Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

International Journal of Adaptive Control and Signal Processing
Volume 35, Issue 3, March 2021, pp. 420-440. DOI:10.1002/acs.3181

SPECIAL ISSUE

Reinforcement Learning Based Closed-Loop Reference Model Adaptive Flight Control System Design

Burak Yuksek*[1] | Gokhan Inalhan[2]

[1]Aerospace Research Center, Istanbul Technical University, Istanbul, Turkey
[2]School of Aerospace, Transport and Manufacturing; Centre for Autonomous and Cyber-Physical Systems, Cranfield University, Bedford, UK

Correspondence
*Burak Yuksek. Istanbul Technical University, Ayazaga Campus, Aerospace Research Center, 34469, Istanbul, Turkey, Email: yuksekb@itu.edu.tr

Summary

In this study, we present a reinforcement learning (RL)-based flight control system design method to improve the transient response performance of a closed-loop reference model (CRM) adaptive control system. The methodology, known as RL-CRM, relies on the generation of a dynamic adaption strategy by implementing RL on the variable factor in the feedback path gain matrix of the reference model. An actor-critic RL agent is designed using the performance-driven reward functions and tracking error observations from the environment. In the training phase, a deep deterministic policy gradient algorithm is utilized to learn the time-varying adaptation strategy of the design parameter in the reference model feedback gain matrix. The proposed control structure provides the possibility to learn numerous adaptation strategies across a wide range of flight and vehicle conditions instead of being driven by high-fidelity simulators or flight testing and real flight operations. The performance of the proposed system was evaluated on an identified and verified mathematical model of an agile quadrotor platform. Monte-Carlo simulations and worst case analysis were also performed over a benchmark helicopter example model. In comparison to the classical model reference adaptive control (MRAC) and CRM-adaptive control system designs, the proposed RL-CRM adaptive flight control system design improves the transient response performance on all associated metrics and provides the capability to operate over a wide range of parametric uncertainties.

KEYWORDS:
Variable Closed-loop Reference Model Adaptive Control, Reinforcement Learning, Adaptive Flight Control System, Resilient Control

1 | INTRODUCTION

Operational safety in urban air mobility (UAM) is a crucial factor that impacts its reliability and sustainability as a viable model for cargo or passenger transportation. While numerous operational requirements are driven by the critical challenges associated with airspace integration, a considerable number of design requirements are directly associated with the requirements related to the performance and robustness of its subsystems. From the perspective of flight control system, the performance and robustness of aerial vehicles define the ability of closed-loop systems for tracking a given reference signal over the flight envelope with minimum error, overshoot and settling time while staying within the limits of actuator saturation. These requirements should be satisfied in all nominal operating conditions covered by the flight envelope. The aforementioned requirements become further

[0]**Abbreviations:** MRAC, model reference adaptive control; CRM, closed-loop reference model; RL, reinforcement learning

complicated owing to the real-life variations in the mass, moment of inertia, aerodynamic properties, or power system properties of aerial vehicles due to the changes observed in a wide range of operating conditions and payload weights. In addition, the control system design should consider the adverse flight conditions caused by the sudden changes in mechanical and aerodynamic properties of aerial vehicles due to faults, failures, and anomalies that occur on the subsystems of aerial platforms such as actuators. To provide operation safety in the urban airspace, it is critical to compensate the possible dynamical variations in the flight control systems along with resilience to maintain stability and performance requirements even in severe and faulty flight conditions. In this paper, we present a new reinforcement learning (RL)-based approach for a closed-loop reference model (CRM) adaptive flight control system design to further enhance the adaptation transient response beyond the existing model reference adaptive control (MRAC) and classical CRM-adaptive systems. The proposed methodology implements RL, through an actor-critic agent to learn the time-varying adaptation policy using the tracking error observations from the environment.

In UAM applications, the use of electric-powered aerial vehicles with vertical takeoff and landing (VTOL) capabilities is very common owing to the infrastructure and operation requirements. The power and propulsion systems of these aerial vehicle concepts consist of battery packs, brushless DC (BLDC) motors, electronic speed controllers (ESC), and propellers. Although they have a simpler system architecture in comparison to conventional helicopter concepts, they are still prone to various types of faults. For example, excessive current requirements may lead to ESC and/or BLDC motor faults, thereby leading to thrust loss on the related BLDC motor - propeller assembly. Moreover, the partial structural loss in the propeller due to mechanical impact or excessive blade flapping causes thrust loss and catastrophic accidents in the urban airspace. In addition to in-flight faults and failures, variations in the mechanical properties of aerial vehicles, such as the center of gravity location, mass, and inertia, also affect the dynamical characteristics of the system. These variations directly affect the system parameters such as the key roll and pitch aerodynamic derivatives; for example, M_q, L_p, and M_{δ_e}. Hence, it is critical to consider these worst-case scenarios and design control systems that consider the modeling uncertainties and modeling variations caused by faults and failures on the aerial platform.

