Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Reinforcement Learning Supervisory Control with Fuzzy-Logic Reward for Multistage Gas Compression

José R. Torraca Neto[a]*, Sergio A. C. Giraldo[a], Mario C. M. M. Campos[c], Gustavo L. R. Caldas[b], Bruno D. O. Capron[a], and Argimiro R. Secchi[b]

[a] EPQB, School of Chemistry, LADES—Software Development Laboratory, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.
[b] Chemical Engineering Program, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.
[c] SmartAutomation, Rio de Janeiro, RJ, Brazil.
* Corresponding Author: joseneto@eq.ufrj.br.

ABSTRACT

Offshore natural gas compression systems are characterized by strong hydraulic coupling, nonlinear behaviour, and strict safety constraints, particularly in high-CO_2 production environments. Conventional decentralized PID control with anti-surge protection ensures reliable local regulation but often leads to poor plant-wide coordination and persistent offsets when multiple compression trains, recycle loops, and separation units interact dynamically. Although multivariable control strategies such as model predictive control can address these issues, their industrial application remains limited by modeling effort, computational demand, and robustness concerns. This work presents a hybrid supervisory control framework in which reinforcement learning (RL) augments an existing PI-based architecture for an offshore gas compression system with membrane-based CO_2 separation. A Proximal Policy Optimization (PPO) agent is trained on a dynamic digital-twin model of export, CO_2, injection, and bypass compression trains with shared headers, recycle flows, and thermal constraints. The RL agent acts exclusively at the supervisory level, providing bounded incremental adjustments to pressure and bypass setpoints, while all regulatory PI loops, anti-surge protections, and safety logic remain unchanged. Closed-loop simulations under representative disturbance scenarios show that the RL-supervised architecture significantly improves plant-wide coordination. Compared to fixed-setpoint PI operation, total tracking error (IAE) and time-weighted error (ITAE) are reduced by approximately 54% and 56%, respectively, using smooth, low-magnitude supervisory actions. In addition, RL achieves performance close to that of a nonlinear MPC benchmark based on a random-shooting, MPPI (Model-Predictive Path Integral) formulation, while requiring substantially lower online computational effort. The results demonstrate that RL can be effectively integrated as a supervisory layer in offshore gas compression systems without replacing established industrial control infrastructure, although additional safety mechanisms are still required to ensure stricter constraint satisfaction.

Keywords: Offshore gas compression, reinforcement learning, supervisory control, PI control

INTRODUCTION

Natural gas compression systems play a central role in offshore oil and gas production, particularly in fields with high CO_2 content such as Brazil's pre-salt reservoirs. In these environments, compression is essential for enabling CO_2 separation, gas export, reinjection, and downstream processing, while simultaneously representing one of the largest contributors to onboard energy consumption and operational cost [1, 2]. Modern offshore facilities typically rely on multistage centrifugal compressors operating under tight constraints on pressure, flow rate, rotational speed, and surge margin, which must be satisfied under highly variable operating conditions.

In industrial practice, gas compression trains are commonly controlled using layered architectures based on proportional–integral–derivative (PID) controllers, complemented by anti-surge protection systems and hard safety interlocks [3]. While such architectures provide reliable local control and are well established in industry, their performance deteriorates in highly coupled systems where multiple compressors, recycle valves, membranes, and headers interact dynamically. In offshore facilities, interactions between parallel compression trains create strong loop interactions, often resulting in oscillatory behaviour, conservative operation, and limited plant-wide coordination [4].

To overcome these limitations, advanced control strategies have been extensively investigated in offshore applications over the last two decades. Model Predictive Control (MPC), in particular, has been successfully applied to gas processing and compression systems, improving constraint handling, multivariable coordination, and operational efficiency compared to decentralized PID control [5, 6]. Nevertheless, the deployment of MPC in large-scale offshore compression systems remains challenging due to the high modeling effort required, the nonlinear nature of compressor maps, frequent changes in operating regimes, the need for robust performance under significant disturbances and uncertainty, and the required computational infrastructure.

