *et al.*, 2024] Alexander D Goldie, Chris Lu, Matthew T Jackson, Shimon Whiteson, and Jakob Foerster. Can learned optimization make reinforcement learning less difficult? *Advances in Neural Information Processing Systems*, 37:5454–5497, 2024.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018.

[Hottung *et al.*, 2025] André Hottung, Mridul Mahajan, and Kevin Tierney. Polynet: Learning diverse solution strategies for neural combinatorial optimization. In *International Conference on Learning Representations*, volume 2025, pages 80417–80435, 2025.

[Jackson *et al.*, 2024] Matthew T Jackson, Chris Lu, Louis Kirsch, Robert Lange, Shimon Whiteson, and Jakob Foerster. Discovering temporally-aware reinforcement learning algorithms. In *International Conference on Learning Representations*, volume 2024, pages 13278–13296, 2024.

[Kaveh *et al.*, 2008] Narjes Shojai Kaveh, Seyed Nezameddin Ashrafizadeh, and Fereidoon Mohammadi. Development of an artificial neural network model for prediction of cell voltage and current efficiency in a chlor-alkali membrane cell. *Chemical Engineering Research and Design*, 86(5):461–472, 2008.

[Kermeli and Worrell, 2025] Katerina Kermeli and Ernst Worrell. Energy efficiency and cost-saving opportunities for the chlor-alkali industry. *US Environmental Protection Agency, Washington, DC*, 2025.

[Lee *et al.*, 2024] Chia-Yen Lee, Yi-Tao Huang, and Peng-Jen Chen. Robust-optimization-guiding deep reinforcement learning for chemical material production scheduling. *Computers & Chemical Engineering*, 187:108745, 2024.

[Li *et al.*, 2021] Kai Li, Qun Fan, Hongyuan Chuai, Hai Liu, Sheng Zhang, and Xinbin Ma. Revisiting chlor-alkali electrolyzers: from materials to devices. *Transactions of Tianjin University*, 27(3):202–216, 2021.

[Liang *et al.*, 2022] Jing Liang, Xuanxuan Ban, Kunjie Yu, Kangjia Qiao, and Boyang Qu. Constrained multiobjective differential evolution algorithm with infeasible-proportion control mechanism. *Knowledge-Based Systems*, 250:109105, 2022.

[Lin *et al.*, 2024] Qian Lin, Zongkai Liu, Danying Mo, and Chao Yu. An offline adaptation framework for constrained multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 37:140292–140319, 2024.

[Liu and Wang, 2019] Zhi-Zhong Liu and Yong Wang. Handling constrained multiobjective optimization problems with constraints in both the decision and objective spaces. *IEEE Transactions on Evolutionary Computation*, 23(5):870–884, 2019.

[Liu *et al.*, 2024] Qidong Liu, Chaoyue Liu, Shaoyao Niu, Cheng Long, Jie Zhang, and Mingliang Xu. 2d-ptr: 2d array pointer network for solving the heterogeneous capacitated vehicle routing problem. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 1238–1246, 2024.

[Lu *et al.*, 2020] Renzhi Lu, Yi-Chang Li, Yuting Li, Junhui Jiang, and Yuemin Ding. Multi-agent deep reinforcement learning based demand response for discrete manufacturing systems energy management. *Applied Energy*, 276:115473, 2020.

[Lu *et al.*, 2024] Yongfan Lu, Bingdong Li, and Aimin Zhou. Are you concerned about limited function evaluations: data-augmented pareto set learning for expensive multi-objective optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14202–14210, 2024.

[Mann *et al.*, 2025] Vipul Mann, Jingyi Lu, Venkat Venkatasubramanian, and Rafiqul Gani. A perspective on artificial intelligence for process manufacturing. *Engineering*, 52:60–67, 2025.

[Marin Moreno *et al.*, 2024] Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard, and Nadia Oudjane. Metacurl: Non-stationary concave utility reinforcement learning. *Advances in Neural Information Processing Systems*, 37:123091–123126, 2024.

[Ming *et al.*, 2021] Mengjun Ming, Anupam Trivedi, Rui Wang, Dipti Srinivasan, and Tao Zhang. A dual-population-based evolutionary algorithm for constrained multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 25(4):739–753, 2021.

