

Digital Twin-Driven TD3-CPER Control for Continuous Casting

Quanhui Qiu^{1, 2, a, #}, Sen Wang^{3, b, #}, Jiacheng Zheng^{1, 2} and Dejun Ning^{1, c *}

¹ Shanghai Advanced Research Institute, Chinese Academy of Sciences, 201210, Shanghai, China;

² University of Chinese Academy of Sciences, 100049, Beijing, China;

³ Shanghai Baoxin Software Co., Ltd, 201999, Shanghai, China

^a qiuqh@sari.ac.cn, ^b wangsen@baosight.com, ^c ningdj@sari.ac.cn,

[#] These authors contributed equally to this work

Abstract. Continuous casting is a key process in intelligent steel manufacturing, and mold level control directly affects slab quality and production stability. However, strong nonlinearity, multivariable coupling, time delays, and operating disturbances make real-time control optimization challenging for conventional control strategies. To address this problem, this paper proposes a digital twin-driven reinforcement learning control method based on Twin Delayed Deep Deterministic Policy Gradient with Composite Prioritized Experience Replay, termed TD3-CPER. A data-driven digital twin environment is constructed using real production data to support closed-loop policy training and evaluation. In TD3-CPER, the replay mechanism integrates temporal-difference error, reward feedback, replay frequency, and K-Medoids clustering to improve sample utilization and training stability. Simulation results show that TD3-CPER increases the number of Grade I slabs by 56.2% compared with manual control and achieves a 14.7% higher average cumulative reward than the strongest baseline reinforcement learning method. These results suggest that TD3-CPER is a feasible approach for quality-oriented control optimization in continuous casting.

Keywords: Digital twin; reinforcement learning; TD3-CPER; continuous casting; intelligent manufacturing; process control.

1. Introduction

Continuous casting connects steelmaking and rolling in modern steel production, and its stability directly affects slab quality, production continuity, and manufacturing efficiency [1][2]. Mold level fluctuation is closely associated with surface defects, inclusion formation, breakout risk, and casting instability [3]. Accurate mold level control is therefore essential for intelligent steel manufacturing.

However, mold level control remains challenging because the continuous casting process involves strong nonlinearity, multivariable coupling, time-varying dynamics, flow disturbances, and strict safety constraints. The mold level is affected by stopper rod position, casting speed, molten steel flow, thermal conditions, and withdrawal behavior. Although artificial intelligence and machine learning have been applied to process optimization in continuous casting and steel manufacturing [4], manual operation and fixed-parameter control strategies still suffer from limited adaptability, delayed response, and insufficient utilization of industrial data.

Digital twin technology provides a virtual representation of physical production systems and supports simulation, monitoring, prediction, and control validation [5][6]. Data-driven digital twins have been investigated for manufacturing quality prediction and coordinated control in continuous casting processes [7][8]. Meanwhile, deep learning and reinforcement learning have been explored for continuous casting prediction, automatic casting control, and decision support [9][10]. However, the application of deep reinforcement learning to continuous casting still faces several challenges, including training instability, limited sample efficiency, and safe policy evaluation.

To address these issues, this paper proposes a digital twin-driven TD3-CPER control framework for continuous casting mold level control. The main contributions are as follows:

- A data-driven digital twin simulation environment is constructed using real continuous casting production data to support closed-loop policy training and evaluation.

- A Composite Prioritized Experience Replay mechanism is integrated into TD3 to improve sample utilization and policy learning stability.
- Digital twin-based simulation results demonstrate improved slab quality and higher cumulative reward compared with manual control and representative reinforcement learning baselines.

2. Digital Twin-Based Continuous Casting Control Framework

The proposed framework optimizes mold level control through closed-loop interaction between a reinforcement learning agent and a data-driven digital twin environment. The physical continuous casting process involves the tundish, mold, stopper rod, withdrawal system, and external disturbances. The mold level is mainly regulated by stopper rod position and casting speed, while it is also affected by coupled variables such as molten steel flow, mold width, argon flow, and cooling-related parameters.