At the same time, increasing attention has been given to the environmental impact of offshore gas production. Compression systems are responsible for a substantial fraction of platform-wide energy consumption, typically supplied by gas turbines, which account for the majority of CO_2 emissions in offshore installations [7]. In addition, inefficient operation may lead to increased flaring, recycle flows, and pressure violations, further increasing emissions and reducing overall efficiency. As a result, improving the operational efficiency and stability of gas compression systems is directly aligned with broader efforts to reduce greenhouse gas emissions and comply with international climate targets [8].

In this context, data-driven control methods based on Reinforcement Learning (RL) have emerged as a promising alternative for the control of complex, nonlinear, and highly coupled process systems. By learning control policies through interaction with a dynamic environment, RL techniques can address multivariable coordination problems without relying on explicit linearized models, making them attractive for systems characterized by nonlinear dynamics, strong interactions, and changing operating conditions [9, 10]. Recent studies have reported encouraging results for RL-based control and optimization in energy systems and chemical processes, including continuous control problems with constraints, disturbances, and nonlinear dynamics [11]. These works show the potential of modern policy-gradient and actor–critic methods to outperform conventional controllers in terms of tracking performance, efficiency, and adaptability.

Nevertheless, the direct application of RL to safety-critical industrial systems remains limited. Key challenges include guaranteeing constraint satisfaction, ensuring robustness under unmodelled disturbances, maintaining interpretability for operators, and safely integrating learning-based components with existing control infrastructures [12]. These concerns have motivated increasing interest in hybrid and hierarchical architectures that combine RL with conventional control layers.

A pragmatic approach to mitigate these limitations consists in embedding RL within hierarchical control architectures, in which learning-based agents operate at a supervisory level, while conventional controllers remain responsible for direct actuator manipulation. In such hybrid schemes, the RL agent adjusts setpoints or high-level targets, whereas low-level PID controllers, anti-surge loops, and safety interlocks enforce fast dynamics and operational constraints [13]. This structure preserves the reliability and transparency of industrial control systems while enabling adaptive, plant-wide coordination and optimization. Related concepts have been explored in supervisory, economic, and learning-based control layers; however, their application to multistage offshore gas compression systems with CO_2 separation and strong hydraulic coupling between trains remains relatively underexplored.

This work proposes a hybrid PI–RL supervisory control strategy for an offshore gas compression system. It comprises multiple interacting compression trains and a membrane-based CO_2 separation unit. A dynamic plant model was developed as a simplified digital twin, incorporating centrifugal compressor models, composition-dependent thermodynamics, valve flow dynamics, recycle loops, and operational constraints representative of industrial practice [1, 4]. The RL agent, trained using Proximal Policy Optimization (PPO) [14], acts exclusively on supervisory setpoints related to pressure and capacity, while conventional PI controllers and anti-surge protections remain unchanged.

To ensure learning stability and reflect operational priorities, a structured reward-shaping strategy was adopted. In addition to tracking export flow and key suction pressures, the reward incorporates soft penalties for violations of compressor discharge temperature limits, excessive anti-surge recycle activity, undesirable valve usage, and abrupt setpoint movements. Temperature constraints are enforced through continuous penalties activated beyond predefined soft limits, discouraging sustained operation at elevated discharge temperatures that could compromise equipment integrity and efficiency. This multi-term reward formulation, inspired by fuzzy-logic concepts and robust loss functions, introduces tolerance regions around nominal operating targets while penalizing asymmetric deviations, thereby preventing overly aggressive control actions and improving training robustness under disturbances [15].

The proposed approach was evaluated using closed-loop simulations based on real offshore gas compression system subject to disturbances in inlet flows and operating conditions. Performance was assessed relative to conventional fixed-setpoint PI operation using standard control performance metrics.

METHODOLOGY

1. Case study and control objective

This work considers the offshore gas-compression and high-CO_2 management system of an offshore platform. The unit has strong interactions between separation, compression, export, and reinjection. Figure 1 presents the process flow diagram (PFD) used as the reference for the modeling and control design. The system comprises: (i) a membrane-based CO_2 separation unit generating retentate (treated gas) and permeate (CO_2-rich gas), (ii) export compression, (iii) CO_2 compression for reinjection/EOR, (iv) injection compression, and (v) bypass routing that allows partial or full diversion around the membrane unit depending on the operating mode (export vs. injection priority).

