Hierarchical Satellite Attitude Control via Reinforcement Learning-Guided Model Predictive Control*

Hicham Henna[1], Houari Toubakh[2], Aissa Boutte[1], Mohamed Redouane Kafi[2] and Moamar Sayed-Mouchaweh[3]

***Abstract*— Satellite attitude control under uncertainties and actuator degradation poses significant challenges to fixed-gain model-based controllers. Although MPC offers constraint handling and stability properties, its performance strongly depends on offline tuning of the cost function. This paper proposes a hierarchical RL–MPC framework in which a TD3 agent supervises the controller by adaptively modulating the MPC weighting matrices within bounded ranges, while the nominal MPC structure remains unchanged. The RL layer operates at a slower timescale and exploits long-term performance feedback to enhance control adaptability without compromising safety. Simulation results on a nonlinear satellite attitude model show that the proposed approach improves stabilization and fault-tolerant performance compared to conventional fixed-weight MPC, particularly under attitude sensor accuracy loss. The results demonstrate that hierarchical RL supervision can effectively enhance MPC performance while maintaining real-time feasibility and control reliability.**

I. INTRODUCTION

Satellite attitude control is a critical subsystem for ensuring mission success in space applications, including Earth observation, communication, and scientific exploration [1]. Accurate attitude regulation directly impacts payload pointing precision, power management, and overall mission lifetime. However, the space environment is inherently harsh and uncertain, exposing spacecraft to disturbances, model uncertainties, actuator limitations, and gradual performance degradation of onboard components [2]. These challenges motivate the development of control strategies that are not only high-performing but also adaptive and robust over long-term operation [3], [4].

Model Predictive Control (MPC) has emerged as an attractive solution for satellite attitude control due to its ability to explicitly handle multivariable dynamics, actuator constraints, and performance trade-offs through an optimization-based formulation [5], [6]. MPC has been successfully applied to rigid-body attitude regulation using quaternion representations, enabling constraint-aware control while avoiding singularities associated with Euler angles [7]. Nevertheless, the effectiveness of MPC critically depends on the selection of weighting matrices in the cost function, which balance tracking accuracy against control effort. These parameters are typically tuned offline and remain fixed during operation, limiting performance under varying operating conditions, large-angle maneuvers, or changes in system dynamics.

To address such limitations, adaptive and fault-tolerant control approaches have been widely investigated. Fault-tolerant attitude control schemes aim to preserve stability and acceptable performance in the presence of actuator or sensor degradations [3], [8], [9], [10], [11]. While these methods enhance robustness, they often rely on explicit fault models or conservative design margins, which may lead to suboptimal performance in nominal conditions. Moreover, retuning MPC parameters online in a principled and stable manner remains a challenging problem.

Recent advances in Reinforcement Learning (RL) have opened new perspectives for adaptive spacecraft control. RL-based attitude control methods have demonstrated the capability to learn control policies directly from interaction data, achieving competitive performance in both simulation and experimental settings [12], [13]. However, end-to-end RL approaches raise concerns regarding safety, interpretability, and sample efficiency, particularly for safety-critical aerospace systems. Direct torque generation by an RL agent may violate physical constraints or lead to unstable transient behaviors during learning.

To overcome these challenges, hierarchical control architectures that combine RL with classical control techniques have gained increasing attention [14], [15]. In such frameworks, RL is employed as a high-level supervisor, while a well-understood control law ensures low-level stability and constraint satisfaction. This paradigm preserves the reliability and interpretability of model-based control while leveraging the adaptability of learning-based methods.

In this paper, a hierarchical RL–MPC framework for satellite attitude regulation is proposed. The MPC controller operates as the primary control loop, generating torque commands based on a quaternion-based attitude error formulation. A Twin Delayed Deep Deterministic Policy Gradient (TD3) agent [16] acts as a supervisory layer, providing incremental and bounded adjustments to the MPC weighting matrices in real time. This design allows the controller to adapt its performance trade-offs according to the current operating regime, without granting the RL agent direct authority over actuator commands. The proposed scheme ensures safety by delayed RL actions, maintains geometric consistency in quaternion space, and mitigates the limitations associated with fixed-parameter MPC.