A data-driven digital twin model is constructed using real industrial production data. In the original dataset, 122 process variables were collected from the continuous casting production line. A random forest-based feature importance analysis was adopted to select 30 key variables highly correlated with mold level fluctuations. Based on these variables, a multilayer perceptron model was trained to predict the dynamic mold level. The digital twin model achieved an R^2 of 0.8006, a normalized MSE of 8.36×10^{-4} , and an MAE of 0.3242 mm on the test set, providing a feasible simulation environment for reinforcement learning-based policy training and evaluation.

As shown in Fig. 1, the reinforcement learning agent observes the mold level state from the digital twin environment and outputs continuous control actions, including stopper rod adjustment and casting speed regulation. The digital twin then generates the next state and reward according to the simulated process response. This closed-loop structure enables policy training and simulation-based validation before potential industrial implementation.

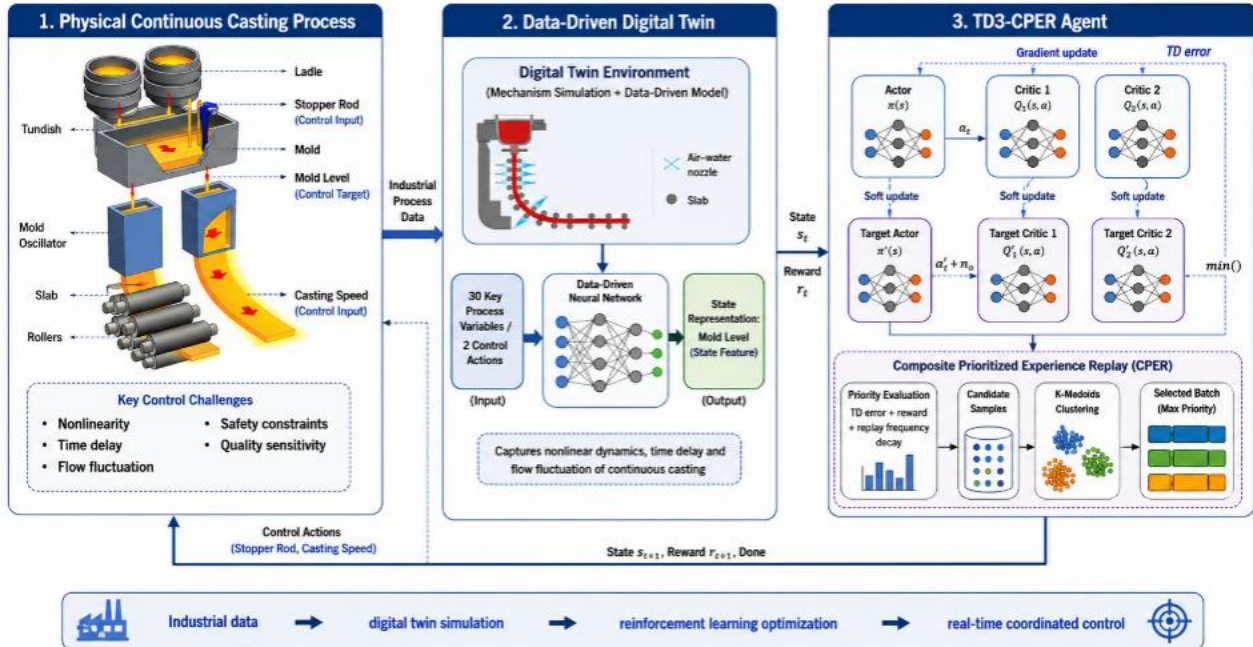


Fig. 1 Framework of the digital twin-driven TD3-CPER control method.

The control problem is formulated as a Markov decision process (MDP). To reduce model complexity and improve real-time feasibility, the state is represented by the mold level:

$$S_t = [L_{mold}(t)]$$

where $L_{mold}(t)$ denotes the mold level at time t . The action space includes two adjustable control variables:

$$A_t = [H_{stopper}(t), V_{casting}(t)]$$

where $H_{stopper}(t)$ is the stopper rod position and $V_{casting}(t)$ is the casting speed. The reward function jointly considers mold level stability, casting speed stability, and abnormal mold level events:

$$R_t = R_v(t) + R_h(t) + R_l(t)$$

where $R_v(t)$ denotes the reward related to casting speed stability, $R_h(t)$ denotes the reward related to mold level stability, and $R_l(t)$ denotes the penalty term for abnormal mold level events.