The routing and coupling structure that drives the control problem can be seen in Figure 1: after the membrane split step, the retentate flow is routed for compression and exportation, while the permeate flows for CO_2 compression. The bypass stream modifies both membrane feed and downstream load. Shared headers link export and injection trains; thus, changes in exportation demand or bypass fraction propagate through suction pressures, discharge pressures, and recycle activity.

The operational goal is to maintain safe and efficient operation under disturbances (production flow swings, routing changes, header pressure shifts) while ensuring:

- Header and train stability: suction/discharge pressures remain within allowable envelopes to prevent compressor trips and surge;
- Export delivery tracking: export flow (or export capacity) follows a target consistent with pipeline constraints and gas quality;

CO_2 management: membrane operation and bypass routing are coordinated so that the CO_2 content in the treated gas (when exporting) is kept within specification, while CO_2-rich permeate is compressed for reinjection when required;

- Coupling-aware coordination: shared headers and routing decisions couple export and injection demands, and membrane permeate pressure is coupled to the suction setpoint of the CO_2 train, impacting separation driving force and downstream compression load.

These couplings motivate a hierarchical control architecture. Fast stabilization and protection remain handled by conventional regulatory control (PI loops, anti-surge and operational interlocks), while a supervisory layer performs slower coordination by adjusting a small set of economically and operationally meaningful setpoints (pressures and bypass routing), targeting global feasibility, constraint compliance, and reduced recycle/energy waste.

2. Dynamic simulation model

A dynamic simulator was implemented in Python to enable closed-loop testing and training of the supervisory controller. The model follows a simplified digital-twin philosophy: it preserves the dominant pressure–flow accumulation dynamics, train couplings (shared headers, membrane routing), and actuator nonlinearities needed for control design, while remaining computationally efficient for RL rollouts.

2.1 Compressor trains and stage calculations

Each compression section is modeled as a series of centrifugal compressor stages, each represented by a performance map embedded within a dynamic mass and energy balance. For each stage, an iterative procedure matches volumetric flow and operating point to compute polytropic head, efficiency, power, and discharge temperature.

Composition effects are captured through mixture properties (molecular weight, Mw, and compressibility factor, *Z*), allowing the simulator to represent performance shifts due to varying CO_2 fraction (especially relevant across the membrane split and bypass changes).

2.2 Vessel/header pressure dynamics

Dynamic pressure behavior is represented by vessels/headers that capture the dominant accumulation effects in each section, providing the state evolution used by the control loops and the RL observation vector. (These dynamics are exposed to the RL agent primarily through suction and discharge pressures per train, see Section 4.)

2.3 Valve and recycle-flow modeling

Valve flows are modeled using a compressible-flow relationship, given by Equation (1).

$$\dot{m} = \phi(u) K p_{up} C_v \sqrt{\frac{(p_{up} - p_{down})^+}{p_{up} T Z}} \quad (1)$$

where u is the valve opening, $\phi(u)$ represents the installed valve characteristic (linear, equal-percentage, or quick-opening), p_{up} and p_{down} are upstream and downstream pressures, T is absolute temperature, Z is compressibility factor, C_v is the valve coefficient, and K is a unit-conversion constant. The "$(\bullet)^+$" operator prevents non-physical negative pressure drops.

Anti-surge behavior is represented through recycle flows per train/stage, which are also included as monitored variables and (optionally) penalized in the reward

