Fault-Aware Hybrid MPC–Reinforcement Learning Framework for Adaptive Sensor Bias Handling in Industrial Furnace Temperature Control

Bibi Amna [*], Muhammad Zeeshan[†], Kiran Raheel[‡]

Abstract

Temperature control in industrial furnaces is an important issue in view of product quality and efficiency. In the presence of sensor bias faults, the performance of the controller declines due to the use of incorrect feedback data. This paper presents a hybrid system of Model Predictive Control (MPC) combined with Reinforcement Learning (RL) for the detection and compensation of sensor bias faults in the control system of the furnace temperature. It uses MPC to achieve stable tracking performance as well as the creation of residuals, which indicate sensor abnormalities. It also includes a supervisory Deep Q-Network (DQN) agent, which chooses the best tuning strategy of MPC from the set of available options on the basis of tracking and residuals. The proposed methodology was tested on a simulated system in MATLAB/Simulink with the use of constant, periodic, and multilevel reference temperature profiles with both negative and positive sensor bias. The training procedure took 57 episodes in the case of negative bias and 77 episodes in the absence of any faults when a constant reference is provided. Other types of the reference temperature profile were more difficult to be learned by the DQN agent. They took 120 episodes without achieving convergence.

***Keywords*:** Model Predictive Control, Reinforcement Learning, Sensor Bias Fault, Industrial Furnace, Fault-Tolerant Control, Deep Q-Network, Temperature Regulation.

Introduction

Furnaces have gained considerable importance in industries like metalworking, thermal treatment, chemicals production, and materials processing (Qin & Badgwell, 2003; Zhang et al., 2014). In all of these industries, accurate temperature control does not only serve the purpose of performing optimally but also is of high importance from the perspective of quality assurance, energy efficiency, and safety issues (Camacho & Alba, 2023; Tian et al., 2024). With advancements in automation, the demands on temperature control systems have increased immensely. Now, temperature control systems must be intelligent, adaptive, and robust

[*]Department of Electrical Engineering, CECOS University of IT & Emerging Technology, Peshawar 25000, Pakistan, bibi.amna.msee-s2024a@cecosian.edu.pk

[†]Department of Electrical & Electronics Engineering, Karadeniz Technical University, Trabzon 61080, Türkiye, 449922@ogr.ktu.edu.tr

[‡]Corresponding Author: Department of Electrical Engineering, CECOS University of IT & Emerging Technology, Peshawar 25000, Pakistan, kiran@cecos.edu.pk

enough to perform under sub-optimal operating conditions (Mesbah, 2016; Hewing et al., 2020).

Bias is one of the most common problems in furnace temperature control systems (Han et al., 2023; El-Mahdy et al., 2024). Unlike a complete sensor failure, bias in sensors problem, bias in sensors is characterized by an offset value in the readings. The control system continues running, and the system looks normal, but, at the same time, the furnace runs below or above the set point. What makes bias hard to recognize is that such problems usually cause no alarm signals. Eventually, such problems result in errors and can lead to serious consequences (Du et al., 2024).

Model Predictive Control (MPC) is widely used in industrial applications due to its capability to predict future system behavior, optimize control actions over a finite horizon, and manage multiple constraints simultaneously (Qin & Badgwell, 2003; Rawlings et al., 2020). The technique is particularly important in furnace temperature control due to the structure and reliability that it provides (Zhang et al., 2014). Additionally, it creates residuals as a result of differences between prediction and measurement.

A distinct strength lies in Reinforcement Learning (RL) (Sutton & Barto, 2018; Mnih et al., 2015). In contrast to the approach that depends on pre-defined parameters and a static model, an RL agent gains knowledge based on its interactions with the environment. The process starts when the agent perceives its current state, takes an action, receives a reward, and develops an increasingly effective policy based on the information gathered throughout the learning process (Lillicrap et al., 2016).

The combination of MPC and RL is increasingly gaining popularity for the reasons discussed above (Bertsekas, 2024; Gros & Zanon, 2020). MPC adds structure, consistency, and constraint-aware control, while RL adds flexibility and learning. In tackling the issue of sensor bias faults in industrial furnaces, the fusion of MPC and RL will be able to identify sensor malfunction, assess its magnitude, and adjust the control response without the need for manual adjustments (Jiang et al., 2023; Hosseini & Salahshoor, 2024).

Literature on intelligent fault detection and adaptive control exists, but it mainly concentrates on generalized fault detection, actuator fault detection, or learning-based approaches alone (Leite et al., 2025; Liu et al., 2025). Prior MPC-RL hybrid approaches have concentrated on either actuator fault tolerance, autonomous system applications, or general process control, while problems of persistent sensor bias faults in temperature control in industrial furnaces have not been studied before. In

addition, prior research does not take advantage of fault-related residual information in the framework of reinforcement learning to allow adaptive supervisor-based switching between several MPC controllers. Thus, an adaptive fault-tolerant approach is needed that integrates the benefits of MPC and RL to compensate for sensor biases in industrial furnaces. The main contributions of this research work are as follows:

1. The paper represents a novel fault-aware hybrid MPC-RL approach where the role of a Deep Q-Network (DQN) agent is to supervise the selection of multiple MPC controllers instead of controlling the system directly.
2. Residuals and tracking errors are used to incorporate fault-related information into RL observations, making it possible for the algorithm to adaptively choose appropriate controllers when sensor bias is present.
3. This paper analyzes the performance of the proposed approach for constant, periodic, and multi-level temperature reference trajectories under negative, zero, and positive sensor bias conditions.
4. Analysis of how sensor bias sign and reference trajectory type affect the convergence of the learning process is presented.

The rest of this paper is organized as follows. First, the literature review is presented, followed by the methodology, including system modeling, fault injection, controller design, and DQN agent design. Next, the simulation results and their analysis are presented. This is followed by a discussion of the results in relation to the research questions and the existing literature. Finally, the paper concludes with a summary of the main findings.

Literature Review

Fault detection and diagnosis are currently important themes within advanced industrial technologies. In light of Industry 4.0, intelligent fault diagnostics are increasingly recognized as a key part of automated manufacturing processes. Leite et al. (2025) considered machine learning techniques for fault diagnostics in the context of Industry 4.0, whereas Liu et al. (2025) and Seid Ahmed et al. (2025) pointed to the growing importance of data-driven monitoring and intelligent diagnostics in automation technology.

Past research mainly concentrated on sudden failure cases, lost signals, and general fault classifications. However, gradual faults such as sensor bias, which develop continuously over time, have been comparatively less discussed even though they are quite practical for thermal systems (Amin et al., 2024). Sensor faults significantly affect feedback control loops since control actions are directly dependent upon

measured signals. Bias faults are more problematic because they provide deceptive stability conditions that may result in process temperatures being different from expected.

El-Mahdy et al. (2024) developed a fault-tolerant control technique utilizing deep learning to manage faulty sensors in manufacturing processes. On the other hand, Escobar- Jiménez et al. (2023) focused on sensor fault-tolerant control in internal combustion engines, and Han et al. (2023) conducted research into sensor fault-tolerant control in heat-exchanger reactors. Additionally, AlRijeb et al. (2025) have researched soft sensors in industry through machine learning methods. Even though the above papers show that sensor faults can be tolerated, no research exists on the application of hybrid MPC-RL to temperature control in industrial furnaces. MPC has emerged as one of the most commonly used industrial control techniques because of its predictive nature, ability to handle constraints, and optimization formulation (Camacho & Alba, 2023; Rawlings et al., 2020).