In recent years, resilient control has become a critical topic especially in the aerial vehicle technology to provide operational safety and reliability. The development of more complex vehicles has increased the importance of the resilience, defined as the recovering ability of closed-loop systems in the presence of faults, disturbances, and uncertainties. The most important feature of resilient control systems is to guarantee the desired system performance while handling the uncertain system faults and disturbances. This ability requires effective estimation and compensation of the uncertainties in the system dynamics. Hence, the resilient control systems are developed based on two fundamental research topics: a) fault detection, isolation, and fault tolerant control methods[1,2] and b) adaptive control theory[3,4]. This study focuses on the adaptive control theory in the context of resilient control. Specifically, adaptive flight control systems have promising potential in urban airspace applications owing to their critical characteristics. These characteristics include the abilities to a) recover from anomalies without entering into unexpected/undesirable physical states, b) reconfigure/adapt the control parameters to achieve identical or significantly similar performance as the original design specifications and c) ensure graceful degradation of performance as the anomalies increase. Within the context of aerial vehicles, as a part of designing resilient controllers, the transient behavior observed during reconfiguration and adaptation is a very important part of the flight experience. Accordingly, if the error magnitude between the system states and reference model states is high, the oscillatory transient response can lead to uncontrolled and undesirable flight states that can risk the passengers and the operational safety of the platform. However, most of the basic adaptive control algorithms exhibit considerable limitations in providing controlled and bounded transient response within the aerospace context.

The MRAC systems are a fundamental method for implementing adaptive control. MRAC has been successfully applied on several aerial platforms. Moreover, to increase the robustness characteristics of MRAC, several modifications, such as σ and μ modifications, have been developed and applied on various platforms. However, the fundamental limitation of MRAC is the presence of high-frequency oscillations in the control signal and system response at the beginning of the adaptation process. Even if the convergence speed of the adaptation parameters are sufficient, this oscillatory behavior may result in catastrophic accidents, especially in aerial applications. Hence, the transient response of MRAC should be modified and improved to provide operation safety. To accomplish this, several developments have been implemented on the classical MRAC structure and, consequently, improved transient dynamics have been obtained by utilizing the combined/composite model reference adaptive control (CMRAC)[5,6] and CRM[7,8] adaptive control algorithms. Indirect and direct adaptive control methods are combined in the CMRAC algorithm, wherein the adaptation laws include the estimated system parameters. This approach provides robustness in the presence of parameter uncertainties. However, its ability to handle and improve the high-frequency oscillations in the transient response phase is unproven and remains a conjecture[5,7]. In CRM-adaptive systems, the transient response performance of

the closed-loop system is improved by utilizing an observer gain on the feedback path of the reference model. An adequate transient response performance can be obtained by using an optimized feedback gain in the reference model. The oscillatory control signal, adaptive parameters, and system response can be damped if an optimal feedback gain matrix is used on the reference model feedback path.

In literature, there are several studies that focus on the CMRAC and CRM adaptive control systems. Lavretsky[5] introduced the CMRAC design methodology that can be implemented in generic multi-input-multi-output (MIMO) systems. The design methodology does not require online measurements of the system state derivatives. Gregory et al.[6] used CMRAC to augment the baseline flight control system of the NASA Generic Transport Model and performed flight tests to evaluate the system performance. Dydek et al.[9] augmented the baseline flight control system of a quadrotor unmanned aerial vehicle (UAV) by utilizing the CMRAC structure. They evaluated their system's performance in the presence of loss-of-thrust anomaly that may occur owing to propeller damage. Cho et al.[10] utilized a new parameter estimation method based on the regressor filtering scheme in the CMRAC algorithm to relax the persistent excitation requirements. Wiese et al.[11] implemented an adaptive output feedback control algorithm based on CRM on a six degree-of-freedom (DoF) scramjet-powered, blended body, generic hypersonic aerial vehicle model. Zollitsch et al.[12] augmented a linear quadratic regulator (LQR) baseline flight control system using a CRM-adaptive controller that can compensate the uncertainties in the control effectiveness of the FSD ExtremeStar UAV platform. The performance of the proposed system was evaluated in a high-fidelity simulation environment. Gibson et al.[13,8] developed a CMRAC-CO adaptive control system using the CRM model in the CMRAC structure with an observer; the CRM-adaptive control algorithm was extended to output feedback systems in[14]. The proposed structure could provide noise-free state estimation while guaranteeing stability and improved transient performance. However, the inherent nature of CRM presents a trade-off between the improved transient response and convergence speed of the adaptive parameters. In the case of slow adaptation, large tracking errors are observed between the original reference model and system response. Thus, a) the improper selection of reference model parameters and adaptation rates, and b) the inability to dynamically change the CRM feedback gain matrix leads to decreased system performance and results in the well-known water-bed effect[7,15]. This effect corresponds to oscillatory ripples on $|\dot{u}(t)|$, which is the time derivative of the control signal. These oscillations can further result in the saturation of the actuator displacement and its slew rates.