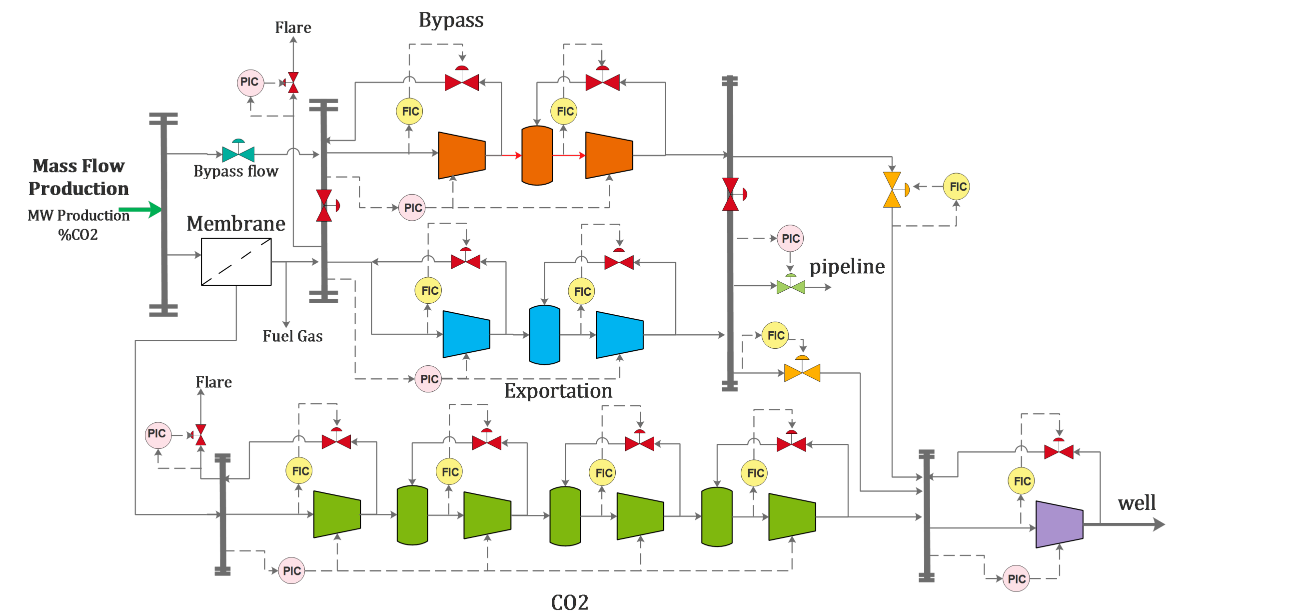

Figure 1. Flowchart of the compression system of the real offshore platform.

to discourage unnecessary recycle activity.

2.4 Membrane CO_2 separation unit

The membrane unit is modeled as a pressure-driven separation process that splits the feed into retentate (treated gas) and permeate (CO_2-rich) streams. An empirical bias term is included to represent non-ideal effects such as aging or performance degradation, while outlet molecular weights are updated based on the resulting gas composition.

The retentate is routed to the export compression section, while the permeate is routed to the CO_2 compression section. The permeate-side pressure is directly linked to the CO_2 compression suction setpoint, explicitly coupling membrane separation performance with downstream compression conditions and supervisory control actions.

3. Disturbance modelling

To assess the robustness and system-level coordination of the supervisory controller, a severe load disturbance is considered. At time t=400s, the production mass flow is subjected to a step increase of 20 kg/s. This scenario imposes an immediate thermal and mechanical load on the compression system, leading to increased discharge pressures and temperatures across all compression trains, including the export, CO_2, and injection sections.

4. Hierarchical control architecture

The proposed control structure is hierarchical and explicitly preserves the regulatory and protection logic of the compression system. The inner layer consists of conventional PI controllers, anti-surge loops at each compression stage, actuator saturation, and split-range logic.

The RL agent operates strictly at the supervisory level and does not bypass or override the inner-loop protections. Instead, it provides bounded adjustments to selected setpoints that are already used by the PI controllers.

4.1 Action space

The RL agent operates in delta (incremental) mode, outputting bounded changes to supervisory setpoints at each control interval. This choice enforces smooth behavior, avoiding abrupt operating-point shifts. The action vector and its numerical bounds are defined in Table 1.

Table 1: Action space configuration.