3. TD3-CPER Reinforcement Learning Method

The proposed TD3-CPER method adopts TD3 as the continuous-control backbone and improves the experience replay process for digital twin-based continuous casting control. Since stopper rod adjustment and casting speed regulation are continuous control actions, TD3 is used to reduce overestimation bias through twin critic networks, target policy smoothing, and delayed policy updates [11][12].

For a sampled transition $(s_i, a_i, r_i, s_{i+1}, d_i)$, the target Q-value is calculated as:

$$y_i = r_i + \gamma(1 - d_i) \min_{j=1,2} Q_{\theta_j'}(s_{i+1}, \tilde{a}_{i+1})$$

where γ is the discount factor, d_i is the terminal flag, $Q_{\theta_j'}$ denotes the j-th target critic network, and $\tilde{a}_{i+1}$ is the smoothed target action. Instead of modifying the original TD3 update structure, this study focuses on selecting valuable and diverse samples from the replay buffer.

Conventional Prioritized Experience Replay samples experiences mainly according to temporal-difference (TD) error [13][14]. Although this mechanism improves the utilization of high-error samples, it may ignore reward feedback and repeatedly reuse a small number of high-priority samples. In continuous casting, high-reward samples often correspond to stable mold level regulation, while excessive replay of similar samples may reduce sample diversity and weaken policy generalization. Therefore, Composite Prioritized Experience Replay is introduced.

First, CPER defines a composite priority by combining TD error and reward:

$$\Gamma_i = |\delta_i| + \lambda \sigma(r_i)$$

where Γ_i is the composite priority of the i -th experience, δ_i is the TD error, r_i is the immediate reward, λ is the reward weight, and $\sigma(\cdot)$ is the sigmoid function. This design prevents negative priorities and emphasizes both learning value and control quality.

Second, CPER introduces a replay-frequency decay term to avoid excessive reuse of specific samples:

$$p_i = \mu^{C(i)} \Gamma_i + \varepsilon$$

where p_i is the final priority, $C(i)$ is the replay count of the i -th experience, μ is the decay factor, and ε is a small constant used to avoid zero priority. This mechanism reduces the priority of repeatedly used samples and encourages learning from fresher or less-replayed experiences [15]. The sampling probability is calculated as:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}$$

where α controls the strength of prioritization.

Third, CPER uses K-Medoids clustering to improve sample diversity. A candidate set is first selected from the replay buffer according to composite priority. Then, K-Medoids clustering divides the candidate samples into several clusters, and representative high-priority samples are selected from different clusters to form the final mini-batch. This reduces redundant sampling and mitigates the diversity limitation of conventional PER [16]. Since priority-based non-uniform sampling may introduce bias, importance sampling weights are used:

$$w_i = \frac{(NP(i))^{-\beta}}{\max_j (NP(j))^{-\beta}}$$

where N is the replay buffer size and β controls the correction strength.

Overall, CPER combines composite priority evaluation, replay-frequency adjustment, and clustering-based sample selection. This replay strategy improves sample utilization and training stability for continuous casting mold level control.

4. Experimental Design and Results

4.1 Experimental Setup

The proposed method is evaluated in a digital twin-based simulation environment constructed from real continuous casting production data. The data were collected from a continuous caster in a steel plant on May 29, 2024. Process variables, including casting speed, mold level, and cooling water flow, were recorded at a sampling frequency of 1 Hz, yielding 43,200 samples over a 12-hour production period. The first 10 hours, containing 36,000 samples, were used for training, and the remaining 2 hours, containing 7,200 samples, were used for testing.

TD3-CPER is compared with actual manual control and three baseline deep reinforcement learning algorithms: DDPG-PER, TD3-PER, and SAC-PER. Algorithm performance is measured by the average cumulative reward and peak reward over five independent runs. Process control performance is evaluated using slab quality grading metrics based on mold level fluctuation and casting speed fluctuation.