In this study, we present a methodology[16] to overcome this challenge by introducing a time-varying scaling parameter $k(t)$ in the feedback path gain of the CRM. This time-varying parameter corresponds to the model adaptation strategy that changes depending on the magnitude of the observations and the errors. The scaling factor is designed through RL by utilizing an actor-critic agent; it is trained by using a deep deterministic policy gradient (DDPG) algorithm across a wide range of possible real flight scenarios and anomalies. This novel control algorithm is called as RL-CRM adaptive control system and its general

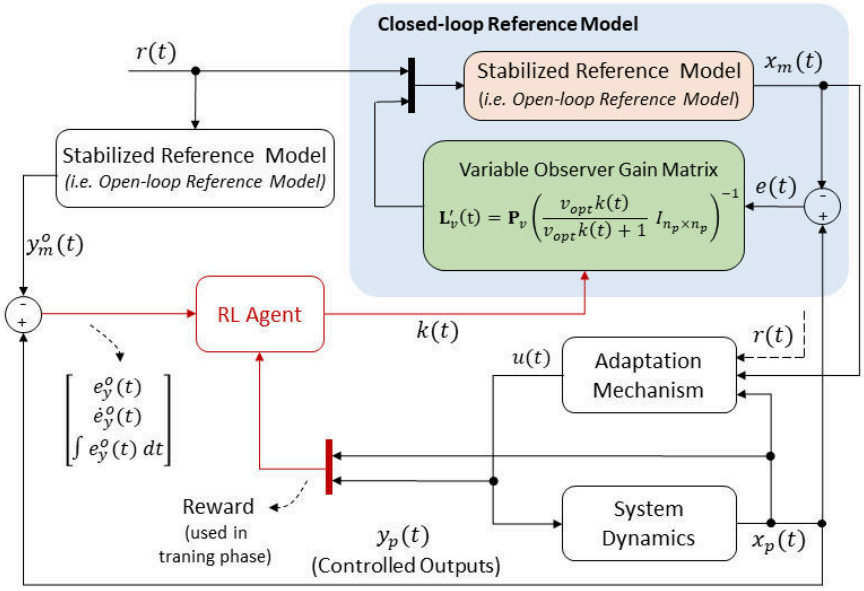

FIGURE 1 General structure of the RL-CRM adaptive control system.

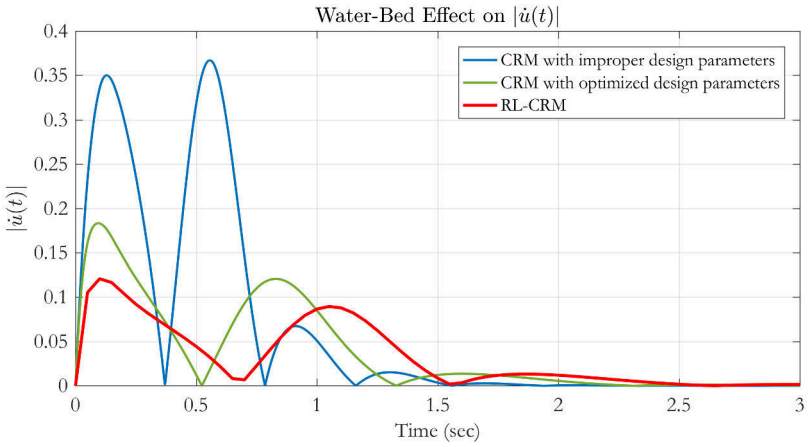

FIGURE 2 Water-Bed effect on $|\dot{u}(t)|$ signal[16].

scheme is illustrated in Figure 1. As shown in Figure 1, closed-loop reference model consists of stabilized reference model (i.e. ideal reference model) and variable observer gain $\mathbf{L}'_v(t)$ in which the time-varying scaling parameter $\big(k(t)\big)$ is utilized to scale the optimized tuning parameter (v_{opt}). Scaling policy is provided by the RL agent which uses reward and observation vector for training process. The adaptation mechanism has the same structure as the classical CRM-adaptive system and it is developed by using the Lyapunov theory. More detailed explanation of the RL-CRM adaptive control scheme including detailed explanations associated with each of the blocks, formulas and signal elements are given in Sections 2 and 3. The proposed approach allows us to learn and design complex adaptation strategies that could not be designed through standard analytic methodologies. Correspondingly, we propose the integration of the fast convergence speed of the MRAC and improved transient dynamics of the CRM-adaptive system. In addition, as illustrated in Figure 2, the proposed control system provides better transient response performance and further suppression of the water-bed effect in terms of $|\dot{u}(t)|$ when compared to the classical CRM-adaptive controller. Here, $\dot{u}(t)$ is time derivative of the control signal, i.e. slew rate.