Action	Description	Lower Bound	Upper Bound	Units
$\Delta \dot{m}_{bypass}$	Bypass flow target	-1	+1	kg/s
$\Delta p^{sp}_{s,exp}$	Export suction pressure setpoint	-500	+500	kPa
$\Delta p^{sp}_{d,exp}$	Export discharge pressure setpoint	-500	+500	kPa
$\Delta p^{sp}_{s,co2}$	CO_2 suction pressure setpoint	-5	+5	kPa
$\Delta p^{sp}_{s,inj}$	Injection suction pressure setpoint	-50	+50	kPa

4.2 Observation vector

The observation vector consists of 28 continuous variables, selected to expose the dominant coupling mechanisms, constraint indicators, and dynamic states relevant for supervisory coordination. Specifically, the environment returns:

- Time t(s);
- Suction, intermediate, and discharge pressures for the bypass, export, CO_2, and injection sections;
- Compressor "rotation proxies" for bypass, export, CO_2, and injection trains;
- Mass flows: bypass, membrane outputs to export and CO_2, and export inlet flow;
- Treated-to-injection valve opening;
- Recycle volumetric flows for bypass and each compression section/stage;
- Discharge temperatures for export stage 2 and CO_2 stage 4;
- Current export capacity setpoint (included as a state variable for consistency).

5. Reward design

A multi-term reward is used to align learning with operational goals: (i) tracking/coordination, (ii) anchoring around desired operating targets, and (iii) safety penalties dominated by discharge temperature constraints. The instantaneous reward is defined by Equation (2).

$$r = r_{track} + r_{temp} + r_{soft} \tag{2}$$

5.1 Tracking

Tracking terms penalize deviations in suction/discharge pressures and export flow, with weights assigned to reflect operational priorities. In particular, CO_2 pressures receive stronger emphasis because they directly affect membrane driving force. An additional target anchoring term penalizes excessive deviation of supervisory setpoints from nominal operating targets.

5.2 Temperature constraints

Thermal safety is encoded explicitly through continuous penalties activated beyond stage-specific discharge temperature limits. The following limits are used:

- $T_{lim,bypass} = 390\,K$, $T_{lim,export} = 410\,K$,
- CO_2 stages: $[400, 440, 420, 420]\,K$ (stages 1-4),
- weights: $w_{temp,co2} = 6.0$, $w_{temp,export} = 2.0$, $w_{temp,bypass} = 2.0$, $w_{temp,injection} = 1.0$.

A soft margin of 5 K is applied before the penalty increases steeply, and temperature exceedances are normalized by a 10 K scale.

5.3 Soft penalties

Additional soft penalties are included to promote operational efficiency and smooth control behavior. Recycle activity is aggregated from recycle flows across sections: bypass, export (stages 1–2), CO_2 (stages 1–4), and injection. Valve usage is represented by the treated-to-injection valve opening, and action smoothness is penalized via the L_1 norm of the action vector.

6. RL algorithm and training procedure

The supervisory policy is trained using PPO [14] with a multilayer perceptron (MLP) policy consisting of two hidden layers of 256 neurons each.

The main hyperparameters are: learning rate 3×10^{-4}, discount factor $\gamma = 0.99$, GAE $\lambda = 0.95$, clipping range 0.2, $n_{steps} = 2048$, $n_{epochs} = 10$, batch size 256, and random seed 42.

Each episode lasts 3000 simulation steps with Δt=0.1 s, corresponding to 300 s of plant operation. Training was performed on a CPU-only workstation (24 logical cores) for a total of 200 episodes ($\approx 6\times10^5$ steps in total).

During training, the best-performing policy is selected based on the rolling mean of episode rewards and saved for post-training evaluation. After training, the selected policy is evaluated in closed-loop simulations using deterministic action selection.

7. NMPC benchmark controller

As a benchmark, a stochastic nonlinear model predictive control (NMPC) strategy was implemented using a random-shooting approach inspired by Model Predictive Path Integral (MPPI) control [16].

The same supervisory action space and objective function used by the RL controller was adopted. The final configuration used a prediction horizon of 8 steps and a control horizon of 2 steps, with 16 candidate sequences evaluated every 6 simulation steps, and 4 elite samples used for action selection.