*This work was supported by the Algerian Space Agency (ASAL)

[1]Hicham Henna and Aissa Boutte are with the Algerian space agency, Algiers, Algeria henna.hicham@univ-ouargla.dz, hhenna@asal.dz, aboutte@cds.asal.dz

[2]Houari Toubakh and Mohamed Redouane Kafi are with the new technologies faculty, Kasdi Merbah University, Ouargla, Algeria toubakh.houari@univ-ouargla.dz, kafi.redouane@univ-ouargla.dz

[3]Moamar Sayed-Mouchaweh is with the High National Engineering School of Mines, Douai, France moamar.sayed-mouchaweh@imt-nord-europe.fr

Simulation results demonstrate that the proposed hierarchical approach improves transient performance and adaptability compared to conventional MPC. The framework also provides a practical pathway toward learning-assisted, fault-aware attitude control architectures suitable for future autonomous spacecraft.

II. METHODOLOGY

A. Model Predictive Control for satellite attitude

Model Predictive Control (MPC) is employed to regulate the satellite attitude by optimizing future control actions over finite prediction and control horizons. At each discrete time instant $\boldsymbol{k}$, the controller relies on a discrete-time prediction model of the form:

$$\mathbf{x}_{k+1} = \mathbf{f}_d(\mathbf{x}_k, \mathbf{u}_k) \tag{1}$$

where $\mathbf{x}_k \in \mathbb{R}^{n_x}$ denotes the MPC state vector, typically composed of attitude and angular velocity tracking errors, and $\mathbf{u}_k \in \mathbb{R}^{n_u}$ represents the control torque input.

The MPC optimization minimizes a quadratic cost function over a prediction horizon $\boldsymbol{N_p}$:

$$J = \sum_{i=0}^{N_p-1} \left(\mathbf{x}_{k+i}^{\top} Q \mathbf{x}_{k+i} + \mathbf{u}_{k+i}^{\top} R \mathbf{u}_{k+i} \right) \tag{2}$$

where $Q \geqslant 0$ and $R > 0$ are weighting matrices that penalize state deviation and control effort respectively. Beyond the control horizon $\boldsymbol{N_c}$, the control input is held constant, i.e.,

$$\mathbf{u}_{k+i} = \begin{cases} \mathbf{u}_{k+i}, & i = 0, \ldots, N_c - 1, \\ \mathbf{u}_{k+N_c-1}, & i = N_c, \ldots, N_p - 1. \end{cases} \tag{3}$$

which reduces computational complexity while maintaining adequate prediction accuracy.

The optimization is subject to actuator constraints:

$$\mathbf{u}_{min} \leq \mathbf{u}_{k+i} \leq \mathbf{u}_{max}$$

ensuring feasibility and compliance with hardware limitations. In a receding-horizon implementation, only the first control action is applied at each time step, and the optimization problem is solved again at the next sampling instant [17], [18].

The closed-loop performance of MPC is therefore strongly influenced by the selection of the weighting matrices Q and R, which motivates the hierarchical RL–based supervisory tuning strategy developed in this work.

B. Hierarchical Reinforcement Learning-Guided MPC

To address this limitation, a hierarchical RL–MPC architecture is proposed, where a reinforcement learning agent operates at a supervisory level and adapts selected MPC tuning parameters online, while the MPC remains responsible for real-time attitude stabilization. In this framework, the RL agent does not replace the MPC controller nor directly generate control torques. Instead, it modulates the MPC cost structure through low-dimensional tuning variables, preserving interpretability. This separation of roles allows the RL agent to focus on long-term performance optimization and fault accommodation, while the MPC ensures reliable short-term control performance.

The proposed architecture consists of two control layers:

- Inner loop (MPC controller): Executes at a fast sampling rate T_s, solves a constrained optimization problem, and generates control torques for satellite attitude regulation.
- Outer loop (TD3 supervisor): Operates at a slower decision rate $T_{RL} = nT_s, n > 1$, and updates MPC tuning parameters based on observed closed-loop behavior.