The use of state-space MPC was demonstrated by Zhang et al. (2014) to be used for temperature control in coke furnaces, whereas MPC-based optimization was proposed by Tian et al. (2024) for municipal solid waste incineration furnaces. Some of the recent research works have considered MPC techniques enhanced by machine learning. Data-driven MPC methods have been analyzed by Morato & Felix (2024) and Dai et al. (2024) combined deep RL with MPC.

As was shown by Soza Mamani & Prado Romo (2025), coupling nonlinear MPC with deep RL algorithms can enhance robustness in thermal processes with long delays. However, the authors did not specifically address to sensor bias compensation. Another significant benefit of using MPC includes residual generation. Indeed, residuals may help in fault monitoring because comparison of residual signals with measured outputs will show whether there are any faults in the process (Bagla et al., 2023; Du et al., 2024). In recent decades, RL has received considerable attention as a method of designing control systems, which makes it possible to obtain control policies based on the interaction of the agent with the environment without using process models. According to Nievas et al. (2024), besides the benefits of flexibility and nonlinear decision making, there are difficulties associated with safety concerns and slow initial learning.

Furnace control systems can solve these issues in case of using RL as a supervisory level over an already existing stable MPC controller. Hosseini & Salahshoor (2024) showed that RL employing double Q-learning can enhance the performance of fault-tolerant control in cases of sensor faults. In addition, the application of RL to intelligent industrial

processes and optimization and control has been highlighted by the works of Li & Qiu (2024), Ogunseye et al. (2023), and Razmi (2025). The combination of MPC and RL techniques has been studied extensively in recent times. For example, Jiang et al. (2023) presented an approach to adaptive fault-tolerant MPC-RL for quadcopter actuator faults.

A theoretical study on the integration of MPC and RL via dynamic programming was carried out by Bertsekas (2024). On the other hand, Gros & Zanon (2020) have focused their research on the use of RL in non-linear economic MPC systems, and Federici et al. (2025) utilized an MPC system with RL in autonomous landing applications. However, there has been little progress in the development of MPC-RL models to detect and compensate for sensor biases in temperature control systems of industrial furnaces. Hybrid MPC/RL modeling techniques have found applications in other areas such as greenhouse climate control (Mallick et al., 2025), power electronics control (Wan et al., 2025), and intelligent robot manufacturing systems (Schaal et al., 2019; Saeed et al., 2025).

According to literature sources, intelligent fault diagnosis is gaining importance in industrial automation systems. Literature review reveals that RL enables adaptability whereas MPC ensures stability and structured control in the system. Yet, a safe and fault-tolerant hybrid approach needs to be developed (Ham & Ah, 2025; Hewing et al., 2020).

Previous research studies have examined MPC and RL techniques along with fault tolerant approaches individually but research on sensor bias compensation in industrial furnaces through a hybrid MPC and RL approach is inadequate. Thus, this paper aims at bridging this knowledge gap by developing a hybrid MPC-RL framework with fault detection and bias estimation abilities.

Methodology

Research Design

Experimental simulation approach was used in the study. The experiment was carried out through simulation done on MATLAB/Simulink. In the research, modeling and simulation were conducted in MATLAB/Simulink without actually building the furnace system. The advantage of this methodology is that faults can be introduced safely and systematically to observe how the system behaves with and without faults. The main comparison in this study was between the normal MPC control and the suggested hybrid MPC-RL system with variations in the reference temperature and sensor bias.

Mathematical Modeling of the Furnace System

The furnace was considered a first-order dynamic system, an appropriate assumption for industrial heating systems that exhibit dominant slow thermal dynamics and gradual temperature variations. Although real industrial furnaces may contain nonlinearities, transport delays, thermal losses, and material-dependent dynamics, first-order models are commonly used in preliminary control design and simulation studies due to their simplicity and ability to capture the dominant thermal behavior of the process (Zhang et al., 2014; Tian et al., 2024). The simplified model used in this study therefore provides a suitable proof-of-concept platform for evaluating the proposed hybrid MPC-RL framework before future validation on more complex furnace models and real industrial systems. The continuous-time transfer function is:

$$G(s) = \frac{1}{10s+1} \tag{1}$$

This was converted into continuous-time state-space form:

$$x(t) = Ax(t) + Bu(t) \tag{2}$$

$$y(t) = Cx(t) + Du(t) \tag{3}$$

With matrices $A = -0.1$, $B = 0.1$, $C = 1$, $D = 0$. The system was then discretized using MATLAB's c2d function with a sampling time of $T_s = 0.1$ s, yielding a discrete-time state-space model:

$$x(k+1) = \Phi x(k) + \Gamma u(k) \tag{4}$$

$$y(k) = Cx(k) + Du(k) \tag{5}$$

The discrete model was used to predict the output from the MPC controllers, generate the residuals, and represent the plant model within the closed-loop Simulink structure. The overall structure of the proposed hybrid system is shown in Figure 1.

Sensor Bias Fault Modeling

Bias in sensors was chosen as the major fault due to its prevalence and practical significance in industrial temperature control systems. Sensor bias does not imply that the sensor is unable to provide an output, rather, there is a constant shift in the signal provided by the sensor. The measured temperature under fault conditions is expressed as:

$$y_m(k) = y(k) + b(k) \tag{6}$$

where $y_m(k)$ is the measured temperature, $y(k)$ is the true furnace temperature, and $b(k)$ is the sensor bias. Three bias conditions were investigated:

$$b(k) \in \{-5, 0, +5\} \tag{7}$$

Negative bias means that the controller perceives the furnace temperature as being lower than what it really is, resulting in overheating. Positive bias means the exact opposite, where the controller cools the furnace too quickly because it perceives. the temperature is higher than it

really is, resulting in low temperatures. Zero bias means fault-free operation.

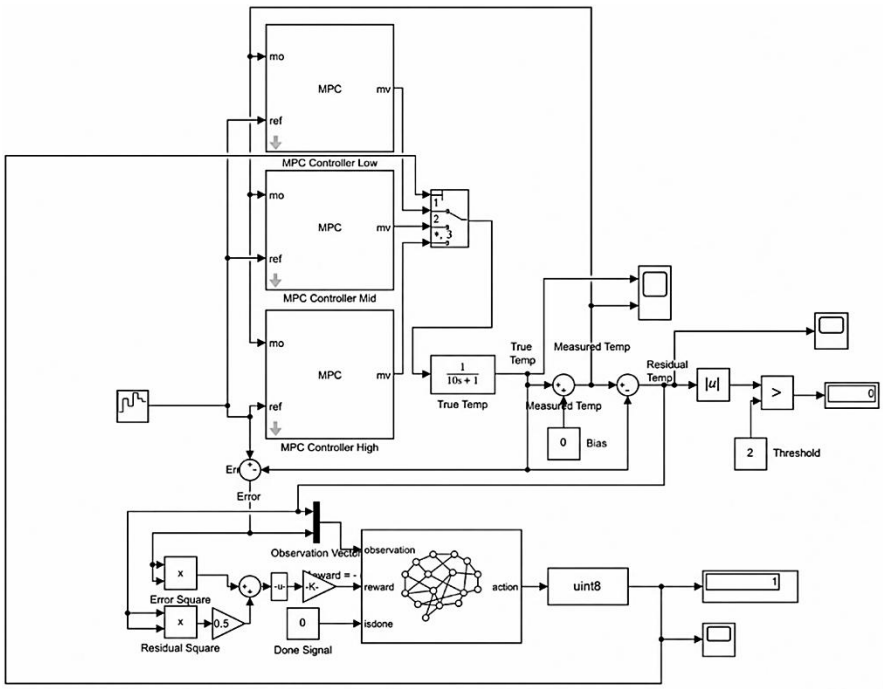

Figure 1: Overall Simulink Model of the Proposed Hybrid MPC-RL System.