RESULTS AND DISCUSSION

8 Training convergence

The PPO-based supervisory policy converged within approximately 80–100 episodes (Figure 2). With an average simulation throughput of 129 steps/s, the complete training required approximately 77.5 minutes.

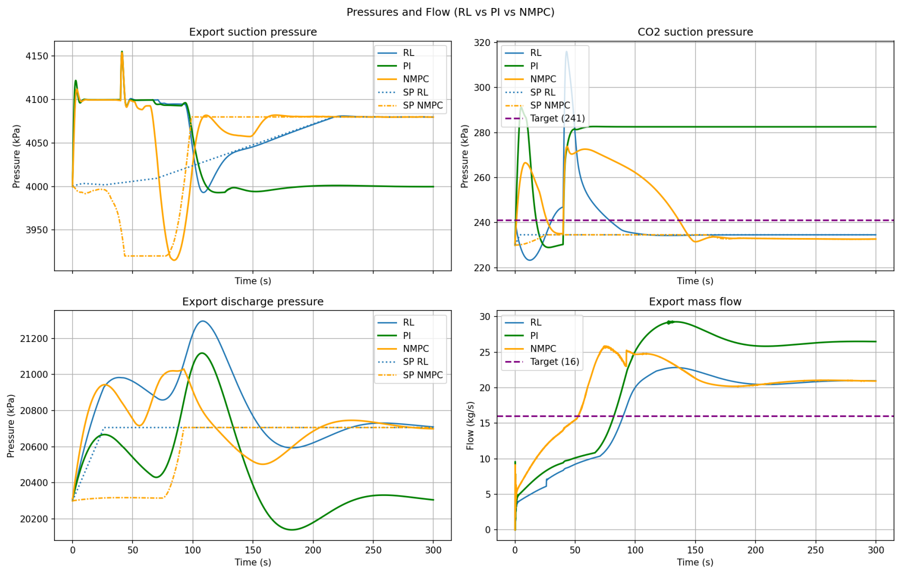

Figure 3. Closed-loop pressures and export flow under PI, RL, and NMPC control.

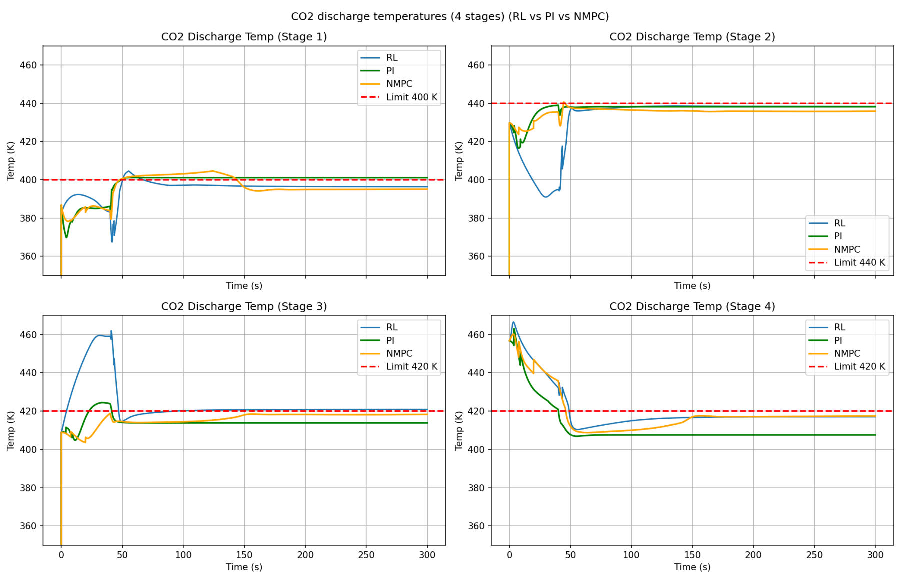

Figure 4. CO_2 compressor discharge temperatures (stages 1–4) under PI, RL, and NMPC control; dashed lines indicate temperature limits.