This time-scale separation ensures that the TD3 agent does not interfere with the inner MPC loop dynamics and avoids destabilizing rapid parameter fluctuations.

1) Observation space: The TD3 agent receives a compact observation vector summarizing the closed-loop system behavior over the MPC horizon. A representative observation vector is defined as:

$$\mathbf{o}_k = \begin{bmatrix} \mathbf{q}_e \\ \omega_e \end{bmatrix} \tag{4}$$

where $\mathbf{q}_e$ and ω_e denote quaternion and angular velocity errors, respectively.

2) Action space: The TD3 agent outputs a continuous action vector:

$$\mathbf{a}_k = \begin{bmatrix} \Delta\alpha_Q \\ \Delta\alpha_R \end{bmatrix}, \quad \mathbf{a}_k \in [-1, 1]^N \tag{5}$$

which represents incremental adjustments to scaling factors applied to the MPC weighting matrices. N is the number of diagonal elements of Q and R.

3) MPC parameter adaptation: The scaling factors are updated incrementally as:

$$\begin{aligned} \alpha_Q(k+1) &= \mathrm{clip}\big(\alpha_Q(k) + \eta_Q \Delta\alpha_Q, [\alpha_Q^{\min}, \alpha_Q^{\max}]\big), \\ \alpha_R(k+1) &= \mathrm{clip}\big(\alpha_R(k) + \eta_R \Delta\alpha_R, ; [\alpha_R^{\min}, \alpha_R^{\max}]\big) \end{aligned} \tag{6}$$

and applied to the nominal weighting matrices as

$$\begin{aligned} \mathbf{Q}_k &= \alpha_Q(k)\mathbf{Q}_0, \\ \mathbf{R}_k &= \alpha_R(k)\mathbf{R}_0 \end{aligned} \tag{7}$$

4) Reward Function Design: Let $\mathbf{q}_e(k)$ and $\omega_e(k)$ denote the quaternion error and angular velocity error at discrete time step $\boldsymbol{k}$, respectively. Define the quaternion error magnitude as

$$n_q(k) = \|\mathbf{q}_e(k)\|. \tag{8}$$

The instantaneous reward is constructed to encourage monotonic reduction of the attitude error, penalize angular velocity error, and regularize deviations of the MPC tuning parameters from their nominal values. To explicitly reward progress rather than absolute performance, the reward incorporates the difference between consecutive quaternion error norms:

$$r_{\Delta q}(k) = n_q(k-1) - n_q(k) \tag{9}$$

which yields a positive contribution whenever the attitude error decreases between successive time steps. Residual angular velocity error is penalized through a quadratic or norm-based term:

$$r_\omega(k) = -\lambda_\omega \|\omega e(k)\| \tag{10}$$

where $\lambda_\omega > 0$ is a weighting coefficient. For MPC Tuning regularization, let $\alpha_Q(k)$ and $\alpha_R(k)$ denote the vectors of scaling coefficients applied to the MPC state and control weighting matrices, respectively. Their magnitudes are defined as:

$$\begin{aligned} M\alpha_Q(k) &= \|\alpha_Q(k)\|, \\ M\alpha_R(k) &= \|\alpha_R(k)\| \end{aligned} \tag{11}$$

To prevent unbounded or biased drift of the MPC tuning parameters, a regularization penalty is introduced:

$$r_{\text{reg}}(k) = -\rho \left[(M\alpha_Q(k) - 1)^2 + (M\alpha_R(k) - 1)^2 \right] \tag{12}$$

where $\rho > 0$ is a small regularization coefficient.