4.2 Mold Level Response and Quality Comparison

As shown in Fig. 2, TD3-CPER produces smoother mold level responses and fewer abnormal deviations in the digital twin environment. This indicates that the method can coordinate stopper rod adjustment and casting speed regulation for more stable mold level control.

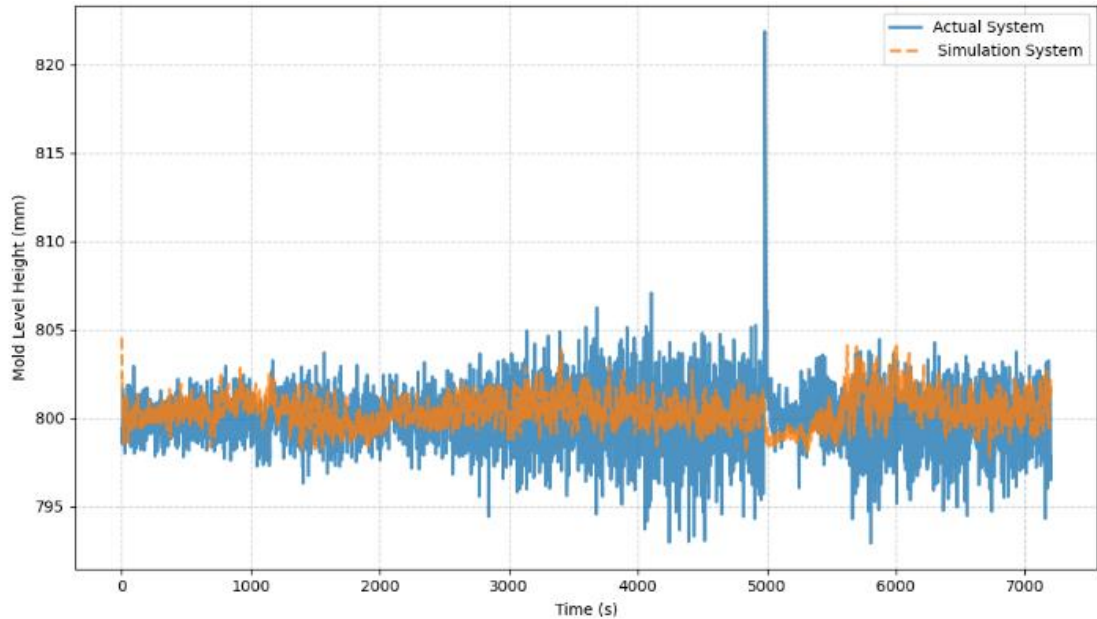


Fig. 2 Comparison of mold level responses between actual production and TD3-CPER simulation.

The quantitative quality comparison is reported in Table 1. The Grade I rate of casting speed fluctuation reaches 100.00% in TD3-CPER simulation, compared with 96.31% in actual production. The Grade I rate of mold level fluctuation reaches 98.63%, compared with 87.10% in actual production. Under the integrated slab quality evaluation, the number of Grade I slabs increases from 185 in actual production to 289 in TD3-CPER simulation, corresponding to a 56.2% increase in the Grade I slab count.

Table 1. Quality comparison between actual production and TD3-CPER simulation

Metric	Actual production	TD3-CPER simulation
--------	-------------------	---------------------

Grade I rate of casting speed fluctuation	96.31%	100.00%
Grade I rate of mold level fluctuation	87.10%	98.63%
Number of Grade I slabs	185	289
Number of Grade II slabs	23	4

4.3 Training Reward Comparison

To evaluate learning performance, TD3-CPER is compared with DDPG-PER, TD3-PER, and SAC-PER under the same simulation settings. The training reward curves are shown in Fig. 3. TD3-CPER achieves faster convergence and more stable reward improvement than DDPG-PER and TD3-PER. Compared with SAC-PER, it obtains a higher average cumulative reward with lower training variance.