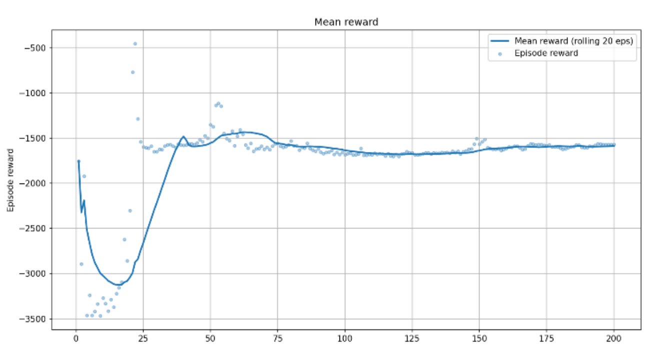

Figure 2. Training mean reward versus episodes.

9 Closed-loop control

Figure 3 shows that RL and NMPC converge to similar steady states, both outperforming PI. Their main difference is in the transient response: NMPC yields smoother but slower corrections, while RL applies more aggressive initial actions before stabilizing.

Figure 4 highlights that RL produces larger temperature excursions, particularly at CO_2 stage 3, where constraint violations are more frequent. These results indicate that RL's aggressive transient actions improve coordination but weaken constraint satisfaction.

10 Quantitative performance comparison

Table 2 compares RL, PI, and NMPC in terms of tracking, energy, safety, control effort, and computational cost.

Table 2: Controllers' performances.

Metric	RL	PI	NMPC
IAE	322, 462	699, 682	310, 400
ITAE	48.94×10^6	111.48×10^6	46.54×10^6
Mean power (HP)	25, 979	25, 090	24, 212
Total abs. power change (HP)	28, 241	30, 572	31, 115
Temp. violation total count	5, 701	3, 178	3, 053
Temp. violation total magnitude	43, 252	35, 423	30, 931
$\int u^2\, dt$	1202.20	0.00	2765.63
Evaluation time (s)	71.04	65.28	3895.77

In terms of tracking, RL and NMPC achieve comparable performance and clearly outperform PI, with similar IAE and ITAE values. From an energy perspective, NMPC achieves the lowest mean power, while RL operates at a slightly higher level. However, RL exhibits the lowest total absolute power variation, indicating less aggressive load changes despite similar steady-state performance.

Regarding safety, RL shows the highest temperature-violation count and magnitude, confirming that the current reward-based formulation allows stronger transient excursions. In terms of control effort, RL is significantly smoother than NMPC, with much lower integral quadratic control effort ($\int u^2 dt$), indicating less aggressive supervisory actions.

Finally, computational cost highlights a key advantage of RL. While RL and PI have similar evaluation times, NMPC is more than one order of magnitude slower.

11 Limitations and scope of applicability

The present work is based on a simplified digital twin and does not account for all sources of uncertainty present in real offshore systems, such as sensor noise, actuator degradation, and long-term equipment aging. The evaluation was conducted under a limited set of disturbance scenarios. Although the RL policy generalizes well within this range, its robustness to broader operating conditions remains to be assessed.

Safety is handled through soft penalties, which leads to transient constraint violations. Practical deployment would require additional mechanisms such as hard constraint enforcement or action shielding. Finally, the study considers a single-agent supervisory policy. Extensions to multi-agent formulations, economic objectives, and online adaptation remain open research directions.

CONCLUSIONS

This work showed the feasibility and benefits of integrating reinforcement learning as a supervisory layer in a multistage offshore gas compression system, improving plant-wide coordination while preserving conventional PI control and safety logic.

The RL-supervised strategy significantly outperforms PI, reducing IAE and ITAE by approximately 54% and 56%, respectively. Its performance is also close to NMPC, with differences of only about 4–5%, and both approaches converge to similar steady-state operating points.

From an operational perspective, RL provides smoother actuation, as the quadratic control effort is reduced by approximately 57% compared to NMPC. In terms of energy, NMPC achieves the lowest mean power consumption, while RL operates about 7% above NMPC and slightly above PI.

A key limitation of the current approach is safety performance. RL exhibits higher temperature violation frequency and magnitude, with approximately 80–90% more violation events and about 40% higher total violation magnitude compared to NMPC. These results reflect the use of soft penalties in the reward function, which allow stronger transient excursions (particularly during early adaptation phases) before converging to safer operation.