The final instantaneous reward is given by:

$$\begin{aligned} r(k) = &\left(n_q(k-1) - n_q(k)\right) - \lambda_\omega \|\omega_e(k)\| \\ &- \rho \left[(M\alpha_Q(k) - 1)^2 + (M\alpha_R(k) - 1)^2 \right] \end{aligned} \tag{13}$$

This reward structure has several desirable properties:

- **Progress-based shaping**: By rewarding reductions in quaternion error rather than its absolute value, the agent is encouraged to improve convergence speed,
- **Dynamic regulation**: Angular velocity penalties discourage oscillatory or aggressive maneuvers,
- **Stability-preserving supervision**: Regularization anchors the MPC tuning parameters around their nominal values, preventing systematic drift toward extreme weightings,
- **Compatibility with hierarchical control**: The reward depends on closed-loop performance rather than directly on the tuning actions, improving learning robustness.

The proposed hierarchical TD3-MPC method is shown in Fig.1.

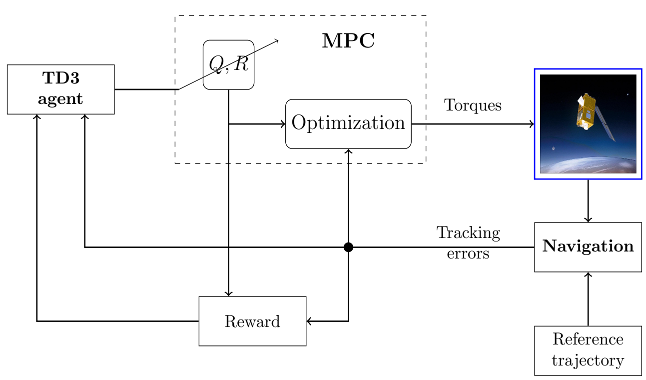


Fig. 1: Hierarchical TD3-MPC for attitude stabilization.

III. RESULTS

All simulations were conducted on a rigid-body satellite model with fixed inertia properties, representative of a small-to-medium class spacecraft. The inertia matrix was assumed constant and diagonal ($[I] = diag([4.0, 3.5, 5.2])Kg.m^2$), while actuator limitations were explicitly enforced through bounded control torques to reflect realistic reaction wheel or torque-rod capabilities ($Tc_i \in [-0.1, 0.1]N.m, for \quad i \in X, Y, Z$)). These constraints were incorporated both in the inner-loop MPC formulation and during the training of the supervisory RL agent to ensure feasibility and consistency between learning and deployment phases.

The TD3 agent was trained in closed loop with the MPC controller under randomized initial attitude conditions and fixed reference orientations per episode. The sampling time of the MPC controller and the supervisory TD3 agent were selected to reflect a multi-rate architecture, with the TD3 agent operating at a slower decision rate than the inner-loop controller. The complete set of MPC design parameters and TD3 hyperparameters employed during training is summarized in Table I, including prediction horizon length, weighting matrices, learning rates, exploration noise characteristics, and discount factors.

TABLE I: Hyper-parameters setting for MPC & TD3.

Hyper-parameter		Value
MPC	Prediction horizon	10
	Control horizon	5
	Q	$diag([10, 10, 10, 1, 1, 1])$
	R	$diag([0.1, 0.1, 0.1])$
	Time step (Ts_{MPC})	0.25
TD3	# of Episodes	2000
	Replay buffer	2×10^6
	Discount factor	0.999
	σ Exploration	0.05
	σ Decay rate	5×10^{-6}
	Optimizer	Adam
	Actor learning rate	3×10^{-4}
	Critic learning rate	3×10^{-3}
	Time step (Ts_{TD3})	$5 \times Ts_{MPC}$

The learning performance of the proposed approach is assessed through the evolution of the cumulative episodic reward during training. Fig. 2 illustrates the resulting learning curve, which reflects the progressive improvement of the closed-loop attitude regulation performance as the TD3 agent learns to adapt the MPC tuning parameters.