[Mirzazadeh *et al.*, 2008] Taher Mirzazadeh, Fereidoon Mohammadi, Mohammad Soltanieh, and Ezzatollah Joudaki. Optimization of caustic current efficiency in a zero-gap advanced chlor-alkali cell with application of genetic algorithm assisted by artificial neural networks. *Chemical Engineering Journal*, 140(1-3):157–164, 2008.

[Rodemann and Augustin, 2024] Julian Rodemann and Thomas Augustin. Imprecise bayesian optimization. *Knowledge-Based Systems*, 300:112186, 2024.

[Rodríguez Soto *et al.*, 2024] Manel Rodríguez Soto, Juan A Rodriguez-Aguilar, and Maite Lopez-Sanchez. An analytical study of utility functions in multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 37:77726–77747, 2024.

[Schweidtmann, 2024] Artur M Schweidtmann. Generative artificial intelligence in chemical engineering. *Nature Chemical Engineering*, 1(3):193–193, 2024.

[Shen *et al.*, 2022] Rendong Shen, Shengyuan Zhong, Xin Wen, Qingsong An, Ruifan Zheng, Yang Li, and Jun Zhao. Multi-agent deep reinforcement learning optimization framework for building energy system with renewable energy. *Applied Energy*, 312:118724, 2022.

[Shi *et al.*, 2025] Yucheng Shi, David Lynch, and Alexandros Agapitos. Ucb-driven utility function search for multi-objective reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 163–178. Springer, 2025.

[Silva *et al.*, 2024] Jonathan R. Silva, Thiago A.M. Euzébio, and Márcio F. Braga. Control of conventional continuous thickeners via proximal policy optimization. *Minerals Engineering*, 214:108761, 2024.

[Sun *et al.*, 2020] Jiaze Sun, Jiahui Deng, and Yang Li. Indicator & crowding distance-based evolutionary algorithm for combined heat and power economic emission dispatch. *Applied Soft Computing*, 90:106158, 2020.

[Sun *et al.*, 2022] Ruiqing Sun, Juan Zou, Yuan Liu, Shengxiang Yang, and Jinhua Zheng. A multistage algorithm for solving multiobjective optimization problems with multiconstraints. *IEEE Transactions on Evolutionary Computation*, 27(5):1207–1219, 2022.

[Tian *et al.*, 2017] Ye Tian, Ran Cheng, Xingyi Zhang, and Yaochu Jin. PlatEMO: A MATLAB platform for evolutionary multi-objective optimization. *IEEE Computational Intelligence Magazine*, 12(4):73–87, 2017.

[Wang *et al.*, 2022] Zhenhui Wang, Juan Lu, Chaoyi Chen, Junyan Ma, and Xiaoping Liao. Investigating the multi-objective optimization of quality and efficiency using deep reinforcement learning. *Applied Intelligence*, 52(11):12873–12887, 2022.

[Xu and Zhu, 2024] Siyuan Xu and Minghui Zhu. Meta-reinforcement learning with universal policy adaptation: Provable near-optimality under all-task optimum comparator. *Advances in Neural Information Processing Systems*, 37:2977–3026, 2024.

[Zaferani *et al.*, 2024] Seyed Peiman Ghorbanzade Zaferani, Mahmoud Kiannejad Amiri, Mohammad Reza Sarmasti Emami, Sasan Zahmatkesh, Mostafa Hajiaghaei-Keshteli, and Hitesh Panchal. Prediction and optimization of sustainable fuel cells behavior using artificial intelligence algorithms. *International Journal of Hydrogen Energy*, 52:746–766, 2024.

[Zhang and Ou, 2025] Wenfan Zhang and Haijiao Ou. Reinforcement learning based multi objective task scheduling for energy efficient and cost effective cloud edge computing. *Scientific Reports*, 15(1):41716, Nov 2025.

[Zheng *et al.*, 2024] Yinan Zheng, Jianxiong Li, Dongjie Yu, Yujie Yang, Shengbo Li, Xianyuan Zhan, and Jingjing Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. In *International Conference on Learning Representations*, volume 2024, pages 3838–3867, 2024.

[Zhu *et al.*, 2024] Zhiwei Zhu, Minglei Yang, Wangli He, Renchu He, Yunmeng Zhao, and Feng Qian. A deep reinforcement learning approach to gasoline blending real-time optimization under uncertainty. *Chinese Journal of Chemical Engineering*, 71:183–192, 2024.