Finally, computational cost strongly favors RL. While NMPC requires approximately 55× more evaluation time, the RL policy can be executed with negligible online overhead. Combined with a moderate offline training time (≈77.5 minutes on CPU), this makes RL a computationally efficient alternative for real-time supervisory control.

Future work will focus on strengthening safety guarantees through hard constraint enforcement, extending evaluation across multiple disturbance scenarios, and assessing robustness and transferability toward real offshore operating conditions.

ACKNOWLEDGEMENTS

Jose Rodrigues Torraca Neto reports financial support provided by CAPES, Finance Code 001, and Petrobras (Cooperation term 0050.0124501.23.9). Argimiro R. Secchi Argimiro R. Secchi is grateful to financial support from CNPq (Grants No. 300744/2025-0). Other authors declare that they have no known competing

financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR IDENTIFIERS

Author ORCIDs:
Torraca JR Neto: 0009-0000-4127-5682
Giraldo SAC: 0000-0001-6253-564X
Campos MCMM: 0000-0002-5746-6915
Caldas GLR: 0000-0003-2554-9561
Capron BDO: 0000-0001-5497-9769
Secchi AR: 0000-0001-7297-3571

REFERENCES

1. Mokhatab S, Poe WA, Mak JY. Phase separation. Handbook of Natural Gas Transmission and Processing :137-167 (2015). https://doi.org/10.1016/b978-0-12-801499-8.00004-3
2. Jones K, Cortinovis A, Mercangoez M, Ferreau HJ. Distributed model predictive control of centrifugal compressor systems. IFAC-PapersOnLine 50:10796-10801 (2017). https://doi.org/10.1016/j.ifacol.2017.08.2343
3. Plucenio A, Normey-Rico J, Vettorazzo C, Campos MM, De Lima ML. MPC advanced control of an offshore gas compression system. OTC Brasil : (2017). https://doi.org/10.4043/28129-ms
4. Gravdahl JT, Willems F, de Jager B, Egeland O. Modeling for surge control of centrifugal compressors: comparison with experiment. Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No.00CH37187) 2:1341-1346 (None). https://doi.org/10.1109/cdc.2000.912043
5. Ribeiro CHP, Miyoshi SC, Secchi AR, Bhaya A. Model predictive control with quality requirements on petroleum production platforms. Journal of Petroleum Science and Engineering 137:10-21 (2016). https://doi.org/10.1016/j.petrol.2015.11.004
6. Giraldo SAC, Supelano RC, d'Avila TC, Capron BDO, Ribeiro LD, Campos MM, Secchi AR. Model predictive control with dead-time compensation applied to a gas compression system. Journal of Petroleum Science and Engineering 203:108580 (2021). https://doi.org/10.1016/j.petrol.2021.108580
7. International Energy Agency. CO_2 emissions in 2023. https://www.iea.org/reports/co2-emissions-in-2023
8. United Nations Framework Convention on Climate Change. Paris agreement. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement
9. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed. MIT Press (2018).
10. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015).
11. Devarakonda VS, Sun W, Tang X, Tian Y. Recent advances in reinforcement learning for chemical process control. Processes 13:1791 (2025). https://doi.org/10.3390/pr13061791
12. García J, Fernández F. A comprehensive survey on safe reinforcement learning. J Mach Learn Res 16:1437–1480 (2015).
13. Qin SJ, Badgwell TA. A survey of industrial model predictive control technology. Control Engineering Practice 11:733-764 (2003). https://doi.org/10.1016/s0967-0661(02)00186-7
14. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017).
15. Ng AY, Harada D, Russell SJ. Policy invariance under reward transformations: theory and application to reward shaping. In: Proceedings of the Sixteenth International Conference on Machine Learning. Morgan Kaufmann Publishers, San Francisco, CA:278–287 (1999).
16. Williams G, Aldrich A, Theodorou EA. Model predictive path integral control: from theory to parallel computation. Journal of Guidance, Control, and Dynamics 40:344-357 (2017). https://doi.org/10.2514/1.g001921