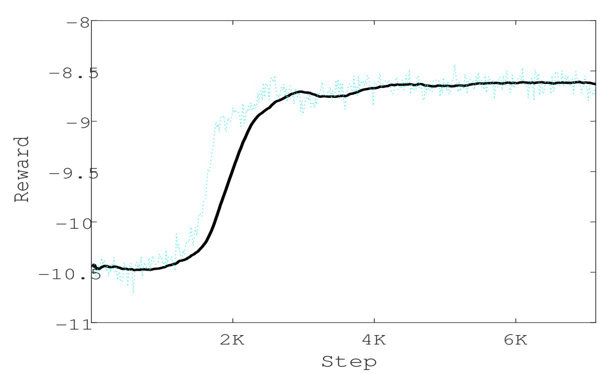


Fig. 2: Training curve of TD3 agent.

A. Attitude stabilization

The first set of simulations evaluates the ability of the proposed hierarchical TD3–MPC scheme to stabilize satellite attitude from both small and large initial orientation errors. The baseline comparison is performed against the conventional MPC controller operating with fixed weighting matrices. For these experiments, the satellite model is initialized with two different attitude deviations: a small initial error, where the quaternion is close to the reference, and a large initial error, corresponding to a significant rotational displacement from the desired orientation. All other simulation parameters, including satellite inertia, actuator torque limits, and MPC prediction and control horizons, are kept identical for fair comparison.

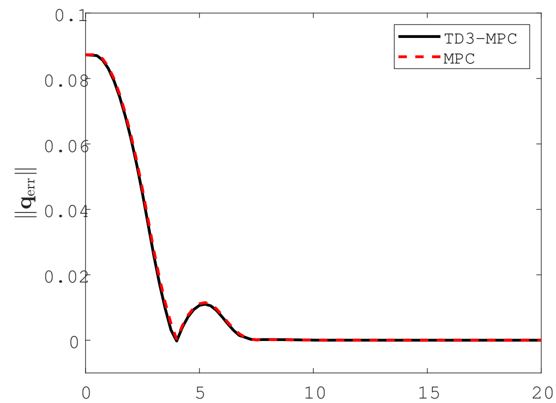
Fig. 3 illustrates the evolution of the quaternion error norm for both initial conditions. Each subplot contrasts the performance of the two control strategies. In the case of small initial errors, both controllers rapidly reduce the orientation error; however, the TD3–MPC achieves slightly faster convergence. For large initial errors, the advantage of the hierarchical RL–MPC is more pronounced: the adaptive adjustment of the MPC weighting matrices by the TD3 agent leads to obvious error reduction, whereas the fixed MPC exhibits larger transient peaks. These results confirm that the RL supervisory layer enhances the stabilization performance, particularly under challenging initial conditions, while preserving the safety and feasibility ensured by the underlying MPC.

B. Fault-tolerant performance

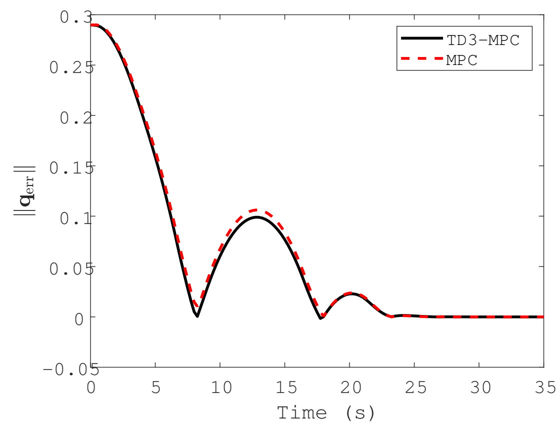
To assess the fault-tolerant behavior of the proposed hierarchical TD3–MPC scheme, a star sensor fault was introduced at $t = 25s$, resulting in a sudden degradation of attitude measurement quality which may be induced by a Big bright object (BBO) entering the sensor's field of view. When a BBO enters the star sensor field of view, the resulting disturbance can be modeled as a degradation of measurement quality, characterized by increased noise levels, potential bias in attitude estimation, and occasional outliers or temporary loss of valid measurements.. This disturbance challenges the controller's ability to preserve stability and regulate control effort in the presence of partial loss of sensing accuracy.