Residual-Based Fault Detection

Abnormal sensor readings were detected using a residual signal, which is the difference between the measured temperature and the real temperature of the furnace:

$$r(k) = y_m(k) - y(k) \tag{8}$$

Under normal operating conditions, this residual stays around zero. In the case of a bias fault, the residual magnitude indicates the direction and amount of the bias. A threshold-based decision rule was applied:

$$|r(k)| > 2 \rightarrow Fault\ detected \tag{9}$$

The residual threshold value of 2 units was chosen based on some initial experiments performed in simulations both without faults and with faults due to sensor biases. In the absence of faults, the residual value was very close to 0 and had minor fluctuations associated with numerical discretization and model uncertainties; on the contrary, in case of faults in the form of a sensor bias of ±5 units, much bigger residual values were observed consistently. Therefore, the value of 2 was considered a good

trade-off between fault sensitivity and fault detection rate since smaller threshold values provided higher fault sensitivity but occasional false alarms, while higher threshold values led to delayed fault detection due to not detecting smaller bias values. The residual was used for two purposes in this framework: as a fault indicator and also as part of the RL agent's observations along with the tracking error, thus making the DQN agent use both performance criteria in choosing the appropriate MPC controller.

Design of the MPC Controllers

Three MPC controllers were designed using the same discrete-time furnace model, with identical prediction and control horizons:

- Prediction horizon: 20 steps
- Control horizon: 5 steps
- Sampling time: $T_s = 0.1$ s

The controllers differed in their output and manipulated variable rate weighting parameters, which produced three distinct response profiles. Table 1 summarizes the weighting parameters used for each controller.

Table 1: MPC Controller Weighting Parameters.

Controller	Output Weight	MV Weight	MV Rate Weight
Conservative	0.2	0	1.0
Normal	1.0	0	0.1
Aggressive	5.0	0	0.01

The weights were chosen in such a way as to ensure that each of the three controllers exhibits a different control action behavior in terms of aggressiveness without compromising the stability of the closed loop system. Smoother control actions arise due to low output weights coupled with high manipulated variable rate weights, while rapid reference tracking results from high output weights with low-rate penalties. The three controllers have different dynamics, allowing the DQN controller to learn different supervisory control policies in case of faults.

The conservative controller emphasizes smoothness of the control process compared to rapid tracking and is ideal where the system works under uncertain conditions. The normal controller offers a middle ground between tracking performance and the level of effort exerted in the process, which is the default mode of operation. The aggressive controller focuses on rapid tracking and correction, which makes it suitable for cases where there are significant errors in tracking. Employing three fixed controllers as opposed to tuning one controller constantly lowers computation demands, enhances stability, and streamlines the RL decision-making process.

Reinforcement Learning Agent Design

The DQN agent was developed utilizing the RL Toolbox that comes with MATLAB. This DQN agent plays the role of an additional supervisory decision-making component, as it does not provide control signals but only determines the active MPC controller.

The observation vector provided to the agent at each step was:

$$o(k) = \begin{bmatrix} e(k) \\ r(k) \end{bmatrix} \tag{10}$$

where $e(k) = r_{\text{ref}}(k) - y(k)$ is the tracking error and $r(k)$ is the residual signal. The action space was discrete:

$$a(k) \in \{1,2,3\} \tag{11}$$

corresponding to the selection of the conservative, normal, and aggressive MPC controllers, respectively. The DQN network used a single hidden layer with 32 units and a Rectified Linear Unit (ReLU) activation function. Key training parameters are listed in Table 2 and Table 3.

Table 2: DQN Agent Parameters.

Parameter	`Value
Sample time	0.1 s
Discount factor	0.99
Mini-batch size	64
Experience buffer length	100,000
Initial epsilon	1.0
Minimum epsilon	0.05
Epsilon decay	5 × 10−4

Table 3: Training Configuration.

Parameter	Value
Maximum steps per episode	300
Maximum episodes	300
Score averaging window	10
Stop training criterion	Average reward ≥-5

DQN algorithm's hyperparameters have been defined using accepted practices in the field of RL as well as simulation tests in order to ensure stability and efficient computation during the learning process. The value of discount factor of 0.99 was used to make sure that the agent will learn to optimize its long-term rewards, which is reasonable when solving sequential temperature control tasks where current control decisions affect the future performance of the system. An experience replays buffer size of 100,000 was chosen to ensure adequate diversity in the training data and minimize the relationship between successive experiences. In terms of the exploration scheme, an ε value of 1.0 at the beginning was used, enabling a large amount of exploration at the start of the training process, then decaying with a rate of 5×10^{-4} until reaching a minimal value of 0.05 to gradually move the learning process from exploration to exploitation. A

maximum number of 300 episodes and 300 steps per episode were set to make sure that the agent interacts enough with the environment for learning the policies while remaining computationally inexpensive. Lastly, a criterion for termination with an average reward of -5 over a sliding window of 10 episodes was used due to the fact that initial experiments showed little gain in the performance after that threshold.

Reward Function Design

The reward function formulated the learning goal for the RL agent. The reward function was designed such that it would penalize both poor tracking performance and high residual values due to sensor bias errors. Quadratic cost function is:

$$J(k) = e^2(k) + 0.5r^2(k) \tag{12}$$

The reward signal was the negative of this cost:

$$\text{Reward}(k) = -J(k) \tag{13}$$

The structure is designed to reward the agent for keeping the tracking process precise and also for reducing the residual, which implicitly indicates that bias correction has been successfully achieved. The value 0.5 used for the weight factor on the residual term was selected as a trade-off between the ability to detect faults and system performance.

Simulation and Evaluation Procedure

Three different simulations were created in order to gradually increase the complexity of the control task. In each setup, the RL agent was trained through repeated interaction with the Simulink environment, and performance was evaluated across three bias conditions (−5, 0, +5). The assessment focused on monitoring behavior, residual response, reward trends for episodes, and learning convergence.

Results

Experimental Setup

Three different configurations were considered for testing the hybrid MPC-RL framework for various complexities of the reference temperature profile and degree of sensor biases. The tests were conducted on the MATLAB/Simulink platform based on the models used in the previous section, i.e., the furnace, sensor bias, and supervisory control model. Tests were performed under both normal conditions and faults to determine the effectiveness of learning of the proposed methodology.

Setup 1: Constant Temperature Reference

For the first experiment, the furnace was set at a constant reference temperature of 100 units. Training could be done in up to 120 episodes. This experiment was conducted to test the steady state performance and the effects of sensor bias in stable operations. Table 4 summarizes the test conditions used for the constant temperature reference experiment.

Table 4: Constant Temperature Reference — Test Conditions.

Reference Temperature	Episodes	Sensor Bias
100	100	-5
100	100	0
100	120	+5

Constant reference situation denoted the easiest operational situation where the desired temperature of the furnace stayed constant at 100 units. The RL algorithm had the task of determining which MPC controller to select for compensation of sensor bias without considering any reference changes. The stopping criterion was satisfied for negative bias and no-fault situations but the positive bias situation did not converge.