We evaluated the performance of the RL-CRM algorithm using an identified and verified high-fidelity mathematical model of an agile multicopter platform. The identified mathematical model was obtained as a part of the desktop-to-flight control system design workflow application shown in Figure 3 which provides a seamless design process. This workflow begins with the frequency-domain system identification process in which the Comprehensive Identification from FrEquency Responses (CIFER)[18] tool is used to obtain the linear mathematical model of the agile multicopter platform for the hover and fast forward flight phases. Then, verified linear models are utilized to optimize the controller parameters in the Control Designer's Unified Interface (CONDUIT)[17] tool which is developed based on Feasible Sequential Quadratic Programming (FSQP) algorithm. In the desktop simulation phase, several preliminary analysis are performed on Matlab/Simulik simulation environment and system performance is evaluated. After that, the proposed control algorithm is embedded into the flight control computer and Hardware-in-the-Loop (HIL) tests are performed. Overall system performance is evaluated in the flight tests and dynamics requirements are validated. This workflow is applied until obtaining a satisfactory flight control system. To increase the intuition about the frequency-domain system identification process, recorded frequency-sweep flight data, non-parametric/parametric modeling and time-domain verification results for pitch axis tests are given in Figure 3. Here, a_x is longitudinal acceleration, q is pitch rate, θ is pitch angle, δ_e is longitudinal input for the mixer. Readers may refer to Yuksek et al.[19] for more information about system identification and classical control system design for the agile multicopter platform.

For an initial insight on the effect of model variation and uncertainty, if we focus on the scalar pitch dynamics model of the test platform, the step responses of the closed-loop system with MRAC, CRM, and RL-CRM adaptive controllers to a unit pitch rate command are shown in Figure 4-a. As evident in this figure, MRAC results in an oscillatory transient response, which is not a desirable situation for an aerial vehicle. The CRM-adaptive control system with optimal gain l_{opt} on the feedback path of the reference model suppresses the high-frequency oscillations and provides improved transient response characteristics in comparison to the MRAC. The response of the closed-loop system with the RL-CRM controller is illustrated in this figure as red solid line. It can be observed that the RL-CRM adaptive control system further improves the transient response of the optimized CRM-adaptive controller (blue solid line) by utilizing a variable gain $k(t)$ on the feedback path to scale the feedback gain l_{opt} dynamically. In other words, the overall feedback gain in the reference model is redefined in a time varying form as