Fig. 4 compares the evolution of the control torque norm generated by the baseline MPC and the TD3–MPC controllers. Following the fault occurrence, both schemes react appropriately to the measurement disturbance and succeed in maintaining closed-loop stability. However, the TD3–MPC exhibits a slightly lower torque norm during the transient phase. Although torque peaks are still present in both cases, the TD3–MPC consistently produces smaller peak magnitudes, indicating a more moderated control action. The difference between the two schemes reflects a tangible benefit of adaptive weighting in reducing excessive actuator effort under faulty sensing conditions.

The adaptation mechanism of the RL supervisory layer is further illustrated in Fig. 5, which depicts the evolution of the nine scaling factors applied to the MPC weighting matrices. These include six coefficients associated with the state weighting matrix Q and three associated with the input weighting matrix R. After the fault is introduced, the TD3 agent adjusts these coefficients to rebalance the trade-off between state regulation and control effort. While the adaptations remain bounded and gradual, they demonstrate the agent's capability to fine-tune the MPC cost function online in response to degraded measurements, thereby enhancing robustness without inducing aggressive control behavior.



(a) Small initial attitude error.



(b) Large initial attitude error.

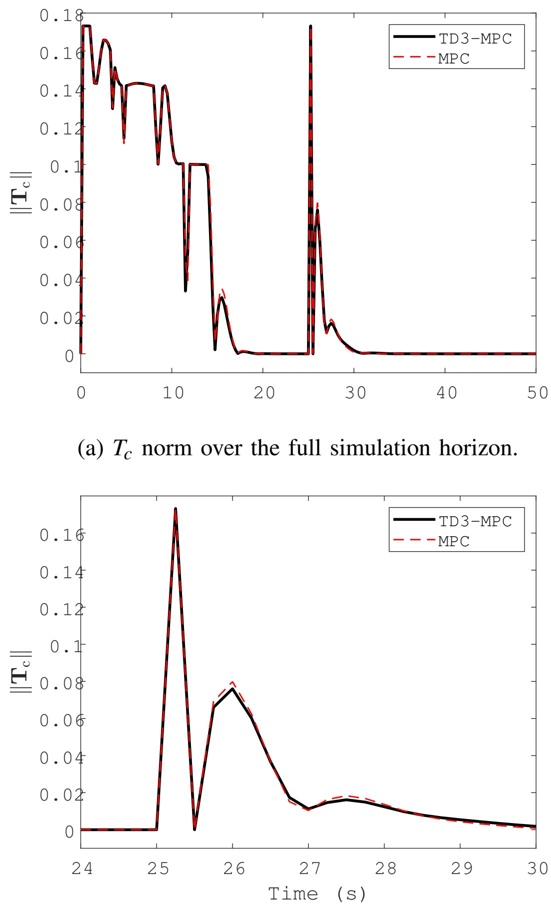
Fig. 3: Attitude stabilization results.

TABLE II: Performance comparison between the proposed approach and conventional MPC.

Criterion	TD3-MPC	MPC
Settling time (s)	23.6	24
Overshoot (%)	11.7	12.3
RMSE (large initial q_{err})	0.472	0.503
RMSE (small initial q_{err})	0.179	0.195

IV. DISCUSSION

The results in Section III show that both MPC and TD3–MPC successfully stabilize the satellite attitude under nominal conditions and in the presence of a star sensor fault. In the stabilization scenarios, the TD3–MPC achieves a slightly faster and smoother reduction of the quaternion error norm (see Fig. 3b), while consistently generating a



(a) T_c norm over the full simulation horizon.

(b) Zoomed view of T_c norm following the fault occurrence.

Fig. 4: T_c norm under the induced fault scenario.