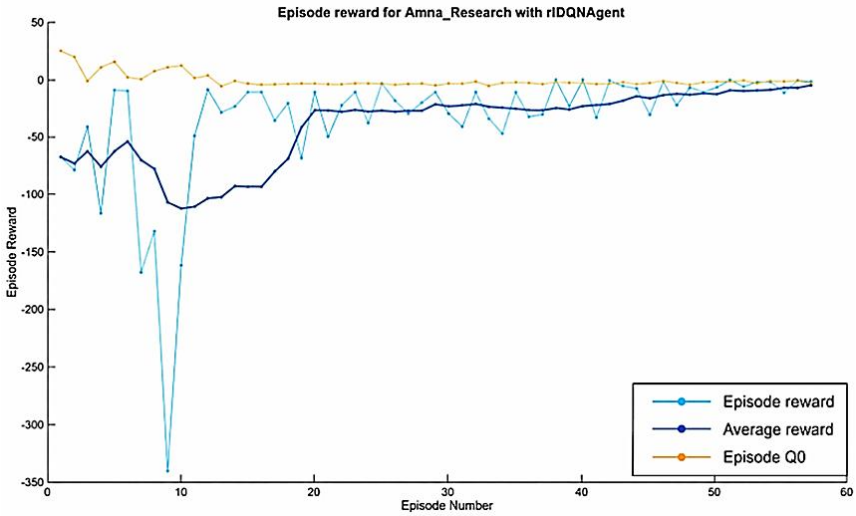

Figure 2: Training reward progression under constant reference temperature with negative sensor bias (−5).

Negative Bias (-5): Training converged at episode 57 based on the stopping criterion. Initially, the reward plot exhibited significant fluctuations owing to exploration while the RL agent was in the learning phase. However, as training continued, the reward plot shown in Figure 2 kept improving consistently as the agent learned how to deal with the

underestimation of temperatures in the environment. The negative bias led the agent to exert more heating control, which gave it a better feedback signal for learning. This shows that the model can efficiently address underestimation bias.

No Bias (0): The training process converged at episode 77 during operation without any faults. In the absence of measurement noise, the control process was initialized with a relatively stable starting point and displayed less reward fluctuations. The learning process continued to be slow because the RL agent still needed to learn the optimal controller selection policy, as shown in Figure 3.

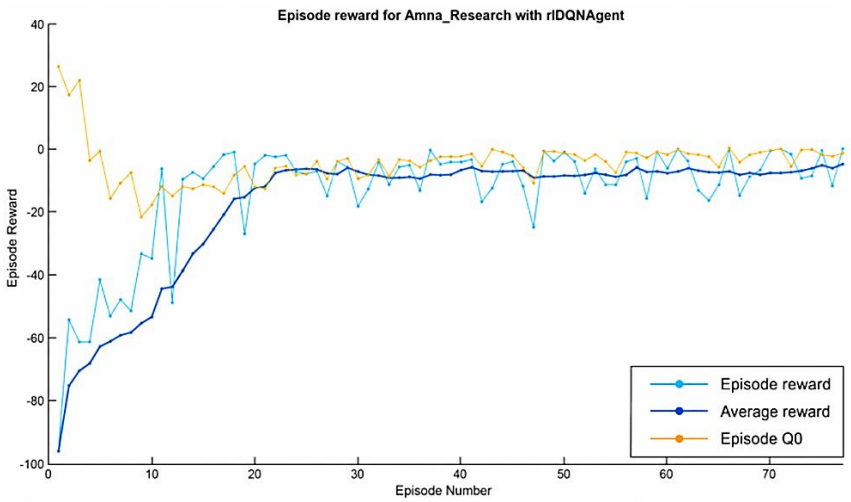

Figure 3: Training reward progression under constant reference temperature without sensor bias.

Positive Bias (+5): The training was stopped only when the maximum number of episodes of 120 had elapsed and no stopping condition had been met. With the positive bias, the observed temperature was higher than the real temperature of the furnace. The early stopping of the corrective control action led to weaker residual action, thereby hampering the learning process. Although there were improvements in the reward values, the behavior of the agent was unstable compared to that of the other scenarios, as shown in Figure 4.

In the constant reference condition, the achieved findings show that the hybrid MPC-RL approach successfully operated throughout the bias tests without any stability issues. In addition, the performance of the controller selection rule was improved based on the residual and tracking errors. As the RL algorithm was trained via numerous episodes with

varying system reactions, the discussion in this part mainly focuses on the learning process of the training rewards and convergence.

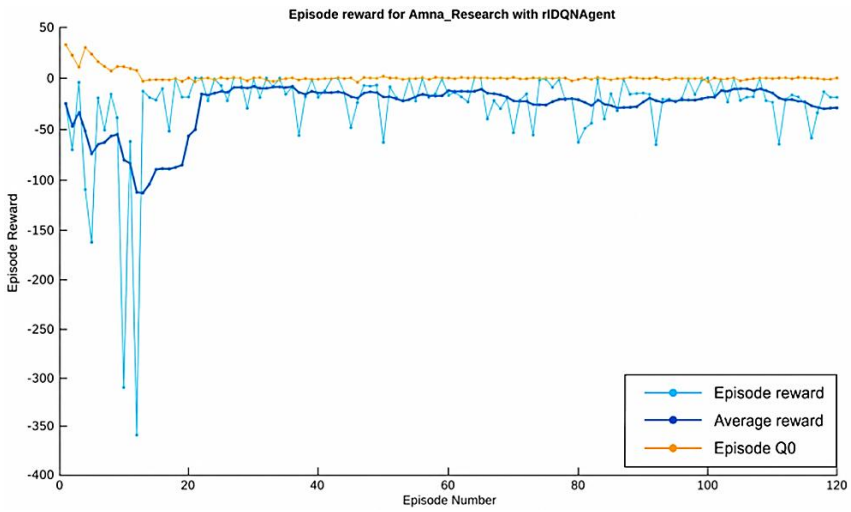

Figure 4: Training reward progression under constant reference temperature with positive sensor bias (+5).

Setup 2: Periodic Temperature Variation

The second case involved changing the reference temperature between 90 and 100 units with a switching interval of 15 seconds. It thus added dynamic tracking capability along with the need for compensation for faulty sensors. The test conditions are summarized in Table 5.

Table 5: Periodic Temperature Reference — Test Conditions.

Reference Temperature	Time Period(s)	Episodes	Sensor Bias
90 and 100	15	100	-5
90 and 100	15	100	0
90 and 100	15	120	+5

The stopping criterion was not satisfied under any of the tested bias conditions. The periodic reference profile introduced an additional control challenge since the RL algorithm had to not only solve the problem of sensor failure but also track time-varying reference values ranging from 90 to 100.

Negative Bias (−5): The reward curves exhibited fluctuations at the beginning of training owing to the impact of reference signal changes in addition to negative compensation for sensor bias. While convergence was not attained within the stipulated number of 100 episodes, there was a gradual improvement in reward values during training sessions. This

implied that the agent improved at choosing a controller, as shown in Figure 5.