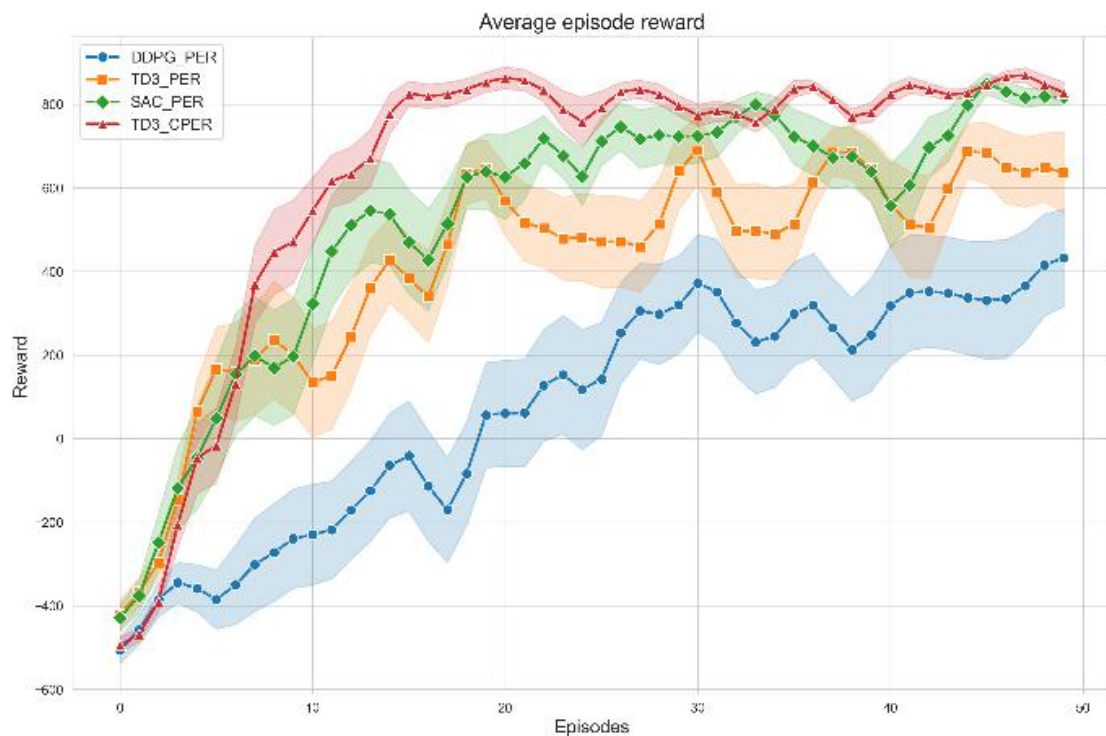


Fig. 3 Training reward curves of TD3-CPER and baseline DRL algorithms.

The final training results show that TD3-CPER achieves an average cumulative reward of 844.6 ± 36.5 and an average peak reward of 973.6 ± 45.0 . In comparison, DDPG-PER, TD3-PER, and SAC-PER achieve average cumulative rewards of 363.8 ± 594.0 , 598.2 ± 326.2 , and 736.5 ± 170.8 , respectively. Compared with the strongest baseline, SAC-PER, TD3-CPER improves the average cumulative reward by 14.7%. This suggests that CPER enhances experience utilization and policy learning stability.

4.4 Effectiveness of TD3-CPER Components

To analyze the contribution of different CPER components, several algorithm variants are compared in Table 2. Removing the reward-feedback component reduces the average cumulative reward from 844.6 to 746.8. Removing replay-frequency adjustment further decreases the reward to 673.9, indicating that replay-frequency adjustment helps avoid excessive reuse of specific samples. When clustering-based sample selection is removed, the average cumulative reward decreases to 693.8, suggesting that sample diversity contributes to training stability.

Table 2. Performance comparison of TD3-CPER variants.

Algorithm variant	Average cumulative reward	Average peak reward
Complete TD3-CPER	844.6 ± 36.5	973.6 ± 45.0
Reward ablated	746.8 ± 185.8	994.4 ± 4.4

Replay frequency ablated	673.9 ± 141.2	948.4 ± 58.8
Clustering ablated	693.8 ± 161.9	969.6 ± 45.9

These results show that the components of TD3-CPER work cooperatively. Reward feedback guides priority evaluation toward high-quality control behavior, replay-frequency adjustment improves sample utilization, and clustering-based selection enhances sample diversity.