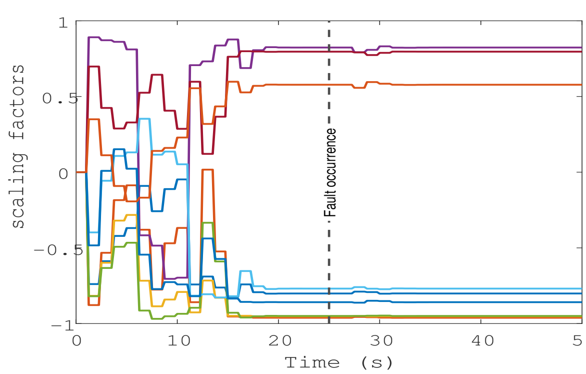


Fig. 5: Q and R adaptation by TD3 agent.

lower control torque norm, as illustrated in Fig. 4a. This improvement is observed even before fault occurrence, highlighting the ability of the RL supervisory layer to enhance control efficiency during nominal operation. Furthermore, an overall performance comparison is detailed in Table II, showing that the proposed scheme provides better time-domain performance metrics.

It is worth-noting here that closed-loop stability is preserved by the underlying MPC, which enforces standard constraints and admits a Lyapunov-based interpretation ensuring recursive feasibility and stability under bounded disturbances. The RL layer only performs bounded adaptation of Q and R, keeping them positive definite and within a predefined compact set; thus, it preserves the MPC cost structure and does not violate stability conditions. Consequently, the scheme can be interpreted as a constrained gain-scheduling of MPC, for which stability arguments are well established in the literature.

Under faulty sensing conditions, both controllers preserve closed-loop stability; however, the TD3–MPC exhibits moderately reduced torque peaks following the fault, as shown in Fig. 4b. Although the performance difference remains limited, the results indicate that the adaptive tuning provided by the TD3 agent improves both attitude regulation and control effort without inducing aggressive behavior. The performance gains introduced by the TD3–MPC scheme are incremental but consistent across all scenarios. This is primarily due to the hierarchical design, in which the RL agent performs bounded scaling of the MPC weighting matrices rather than directly modifying the control law. Such constrained adaptation results in fine adjustments that preserve MPC stability properties while improving efficiency and robustness.

Compared to fixed MPC, the TD3–MPC framework introduces a limited additional computational load associated with RL inference. Since the MPC optimization problem remains unchanged and only a small set of scaling coefficients is updated, the online complexity increase is marginal. Moreover, enforcing bounded tuning ranges for the Q and R matrices ensures positive definiteness of the cost function and prevents unsafe or numerically ill-conditioned control actions. Further, the RL supervisory layer operates at a slower time scale than the MPC controller ($T_{s,TD3} > T_{s,MPC}$), which explains the short adaptation delay observed after fault occurrence in Fig. 5. While this delay slightly postpones tuning updates, it contributes to system robustness by avoiding abrupt parameter changes. The results suggest that this trade-off between responsiveness and safety is acceptable for spacecraft attitude control.

V. CONCLUSIONS

This paper proposed a hierarchical TD3–MPC framework for satellite attitude control, where a reinforcement learning agent supervises a model predictive controller through bounded online tuning of its cost function. This structure retains the stability and constraint-handling advantages of MPC while enabling adaptive behavior. Simulation results showed that TD3–MPC achieves slightly improved attitude stabilization compared to fixed MPC, with reduced quaternion error norms and consistently lower control torque norms under nominal conditions. In the presence of star sensor measurement faults, both controllers remained stable; however, the proposed approach exhibited moderated torque peaks and adaptive tuning, indicating enhanced fault-tolerant capability.

The bounded and multi-rate supervision of the RL agent ensures safe parameter adaptation with limited computational burden. Although performance gains are incremental, they are consistent and obtained without modifying the underlying MPC formulation. While the present study focuses on validating the proposed control architecture under nominal conditions, incorporating disturbance torques, actuator dynamics (e.g., reaction wheel behavior), and orbit-dependent effects is a natural extension. This will be explicitly addressed as part of future work to enhance the practical relevance of the framework. An additional direction for future work involves advanced reward shaping strategies and enriched state representations for the TD3 agent, including augmented feature engineering and the integration of transformer-based architectures to better capture temporal dependencies and improve adaptation under complex and long-horizon attitude dynamics.