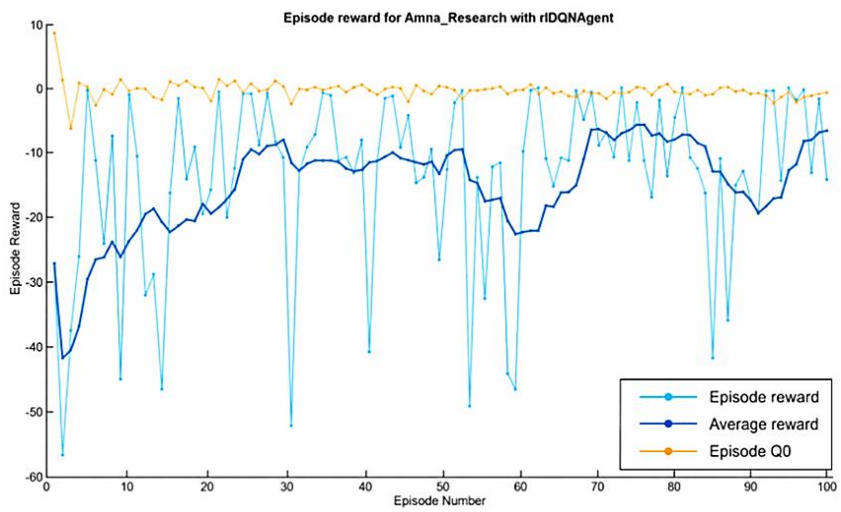

Figure 5: Training reward progression under periodic reference temperature conditions with negative sensor bias (−5).

No Bias (0): In the absence of any faults, the learning task proved more stable in the case of fault-free operation as compared to biased operation, though there was considerable variability in the rewards. As compared to the reference scenario with fixed references, the scenario with varying references made the learning problem more complicated, but rewards consistently improved during learning, as shown in Figure 6.

Positive Bias (+5): The positive bias scenario again proved to be the toughest challenge for the agent in its training. Reward variability remained very high all through the period of training for positive bias. The mean reward also increased more slowly as compared to negative bias and fault-free operation scenarios. The reason was that, due to positive bias, the corrective action taken decreased and therefore provided less information to the agent in its learning process, as shown in Figure 7.

According to the outcomes obtained under the periodic reference condition, the suggested hybrid MPC-RL algorithm operated consistently and adjusted itself to handle sensor faults and periodic reference changes. In contrast to the constant reference, convergence time increased since the learning algorithm needed to adapt to compensate for the faults and track the reference changes.

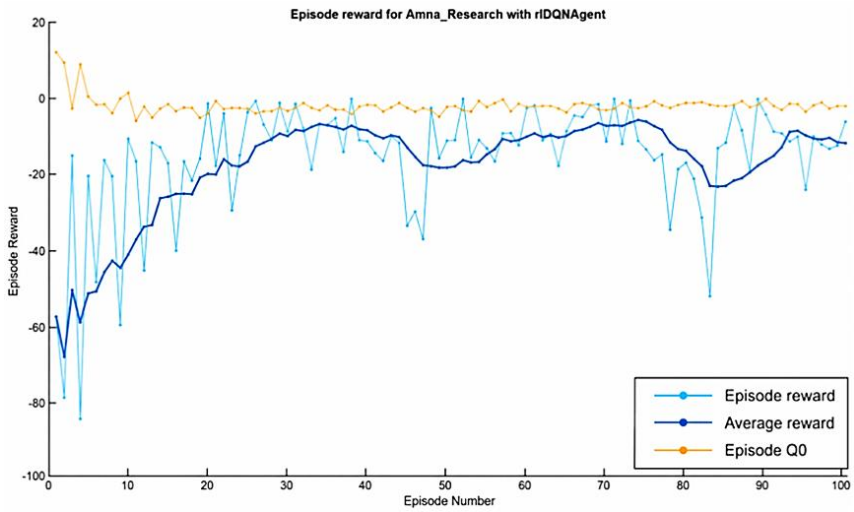

Figure 6: Training reward progression under periodic reference temperature conditions without sensor bias.

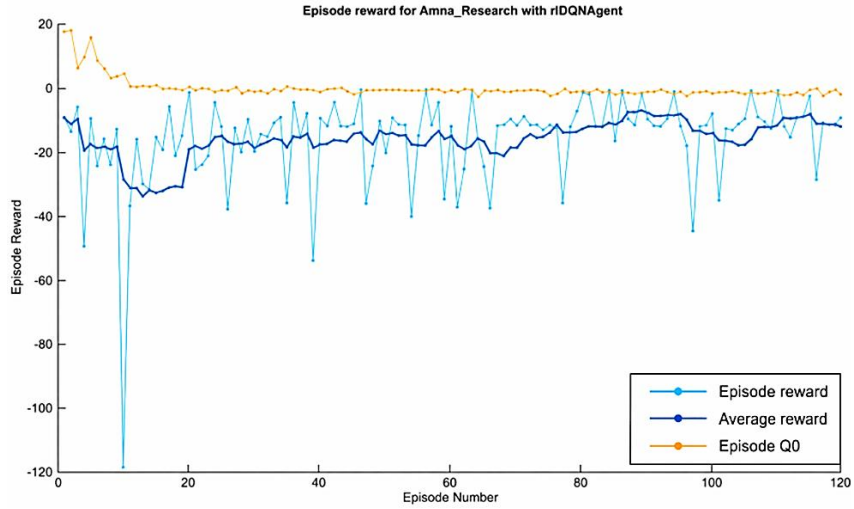

Figure 7: Training reward progression under periodic reference temperature conditions with positive sensor bias (+5).

Setup 3: Multi-Level Temperature Profile

The third configuration involved the use of a reference pattern that was more complex and consisted of [80, 100, 90, 110, 100, 95], with a cycle time of 10 seconds. The third configuration represented the most complex operating condition due to frequent switching between different

temperatures. The test conditions are summarized in Table 6. All three sensor bias configurations were tested for 120 training episodes.

Table 6: Complex Multi-Level Temperature Profile — Test Conditions.

Reference Temperature	Time Period(s)	Episodes	Sensor Bias
80,100,90,110,100,95	10	120	-5
80,100,90,110,100,95	10	120	0
80,100,90,110,100,95	10	120	+5

This scenario represented the most complex operating condition because of the frequent temperature transitions. The reference profile changed between the temperature levels, following the pattern of [80, 100, 90, 110, 100, 95]. This case required the constant adjustment on the part of the two controllers as well as the RL supervisor. None of the tested scenarios achieved convergence within the maximum training limit of 120 episodes.

Negative Bias (−5): Training sessions conducted earlier experienced drastic changes in rewards due to the fact that the RL agent was unable to adapt to fast temperature changes with negative bias sensors. Over time, as more training sessions were conducted, these oscillations diminished, and the general trend showed a positive increase in reward values, as shown in Figure 8.

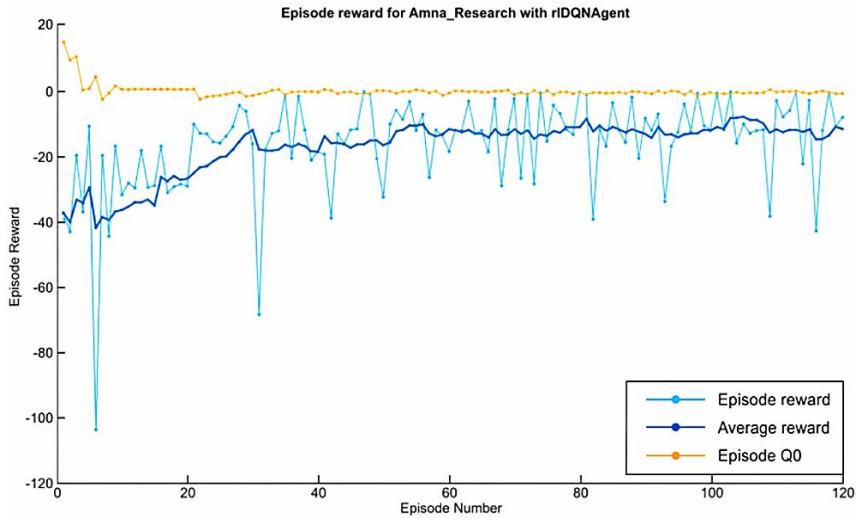

Figure 8: Training reward progression under multi-level reference temperature conditions with negative sensor bias (−5).