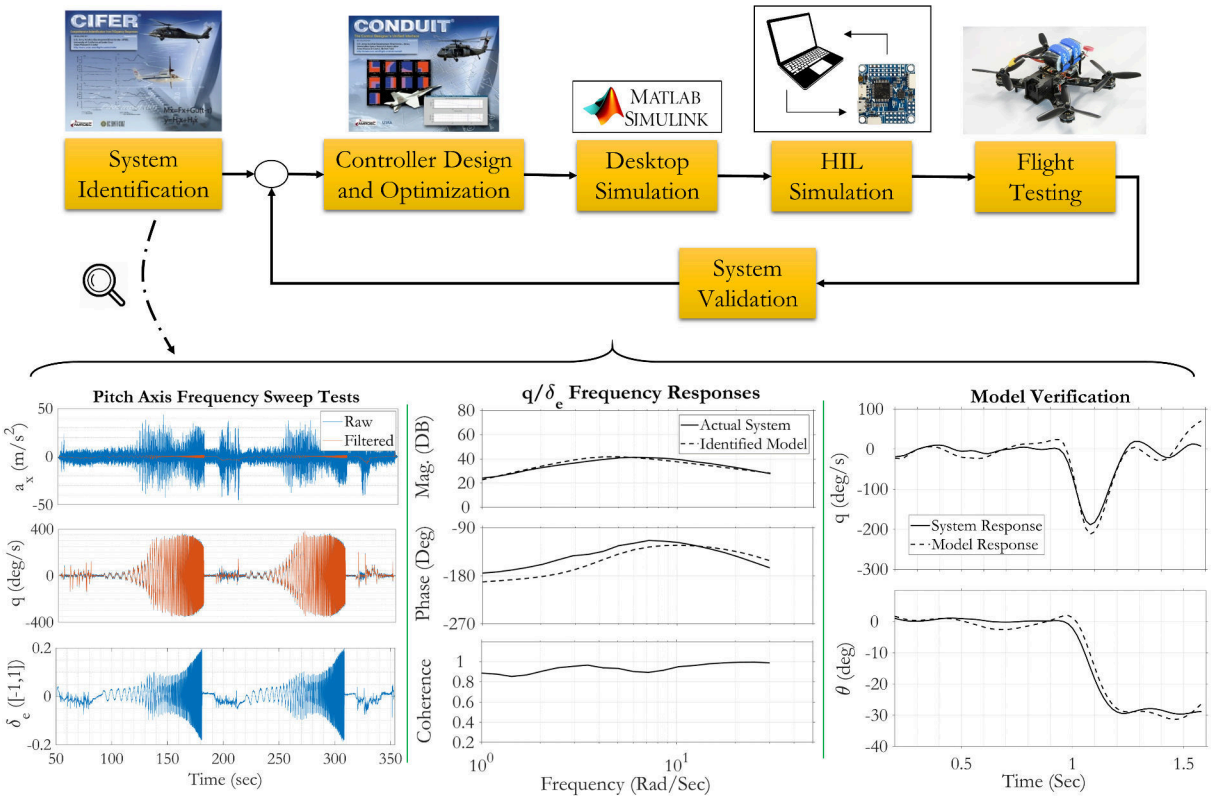

FIGURE 3 General scheme of the desktop-to-flight control system design workflow (adapted from Tischler et al.[17]).

TABLE 1 Transient response comparison of CRM and RL-CRM algorithms on scalar pitch dynamics.

Performance Metrics	CRM (l_{opt})	RL-CRM	Improvement (%)
$\|\dot{\hat{k}}_r\|$	2.6601	2.2130	16.8076
$\|\dot{\hat{k}}_x\|$	1.9200	1.5622	18.6354
$\|\dot{\hat{\theta}}\|$	1.2213	1.0367	15.1150
$\|e\|$	0.1663	0.1383	16.8370
$\|\dot{u}\|$	2.1141	1.7640	16.5602

$l'(t) = k(t)\, l_{opt}$. The time history of the actor signal $k(t)$ is presented in Figure 4-b. For an objective evaluation of the CRM and RL-CRM adaptive control systems, a quantitative performance comparison is presented in Table 1, where $\hat{k}_r$ is the estimated reference signal gain, $\hat{k}_x$ is the estimated feedback gain, $\hat{\theta}$ is the estimated vector of unknown parameters, e is the state tracking error, and u is the control signal. In this analysis, the L_2 norms of several key signals are used as the performance metrics to evaluate the transient behavior of the closed-loop system. A lower L_2 norm indicates damped oscillations in the adaptive parameters, system response, and control signal. Here, it can be observed that the variable reference model feedback gain in the RL-CRM adaptive controller provides an improvement of $15 - 18\%$ in the transient response performance metrics of the dynamical system in comparison to the CRM-adaptive controller. Readers may refer to Yuksek[16] for more information about the RL-CRM adaptive controller design and analysis on the scalar pitch dynamics of a transport class helicopter.

The remainder of this paper is organized as follows. Section 2 discusses the MRAC and CRM-adaptive control system design processes, along with the mathematical model structure and transient response performance metrics. Section 3 elaborates on the general scheme of the proposed RL-CRM adaptive control system, the actor-critic agent structure, and the training process. In

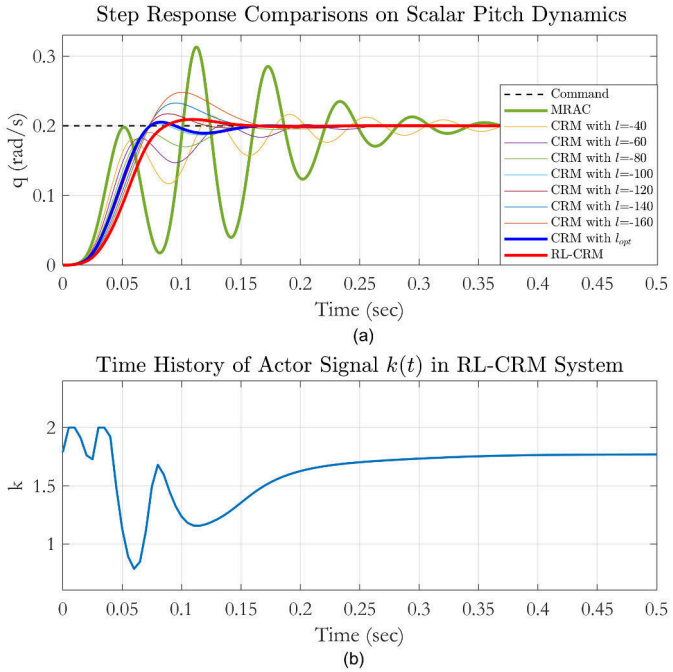

FIGURE 4 a) Step response comparison of MRAC, CRM and RL-CRM adaptive control systems. b) Time history of the actor signal, $k(t)$.

Section 4, utilizing a high-fidelity agile maneuvering quadrotor model and a benchmark example helicopter model, the Monte-Carlo and worst case analysis results are presented and the transient response performances of the MRAC, CRM, and RL-CRM adaptive control systems are compared in terms of the selected key signal norms. The results indicate that the RL-CRM flight control system design not only improves the transient response performance on all associated metrics, but also provides the capability to operate over a wide range of conditions with parametric uncertainties. In Section 5, the concluding remarks and future works are presented.