Data Availability

The datasets, simulation models, and experimental scripts used in this study are openly available on the author's GitHub repository at: `https://github.com/hichamhenna/Hybrid_RL_MPC_for_attitude_control`

References

[1] Orozco, M. L., and Giraldo, B. S. (2024). Attitude determination and control in small satellites: a review. *IEEE Journal on Miniaturization for Air and Space Systems*, 5(3), 182-186.

[2] Pourtakdoust, S. H., Fakhari Mehrjardi, M., Hajkarim, M. H., and Nasihati Gourabi, F. (2023). Advanced fault detection and diagnosis in spacecraft attitude control systems: Current state and challenges. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 237(12), 2679-2699.

[3] Henna, H., Toubakh, H., Kafi, M. R., Sayed-Mouchaweh, M., and Djemai, M. (2024). Hybrid supervision scheme for satellite attitude control with sensor faults. *CEAS Space Journal*, 16(6), 753-767.

[4] Fazlyab, A. R., Fani Saberi, F., and Kabganian, M. (2023). Fault-tolerant attitude control of the satellite in the presence of simultaneous actuator and sensor faults. *Scientific Reports*, 13(1), 20802.

[5] Bordons, C., Garcia-Torres, F., and Ridao, M. A. (2019). Model predictive control fundamentals. In *Model Predictive Control of Microgrids* (pp. 25-44). Springer International Publishing.

[6] Maciejowski, J. M., and Huzmezan, M. (2007). Predictive control. In *Robust Flight Control: A Design Challenge* (pp. 125-134). Springer Berlin Heidelberg.

[7] Markley, F. L., and Crassidis, J. L. (2014). Fundamentals of Spacecraft Attitude Determination and Control (Vol. 33). Springer.

[8] Henna, H., Toubakh, H., Kafi, M. R., and Sayed-Mouchaweh, M. (2020). Towards fault-tolerant strategy in satellite attitude control systems: A review. In *Annual Conference of the PHM Society*, (Vol. 12, No. 1, pp. 14-14).

[9] Henna, H., Gasmi, E., Toubakh, H., Kafi, M. R., Djemai, M., and Sayed-Mouchaweh, M. (2023). Fault tolerant attitude estimation for satellite at low earth orbit. In *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)* (pp. 275-280). IEEE.

[10] Mahboub, M. A., Rouabah, B., Kafi, M. R., and Toubakh, H. (2022). Health management using fault detection and fault tolerant control of multicellular converter applied in more electric aircraft system. *Diagnostyka*, 23(2).

[11] Rouabah, B., Toubakh, H., Djemai, M., Ben-Brahim, L., and Ghandour, R. (2023). Fault diagnosis based machine learning and fault tolerant control of multicellular converter used in photovoltaic water pumping system. *IEEE Access*, 11, 39013-39023.

[12] Vedant, J. T. (2019). Reinforcement learning for spacecraft attitude control. In *70th International Astronautical Congress*.

[13] Elkins, J. G., Sood, R., and Rumpf, C. (2020). Autonomous spacecraft attitude control using deep reinforcement learning. In *International Astronautical Congress*.

[14] Henna, H., Toubakh, H., Kafi, M. R., Gürsoy, Ö., Sayed-Mouchaweh, M., and Djemai, M. (2024). Satellite fault tolerant attitude control based on expert guided exploration of reinforcement learning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 37(6), 987-1011.

[15] Mahmud, S. N., Nivison, S. A., Bell, Z. I., and Kamalapurkar, R. (2021). Safe model-based reinforcement learning for systems with parametric uncertainties. *Frontiers in Robotics and AI*, 8, 733104.

[16] Fujimoto, S., Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *International conference on machine learning* (pp. 1587-1596). PMLR.

[17] Hegrenæs, Ø., Gravdahl, J. T., and Tøndel, P. (2005). Spacecraft attitude control using explicit model predictive control. *Automatica*, 41(12), 2107-2114.

[18] Murilo, A., de Deus Peixoto, P. J., de Souza, L. C. G., and Lopes, R. V. (2021). Real-time implementation of a parameterized Model Predictive Control for Attitude Control Systems of rigid-flexible satellite. *Mechanical systems and signal processing*, 149, 107129.