No Bias (0): In fault-free operation, the learning challenge was due to the wide range of reference changes as opposed to the sensor

correction issue. The reward function exhibited instability during the initial stages of training as the RL algorithm was subjected to varying operating regions in one training instance. As shown in Figure 9, the improvement in the reward function was steady as time passed.

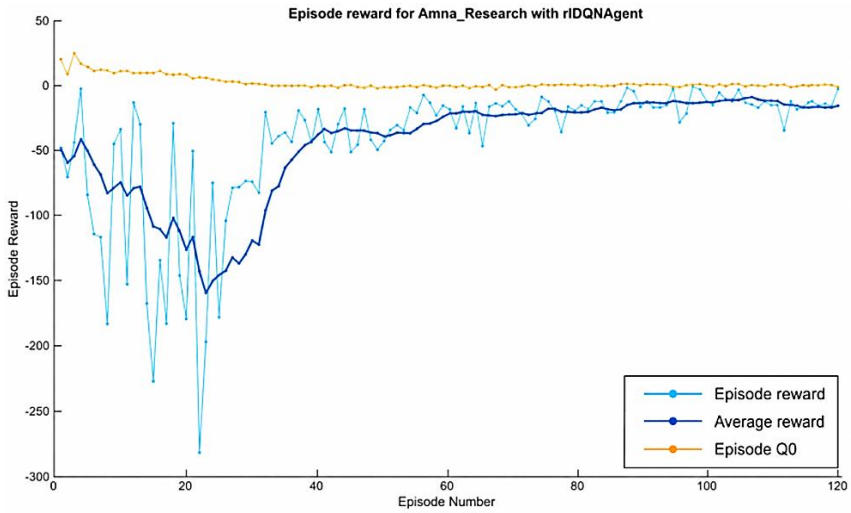

Figure 9: Training reward progression under multi-level reference temperature conditions without sensor bias.

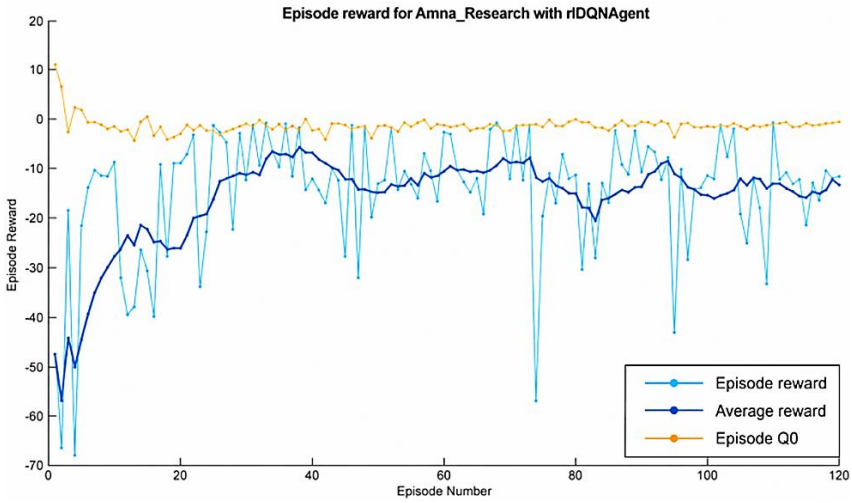

Figure 10: Training reward progression under multi-level reference temperature conditions with positive sensor bias (+5).

Positive Bias (+5): Once more, positive bias caused the slowest and most unstable behavior of the model in learning. The reward curves

showed gradual improvement but never stabilized at a certain level throughout the 120 training episodes. The combination of reduced correction capability and high system dynamics made the process of learning very complicated, as shown in Figure 10.

The outcomes from the multi-level reference configuration prove that the developed hybrid MPC-RL architecture managed to maintain stability despite the high level of dynamics in the operational environment. Despite being slower compared to the previous configurations, there is clear evidence of improvement in reward performance for all bias levels, implying learning and optimization of the controller selection algorithm.

Overall Observations

A number of learning and control regularities occurred across all simulation configurations. In all tested conditions, the ability of the RL agent to learn was evident irrespective of reference temperature and the direction of sensor bias. Nonetheless, the speed and robustness of the learning process greatly depended on both parameters.

For simpler reference profiles, faster convergence occurred. For the constant temperature conditions, two out of three cases met the required criteria within the specified episodes. For the other two kinds of references, no convergence was observed but the tendency toward learning could still be seen. A comparison of the learning effectiveness under different reference profiles and sensor bias conditions is summarized in Table 7.

Table 7: Comparison of the Learning Effectiveness under Various Reference Profile and Sensor Bias Conditions.

Reference Profile	Sensor Bias	Learning Performance	Convergence	Main Observation
Constant	-5	Fastest learning	Episode 57	Strong residuals enabled rapid policy learning.
Constant	0	Stable learning	Episode 77	Smooth convergence under fault-free operation.
Constant	+5	Slow learning	Not converged (120 episodes)	Positive bias reduced learning efficiency.
Periodic	-5	Moderate learning	Not converged	Reference variation increased learning complexity.
Periodic	0	Moderate learning	Not converged	Tracking and controller selection delayed convergence.
Periodic	+5	Slowest learning	Not converged	Bias and varying references hindered learning.
Multi-Level	-5	Moderate learning	Not converged	Frequent operating-point changes required more exploration.
Multi-Level	0	Moderate learning	Not converged	Higher reference complexity delayed learning.
Multi-Level	+5	Slowest learning	Not converged	Highest complexity produced slowest policy improvement.

Negative bias compensation was less complicated than positive bias compensation. This phenomenon can be physically explained by the fact that in case of negative bias, the corrective heating effect of the controller was more intense, resulting in greater feedback for the agent's learning. On the contrary, positive bias resulted in weaker correction.

In all simulations, the MPC module was able to ensure stability in the furnace process even when the RL was still exploring during the earlier stages. This is one key advantage of the proposed framework, wherein the furnace process stayed stable prior to the formation of the RL controller selection policy. A summary of the training convergence across all simulation conditions is presented in Table 8.

Table 8: Training Convergence Summary Across Simulation Conditions.

Setup	Bias	Episodes	Convergence Status
Constant	−5	57	Converged
Constant	0	77	Converged
Constant	+5	120	Not Converged
Periodic	−5	100	Not Converged
Periodic	0	100	Not Converged
Periodic	+5	120	Not Converged
Multi-Level	−5	120	Not Converged
Multi-Level	0	120	Not Converged
Multi-Level	+5	120	Not Converged

Discussion

Interpretation of the Main Findings

The results showed that the suggested hybrid MPC-RL method successfully enhanced control efficiency in the presence of biased sensors without compromising the stability of the closed-loop during the learning phase. This was attributed to the hierarchical structure of the suggested approach, where MPC takes care of the low-level control action and the DQN acts as a supervisory controller. On one hand, the MPC controllers ensure stable constrained control using the process model, while, on the other hand, the RL agent keeps enhancing the performance of the controller through its interaction with the environment. This division of tasks makes sure that both aspects of robustness in model-based control and adaptiveness in RL are achieved, which is in line with the works of Bertsekas (2024) and Gros & Zanon (2020).