2 | MRAC AND CRM-ADAPTIVE CONTROL SYSTEM DESIGN

In this section, a general overview of the MRAC and CRM-adaptive control system design processes are described, and performance metrics, which are used in the transient response analysis of the controllers, are explained. Specifically, the transient response characteristics and the optimization process of the design parameters are discussed in detail. In addition, the fundamental theorem toward the CRM-adaptive system design approach is formulated to show the convergence properties.

We consider a linear time-invariant (LTI) mathematical model, given in Equation 1,

$$\begin{aligned} \dot{\mathbf{x}}_p(t) &= \mathbf{A}_p\,\mathbf{x}_p(t) + \mathbf{B}_p\,\Lambda\left(u(t) + f(\mathbf{x}_p)\right) \\ \mathbf{y}_p(t) &= \mathbf{C}_p\,\mathbf{x}_p(t) \end{aligned} \tag{1}$$

where subscript p is used to define the system dynamics. $\mathbf{x}_p(t) \in \mathbb{R}^{n_p}$, $\mathbf{y}_p(t) \in \mathbb{R}^{l_p}$ and $u(t) \in \mathbb{R}^{m_p}$ are the state, measurement and control signal vectors of the dynamic system, respectively. $\mathbf{A}_p \in \mathbb{R}^{n_p \times n_p}$, $\mathbf{B}_p \in \mathbb{R}^{n_p \times m_p}$ and $\mathbf{C}_p \in \mathbb{R}^{l_p \times n_p}$ are the system, control and output matrices, respectively, which are known and constant. $f(\mathbf{x}_p)$ defines the matched uncertainty and it is modeled as shown in Equation 2

$$f(\mathbf{x}_p) = \mathbf{\Theta}^T\,\mathbf{\Phi}(\mathbf{x}_p) \tag{2}$$

where $\mathbf{\Theta} \in \mathbb{R}^{N \times m_p}$ is a constant matrix of unknown coefficients and $\mathbf{\Phi}(\mathbf{x}_p) \in \mathbb{R}^N$ is the known regressor vector. $\Lambda \in \mathbb{R}^{N \times N}$ is used to represent the control failures and modeling errors on the control system effectiveness. Here, it is assumed that Λ is diagonal, its elements are strictly positive, and the $(\mathbf{A}_p,\ \mathbf{B}_p\Lambda)$ pair is controllable. In this study, controller design and analysis are

performed on single-input-single-output (SISO) system dynamics without losing generality to multiple-input-multiple-output (MIMO) formulations. Hence, the input and output dimensions are set as $m_p = 1$ and $l_p = 1$, respectively.

2.1 | Stabilized Reference Model Design

In the MRAC, CRM, and RL-CRM adaptive control algorithms, a stable reference model is required for calculating the control signal. The reference model represents the desired behavior of the dynamical system. The state tracking error signal, which is necessary to derive the adaptation laws, is generated using the states of the reference model.

The reference mathematical model can be obtained by a) assigning a stable and proper mathematical model or b) designing a stabilizing controller for an unstable dynamical model. For our example cases, we considered the LQR control system design approach to obtain a stabilized reference model with the desired reference dynamics. We did not use an assigned or predefined stable reference model to avoid potential actuator and mechanical limitations owing to the unrealistic assigned dynamics. This stabilized system model can be used as a reference model in adaptive controller design processes, as shown in Figure 1. In the stabilized model, an observer-like feedback gain matrix ($\mathbf{L}_v$), utilized in CRM and RL-CRM adaptive control systems, is not used. Hence, the stabilized reference model is often referred to as the *open-loop reference model*.

From a general point of view, the stabilized reference model is used in all MRAC, CRM, and RL-CRM adaptive control system implementations. However, in the MRAC algorithm, it is used directly to derive the adaptation laws without any feedback gain matrix ($\mathbf{L}_v$). In the CRM-adaptive controller, an outer-loop with the feedback gain matrix is added on the stabilized reference model to obtain the "*closed-loop*" reference model. The major difference between the reference models of the RL-CRM and CRM-adaptive control algorithms is that the RL-CRM adaptive control system uses a time-varying feedback gain ($\mathbf{L}'_v(t)$), which is a modified version of the fixed ($\mathbf{L}_v$).

In addition to the derivation of the adaptation laws, the stabilized reference model (i.e., the open-loop reference model) is also used to calculate the true state tracking error ($\mathbf{e}^o(t)$), which will be described in the following sections. The true state tracking error is utilized to generate the observation vector that provides the tracking error data to the RL agent to give insight on the tracking performance.