5. Conclusion

This paper proposes a digital twin-driven TD3-CPER reinforcement learning control method for continuous casting mold level control. By integrating a data-driven digital twin environment with TD3 and Composite Prioritized Experience Replay, the proposed method enables closed-loop policy training and evaluation for intelligent manufacturing-oriented process control.

Simulation results show that TD3-CPER increases the number of Grade I slabs by 56.2% compared with actual manual control. It also achieves a 14.7% higher average cumulative reward than the strongest baseline reinforcement learning method. These results indicate that the proposed replay mechanism improves both quality-oriented control performance and learning stability.

This study is still limited to simulation-based validation using real production data. The proposed method has not yet been deployed in a real-time industrial closed-loop production line. Future work will focus on safe field validation, safety-constrained reinforcement learning, online adaptation, and generalization testing under multiple casting conditions.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 52335001).

References

- [1] S. Louhenkilpi, "Continuous casting of steel," in *Treatise on Process Metallurgy: Volume 3: Industrial Processes*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2024, pp. 343–383.
- [2] B. G. Thomas, "Review on modeling and simulation of continuous casting," *Steel Research International*, vol. 89, no. 1, Art. no. 1700312, 2018.
- [3] J. K. Brimacombe, "The challenge of quality in continuous casting processes," *Metallurgical and Materials Transactions A*, vol. 30, pp. 1899–1912, 1999.
- [4] D. Cemernek, S. Cemernek, H. Gursch, A. Pandeshwar, T. Leitner, M. Berger, and R. Kern, "Machine learning in continuous casting of steel: A state-of-the-art survey," *Journal of Intelligent Manufacturing*, vol. 33, no. 6, pp. 1561–1579, 2022.
- [5] D. M. Botín-Sanabria et al., "Digital twin technology challenges and applications: A comprehensive review," *Remote Sensing*, vol. 14, no. 6, Art. no. 1335, 2022.
- [6] W. Hu, T. Zhang, X. Deng, Z. Liu, and J. Tan, "Digital twin: A state-of-the-art review of its enabling technologies, applications and challenges," *Journal of Intelligent Manufacturing and Special Equipment*, vol. 2, no. 1, pp. 1–34, 2021.
- [7] D. Liu, Y. Du, W. Chai, C. Lu, and M. Cong, "Digital twin and data-driven quality prediction of complex die-casting manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 8119–8128, 2022.
- [8] J. Yang, Z. Ji, W. Liu, and Z. Xie, "Digital-twin-based coordinated optimal control for steel continuous casting process," *Metals*, vol. 13, no. 4, Art. no. 816, 2023.
- [9] Z. Lu et al., "Real-time prediction and adaptive adjustment of continuous casting based on deep learning," *Communications Engineering*, vol. 2, no. 1, Art. no. 34, 2023.
- [10] X. Wu et al., "Automatic casting control method of continuous casting based on improved soft actor-critic algorithm," *Metals*, vol. 13, no. 4, Art. no. 820, 2023.

- [11] H. Tan, “Reinforcement learning with deep deterministic policy gradient,” in Proc. International Conference on Artificial Intelligence, Big Data and Algorithms, 2021, pp. 82–85.
- [12] S. Fujimoto, H. van Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in Proc. International Conference on Machine Learning, 2018, pp. 1587–1596.
- [13] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, “Prioritized experience replay,” arXiv:1511.05952, 2015.
- [14] Y. Hou, L. Liu, Q. Wei, X. Xu, and C. Chen, “A novel DDPG method with prioritized experience replay,” in Proc. IEEE International Conference on Systems, Man, and Cybernetics, 2017, pp. 316–321.
- [15] J. Ma, D. Ning, C. Zhang, and S. Liu, “Fresher experience plays a more important role in prioritized experience replay,” *Applied Sciences*, vol. 12, no. 23, Art. no. 12489, 2022.
- [16] Y. Pan et al., “Understanding and mitigating the limitations of prioritized experience replay,” in Proc. Uncertainty in Artificial Intelligence, vol. 180, 2022, pp. 1561–1571.