A key observation was the effect of the sensor bias direction on the learning procedure. When there was a negative sensor bias, the measured temperature was smaller than the actual one of the furnaces, resulting in a higher value of the tracking error in the perception of the controller. As a consequence, the MPC applied more correction actions, resulting in a greater variance of the residual that provided the agent with

richer state transitions and faster convergence due to learning. In the case of positive sensor bias, there was a lower perceived tracking error resulting in weaker correction actions and smaller changes in the residual signal. Therefore, the agent received less informative feedback from the system, which negatively affected the policy updating rate and convergence. This problem of fault-tolerant learning has also been noted by Hosseini & Salahshoor (2024), Han et al. (2023), and Wang et al. (2024).

Complexity of the reference trajectory was also evident to have a significant impact on the learning process. When the reference was fixed to constant temperature, the RL algorithm was mainly concentrated on the selection of an optimal MPC controller for sensor bias compensation since the operating point was fixed. However, in case of periodic or multi-level references, continuous changes in operating points occurred and therefore, the agent had to simultaneously track the reference and compensate for faults. This way, the variety of states seen by the DQN agent while training increased as well as the effective state space that should be covered during training.

Relation to the Research Questions

The first research question regarding the impact of sensor bias on temperature control performance was addressed successfully. The simulations showed that both positive and negative bias conditions significantly affected tracking performance and caused deviations from the desired temperature values.

The second research question examined whether the proposed MPC-RL framework could compensate for sensor bias faults. The findings demonstrated that the RL agent gradually improved its controller selection policy under all tested conditions. The effectiveness of residual-based fault information is also supported by previous fault diagnosis studies (Bagla et al., 2023; Du et al., 2024).

Finally, the results confirmed that adaptive MPC selection contributed positively to overall temperature tracking performance, as reflected by the gradual improvement in reward values during training.

Comparison with Prior Research

The results of this study are largely supported by previous studies that have shown that MPC is capable of offering effective control under constraints and good reference tracking within nominal operational conditions (Rawlings et al., 2020; Camacho & Alba, 2023). Nevertheless, the operation of traditional MPC relies heavily on the accuracy of the feedback signal because the future optimal actions are determined from the measured output of the process. As a result, a constant bias of the

sensor leads to poor choices of control actions without any additional measures.

The conventional methods of FTC for dealing with sensor faults include observer-based state estimation, Kalman filtering, analytical redundancy, and fault diagnosis models. Even though these techniques are highly efficient where the properties of the faults are accurately modeled, they tend to be dependent on accurate process models. In cases where there is a change in the operational conditions or in the properties of the faults, the controllers have to be redesigned or the parameters retuned.

Some recent studies have demonstrated the effectiveness of the combination of RL and MPC techniques for enhancing the adaptive control capabilities in dynamic systems (Soza Mamani & Prado Romo, 2025; Hosseini & Salahshoor, 2024). But there are various existing methods in which RL techniques have been applied to directly produce control actions or tune the controller parameters during the control process. This might lead to increased complexity and stability issues during exploration for safety-critical industrial applications.

The suggested approach, in contrast to the other approaches, takes a different approach by using RL in a supervisory role. In this regard, instead of adjusting the control signal, the DQN selects one of the three predefined MPC controllers depending on the tracking error and residual value. The suggested architecture makes it possible to preserve the advantages of MPC in terms of stability, constraint handling, and precision, as well as adaptively select controllers depending on the fault status. Besides, the use of residuals in the RL observation space makes it possible for the supervising policy to make use of fault-sensitive data without modeling sensor bias.

One more characteristic feature of the suggested approach lies in its assessment in three reference trajectory types (constant, periodically varying, and multilevel varying) and in three types of sensor biases (negative, zero, and positive). Instead of proving the efficiency of the approach in only one scenario, the research shows the effect of the complexity of the reference trajectory and the direction of sensor bias on the process of learning and choosing the optimal supervisory controller for RL.

Practical and Theoretical Implications

In terms of the practical approach, the presented MPC-RL hybrid algorithm introduces a robust control design method that increases the fault tolerance of the industrial furnace temperature control system in case of existing sensor bias. The reason behind this is the fact that unlike traditional MPC, which uses a fixed controller, the presented supervisory

DQN agent makes a decision about the selection of one out of several possible MPC controllers based on tracking error and residual data. Thus, the control action becomes adaptable to different types of faults, without any need to reconfigure controllers continuously and estimate the sensor bias online. In this way, the supervisory approach becomes highly applicable for the industrial thermal processes.

From a theoretical point of view, the current study proves the possibility of applying RL as a supervisory decision layer in a way that does not interfere with the optimization process used by MPC. In other words, RL is suggested as an additional level of decision-making, where MPC is responsible for constrained optimal control, and DQN decides which controller is better in a particular situation. Such a hierarchy helps reduce exploration-related problems in RL control as the agent affects the controller selection, not the control itself.

Additionally, the utilization of residual data in the RL observation space helps to show how fault-tolerant attributes can help in learning the supervisory policies without having to explicitly estimate the faults, hence broadening the idea of learning-based predictive control to adaptive fault-tolerant temperature control. This is aligned with the new developments in safe and learning-based predictive control by Ham & Ah (2025) and Hewing et al. (2020).

Limitations

This research has certain limitations. Firstly, the proposed framework was validated using MATLAB/Simulink simulations, and there have been no physical hardware tests conducted. Consequently, even though the results from the simulations confirm the feasibility of the framework, they cannot yet be considered proof of its industrial applicability. Secondly, the mathematical model of the furnace is rather simplified and describes the most important dynamics of the heating process but not the whole complexity of the process such as nonlinear heat exchange, heat losses, or dependence on the material. Lastly, the faults considered in this paper are only static biases of the sensors, while industrial faults can include dynamic faults, intermittent faults, and even nonlinear ones (Amin et al., 2024; Amin et al., 2025).

Conclusion

In this study, a hybrid MPC-RL approach for the detection and compensation of sensor bias faults in industrial furnace' temperature control systems were proposed. The MPC algorithm ensured steady-state temperature control and produced residual signals as fault indicators. The RL agent trained via the DQN algorithm made adaptive choices regarding

the selection of an optimal MPC controller based on residual signals and tracking errors.

Experimental simulations conducted in MATLAB/Simulink proved that the framework operated successfully under normal and sensor bias states. Convergence was achieved faster when a constant reference temperature value was selected, while a gradual increase in the reference's complexity prolonged the learning process. It was found that negative biases were more straightforward to compensate than positive biases due to different control signal magnitudes and learning feedback quality.

In all scenarios, the proposed hybrid framework maintained stable performance during the entire learning process. Thanks to the MPC module, the furnace could not become uncontrollable until the RL algorithm had learned a good policy. It is an important characteristic because it allows using the designed method right away, improving the performance continuously with no need to wait for full convergence.

The major drawback of the presented design is that it operates purely based on simulations. Validation using actual hardware under real-world conditions is needed before any practical implementation. Moreover, studies can include situations of sensor drifts, noise, and faulty actuators. Lastly, other RL algorithms must be investigated for faster convergence and stabilization of policy. On the whole, results prove that combining predictive control with machine learning is effective for temperature regulation in industrial furnaces, and it would be interesting further investigate this hybrid approach as well.

References

AlRijeb, M. F. S., Othman, M. L., Ishak, A., Hassan, M. K., & Albaker, B. M. (2025). Machine learning-driven soft sensor implementation for real-time fault detection in CDU of oil refinery. *Engineering, Technology & Applied Science Research*, *15*(1), 20425–20432.

Amin, A. A., Iqbal, M. S., & Shahbaz, M. H. (2024). Development of intelligent fault-tolerant control systems with machine learning, deep learning, and transfer learning algorithms: A review. *Expert Systems with Applications*, *238*, 121956.