The LQR control law is given in Equation 3,

$$u_{lqr}(t) = -\mathbf{K}\mathbf{x}_m^o(t) + \bar{N}r(t) \tag{3}$$

where $\mathbf{K} \in \mathbb{R}^{m_p \times n_p}$ is the feedback gain, $\mathbf{x}_m^o(t) \in \mathbb{R}^{n_p}$ is the state vector of the open-loop reference model, $\bar{N} \in \mathbb{R}$ is the feed-forward gain calculated for reference tracking application with the LQR controller, and $r(t) \in \mathbb{R}^{m_p}$ is the reference signal. Superscript o and subscript m are used to indicate that it is the stabilized *reference model*, which is actually the *open-loop* reference model. The stabilized reference model dynamics are given in Equation 4,

$$\begin{aligned} \dot{\mathbf{x}}_m^o(t) &= \left(\mathbf{A}_p - \mathbf{B}_p\,\mathbf{K}\right)\mathbf{x}_m^o(t) + \mathbf{B}_p\,\bar{N}r(t) \\ \mathbf{y}_m^o(t) &= \mathbf{C}_m\,\mathbf{x}_m^o(t) \end{aligned} \tag{4}$$

where $\mathbf{y}_m^o(t) \in \mathbb{R}^{l_p}$ and $\mathbf{C}_m \in \mathbb{R}^{l_p \times n_p}$ denote the output vector and output matrix of the stabilized reference model, respectively. Using these definitions, the state and control matrices of the stabilized reference model ($\mathbf{A}_m$, $\mathbf{B}_m$) can be obtained as

$$\begin{aligned} \mathbf{A}_m &= \mathbf{A}_p - \mathbf{B}_p\,\mathbf{K} \\ \mathbf{B}_m &= \mathbf{B}_p\,\bar{N} \end{aligned} \tag{5}$$

Thus, the state-space representation of the stabilized system model given can be given by

$$\dot{\mathbf{x}}_m^o(t) = \mathbf{A}_m\,\mathbf{x}_m^o(t) + \mathbf{B}_m\,r(t) \tag{6}$$

where $\mathbf{A}_m \in \mathbb{R}^{n_p \times n_p}$ and $\mathbf{B}_m \in \mathbb{R}^{n_p \times m_p}$ are the state and control matrices of the stabilized reference model, respectively. Here, $\mathbf{A}_m$ is a Hurwitz matrix and $r(t)$ is the bounded reference signal.

2.2 | Model Reference Adaptive Control System Design

In the MRAC system, the control input $u(t)$ is designed such that the true state tracking error signal $\mathbf{e}^o(t) \in \mathbb{R}^{n_p}$, given by

$$\mathbf{e}^o(t) = \mathbf{x}_p(t) - \mathbf{x}_m^o(t) \tag{7}$$

globally, uniformly, and asymptotically converges to zero[4];

$$\lim_{t\to\infty} \|\mathbf{e}^o(t)\| = 0 \tag{8}$$

The control law of the MRAC algorithm is given by

$$u_{mrac}(t) = \hat{\mathbf{K}}_x^T\, \mathbf{x}_p(t) + \hat{\mathbf{K}}_r^T\, r(t) - \hat{\boldsymbol{\Theta}}^T \boldsymbol{\Phi}(\mathbf{x}_p) \tag{9}$$

where $\hat{\mathbf{K}}_x \in \mathbb{R}^{n_p \times m_p}$, $\hat{\mathbf{K}}_r \in \mathbb{R}^{m_p \times m_p}$ and $\hat{\boldsymbol{\Theta}} \in \mathbb{R}^{N \times m_p}$ are the estimated feedback gain matrix, feed-forward gain matrix, and estimated unknown coefficients in the matching uncertainty model, respectively. The adaptation laws in the MRAC algorithm, derived using the Lyapunov theory, are given below[4];

$$\begin{aligned}
\dot{\hat{\mathbf{K}}}_x(t) &= -\boldsymbol{\Gamma}_x\, \mathbf{x}_p(t)\, (\mathbf{e}^o(t))^T\, \mathbf{P}\, \mathbf{B}_p \\
\dot{\hat{\mathbf{K}}}_r(t) &= -\boldsymbol{\Gamma}_r\, r(t)\, (\mathbf{e}^o(t))^T\, \mathbf{P}\, \mathbf{B}_p \\
\dot{\hat{\boldsymbol{\Theta}}}(t) &= \boldsymbol{\Gamma}_\Theta\, \boldsymbol{\Phi}(\mathbf{x}_p)\, (\mathbf{e}^o(t))^T\, \mathbf{P}\, \mathbf{B}_p
\end{aligned} \tag{10}$$

where $\boldsymbol{\Gamma}_x \in \mathbb{R}^{n_p \times n_p}$, $\boldsymbol{\Gamma}_r \in \mathbb{R}^{m_p \times m_p}$ and $\boldsymbol{\Gamma}_\Theta \in \mathbb{R}^{N \times N}$ are the adaptation rates and $\mathbf{P} = \mathbf{P}^T > 0$ satisfies the algebraic Lyapunov equation given in Equation 11 for $\mathbf{Q} = \mathbf{Q}^T > 0$;

$$\mathbf{P}\, \mathbf{A}_m + \mathbf{A}_m^T\, \mathbf{P} = -\mathbf{Q} \tag{11}$$