Amin, A. A., Mubarak, A., & Waseem, S. (2025). Application of physics-informed neural networks in fault diagnosis and fault-tolerant control design for electric vehicles: A review. *Measurement*, *246*,116728.

Bagla, G., Patwardhan, S. C., & Valluru, J. (2023). Intelligent state estimation for fault tolerant integrated frequent RTO and adaptive nonlinear MPC. *Journal of Process Control*, *131*, 103092.

Bertsekas, D. P. (2024). Model predictive control and reinforcement learning: A unified framework based on dynamic programming. *IFAC-PapersOnLine*, *58*(18), 363-383.

Camacho, E. F., & Alba, C. B. (2023). *Model predictive control* (3rd ed.). Springer.

Dai, B., Wang, F., Chang, Y., Chu, F., & Song, S. (2024). Process model-assisted optimization control method based on model predictive control and deep reinforcement learning for coordinated control system in coal-fired power unit. *IEEE Transactions on Automation Science and Engineering*, *22*, 228–239.

Du, P., Wilhite, B., & Kravaris, C. (2024). Model-based fault diagnosis for safety-critical chemical reactors: An experimental study. *AIChE Journal*, *70*(12). *Journal*, *70*(12).

El-Mahdy, M. H., Awad, M. I., & Maged, S. A. (2024). Deep-learning-based design of active fault-tolerant control for automated manufacturing systems subjected to faulty sensors. *Transactions of the Institute of Measurement and Control*, *46*(12), 2289–2299.

Escobar-Jiménez, R. F., Garcí a-Morales, J., & Gómez-Aguilar, J. F. (2023). Sensor fault-tolerant control for an internal combustion engine. *International Journal of Adaptive Control and Signal Processing*, *38*(1), 1–20.

Federici, L., Benedikter, B., & Furfaro, R. (2025). Reinforcement Learning Enhanced Model Predictive Control for Autonomous Planetary Landing. In *2025 AAS/AIAA Space Flight Mechanics Meeting* (pp. 2025-420).

Gros, S., & Zanon, M. (2020). Data-driven economic NMPC using reinforcement learning. *IEEE Transactions on Automatic Control*, *65*(2), 636–648.

Ham, H., & Ahn, H. (2025). A survey on safe model predictive control and reinforcement learning. *Journal of Institute of Control, Robotics and Systems*, *31*(2), 87-97.

Han, X., Rammal, R., He, M., Li, Z., Cabassud, M., & Dahhou, B. (2023). Dynamic and sensor fault tolerant control for an intensified heat-exchanger/reactor. *European Journal of Control*, *69*, 100736.

Hewing, L., Wabersich, K. P., Menner, M., & Zeilinger, M. N. (2020). Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, *3*, 269–296.

Hosseini, S. A., & Salahshoor, K. (2024). A reinforcement learning-based fault tolerant control design approach using the double Q-learning algorithm. In *Science, Engineering Management and Information Technology* (pp. 184–202).

Jiang, H., Xu, F., Wang, X., & Wang, S. (2023). Active fault-tolerant control based on model predictive control and reinforcement learning for quadcopter with actuator faults. *IFAC-PapersOnLine*, *56*(2), 11853–11860.

Leite, D., Andrade, E., Rativa, D., & Maciel, A. M. A. (2025). Fault detection and diagnosis in Industry 4.0: A review on challenges and opportunities. *Sensors*, *25*(1), 60.

Li, H., & Qiu, T. (2024). Supply responsive scheduling for ethylene cracking furnace systems based on deep reinforcement learning. *AIChE Journal*, *70*(12).

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2016). Continuous control with deep reinforcement learning. *International Conference on Learning Representations (ICLR)*.

Liu, G., Li, X., & Chen, W. (2025). Research progress on data-driven industrial fault diagnosis methods. *Sensors*, *25*(9), 2952.

Mallick, S., Airaldi, F., Dabiri, A., Sun, C., & De Schutter, B. (2025). Reinforcement learning-based model predictive control for greenhouse climate control. *Smart Agricultural Technology*, *10*, 100751.

Mesbah, A. (2016). Stochastic model predictive control: An overview and perspectives for future research. *IEEE Control Systems Magazine*, *36*(6), 30–44.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Morato, M. M., & Felix, M. S. (2024). Data science and model predictive control: A survey of recent advances on data-driven MPC algorithms. *Journal of Process Control*, *144*, 103327.

Nievas, N., Page`s-Bernaus, A., Bonada, F., Echeverria, L., & Domingo, X. (2024). Reinforcement learning for autonomous process control in Industry 4.0: Advantages and challenges. *Applied Artificial Intelligence*, *38*(1).

Ogunseye, O. S., Dabiri, Y. E., Osoko, E. A., Rahmon, S. O., & Brown, M. (2023). Reinforcement learning-based control improves efficiency and precision in smart manufacturing robotics. *World Journal of Advanced Engineering Technology and Sciences*, *9*(1), 421-431.

Qin, S. J., & Badgwell, T. A. (2003). A survey of industrial model predictive control technology. *Control Engineering Practice*, *11*(7), 733–764.

Rawlings, J. B., Mayne, D. Q., & Diehl, M. (2020). *Model predictive control: Theory, computation, and design* (2nd ed.). Nob Hill Publishing.

Razmi, D., Babayomi, O., & Zhang, Z. (2025). Reinforcement learning-driven dynamic Model Predictive Control for adaptive real-time multi-agent management of microgrids. *International Journal of Electrical Power & Energy Systems*, *170*, 110823.

Saeed, A., Khan, M. A., Obidallah, W. J., & Jawed, S. (2025). Deep learning based approaches for intelligent industrial machinery health management and fault diagnosis in resource-constrained environments. *Scientific Reports*, *15*, 1114.

Schaal, S., Atkeson, C. G., & Vijayakumar, S. (2019). Scalable techniques from nonparametric statistics for real-time robot learning. *IEEE Robotics & Automation Magazine*, *26*(1), 50–64.

Seid Ahmed, Y., Abubakar, A. A., Arif, A. F. M., & Al-Badour, F. A. (2025). Advances in fault detection techniques for automated manufacturing systems in industry 4.0. *Frontiers in Mechanical Engineering*, *11*, 1564846.

Soza Mamani, K. M., & Prado Romo, A. J. (2025). Integrating model predictive control with deep reinforcement learning for robust control of thermal processes with long time delays. *Processes*, *13*(6), 1627.

Sutton, R. S., Barto, A. G., & Barto, A. (1998). *Reinforcement learning: An introduction* (Vol. 1, No. 1, pp. 9-11). Cambridge: MIT press.

Tian, H., Tang, J., & Wang, T. (2024). Furnace temperature model predictive control based on particle swarm rolling optimization for municipal solid waste incineration. *Sustainability*, *16*(17), 7670.

Wan, Y., Xu, Q., & Dragičević, T. (2025). Reinforcement learning-based predictive control for power electronic systems. *IEEE Transactions on Industrial Electronics*.

Wang, S., Li, H., Shi, H., Sun, Q., & Li, P. (2024). Dynamic output feedback robust MPC hybrid fault-tolerant control for discrete uncertain systems: A switching control strategy. *International Journal of Control*, *98*(7), 1518–1531.

Zhang, R., Xue, A., & Gao, F. (2014). Temperature control of industrial coke furnace using novel state space model predictive control. *IEEE Transactions on Industrial Informatics*, *10*(4), 2084–2092.