2.3 | CRM Adaptive Control System Design

In the CRM-adaptive control system, an observer-like reference model, which includes a gain matrix $\mathbf{L}_\upsilon$ on the feedback path of the reference model, is used. Hence, this structure is called the *closed-loop reference model*. Mathematical description of the closed-loop reference model is given in Equation 12:

$$\dot{\mathbf{x}}_m(t) = \mathbf{A}_m \mathbf{x}_m(t) + \mathbf{L}_\mathbf{v}\big(\mathbf{x}_p(t) - \mathbf{x}_m(t)\big) + \mathbf{B}_m r(t) \tag{12}$$

where $\mathbf{x}_m \in \mathbb{R}^{n_p}$ is the state vector of the closed-loop reference model, $r \in \mathbb{R}^{m_p}$ is the command signal, and $\mathbf{L}_\upsilon \in \mathbb{R}^{n_p \times n_p}$ is the feedback gain matrix of the reference model tracking error, denoted as

$$\mathbf{e}(t) = \mathbf{x}_p(t) - \mathbf{x}_m(t) \tag{13}$$

$\mathbf{L}_\upsilon$ is parameterized by a scalar $\upsilon > 0$, as given in Lavretsky and Wise[4], and the error feedback gain is selected as

$$\mathbf{L}_\upsilon = \mathbf{P}_\upsilon \mathbf{R}_\upsilon^{-1} \tag{14}$$

where $\mathbf{P}_\upsilon = \mathbf{P}_\upsilon^T > 0$ is the solution of the Algebraic Riccati Equation (ARE):

$$\mathbf{P}_\upsilon \mathbf{A}_m^T + \mathbf{A}_m \mathbf{P}_\upsilon - \mathbf{P}_\upsilon \mathbf{R}_\upsilon^{-1} \mathbf{P}_\upsilon + \mathbf{Q}_\upsilon = 0 \tag{15}$$

whose weight matrices $\mathbf{Q}_\upsilon, \mathbf{R}_\upsilon$ are selected according to Equation 16:

$$\mathbf{Q}_\upsilon = \mathbf{Q}_0 + \frac{\upsilon + 1}{\upsilon} I_{n_p \times n_p};\ \mathbf{R}_\upsilon = \frac{\upsilon}{\upsilon + 1} I_{n_p \times n_p} \tag{16}$$

Here, $\mathbf{Q}_\upsilon \in \mathbb{R}^{n_p \times n_p}$, $\mathbf{Q}_0 \in \mathbb{R}^{n_p \times n_p}$ and $\mathbf{R}_\upsilon \in \mathbb{R}^{n_p \times n_p}$. The constant parameter $\upsilon > 0$ becomes a "tuning knob" in the CRM-adaptive control system design. Low values of υ damp the oscillations in the system response but increase the state tracking error. In contrast, higher values of υ decrease the state tracking error but increase the oscillations in the system response. Hence, there is a trade-off between the damping of the oscillatory response and state tracking error. To obtain a desirable transient response, an optimization process will be performed to calculate a suitable value of υ, which is described in Section 2.3.2.

The adaptation laws for the CRM-adaptive controller were obtained owing to the Lyapunov stability analysis, as given in Equation 17.

$$\begin{aligned}
\dot{\hat{\mathbf{K}}}_x(t) &= -\boldsymbol{\Gamma}_x\, \mathbf{x}_p(t)\, \mathbf{e}^T(t)\, \tilde{\mathbf{P}}_\upsilon\, \mathbf{B}_p \\
\dot{\hat{\mathbf{K}}}_r(t) &= -\boldsymbol{\Gamma}_r\, \mathbf{r}(t)\, \mathbf{e}^T(t)\, \tilde{\mathbf{P}}_\upsilon\, \mathbf{B}_p \\
\dot{\hat{\boldsymbol{\Theta}}}(t) &= -\boldsymbol{\Gamma}_\Theta\, \boldsymbol{\Phi}(\mathbf{x}_p)\, \mathbf{e}^T(t)\, \tilde{\mathbf{P}}_\upsilon\, \mathbf{B}_p
\end{aligned} \tag{17}$$

where $\tilde{\mathbf{P}}_\upsilon = \mathbf{P}_\upsilon^{-1}$ exists for any $\upsilon \geq 0$. Similar to the MRAC, the control signal is given in Equation 18. For more information about the CRM-adaptive system design approach, readers may refer to Lavretsky et al.[4].

$$u_{crm}(t) = \hat{\mathbf{K}}_x^T \mathbf{x}_p(t) + \hat{\mathbf{K}}_r^T \mathbf{r}(t) - \hat{\mathbf{\Theta}}^T \mathbf{\Phi}(\mathbf{x}_p) \tag{18}$$

Theorem 1. Consider the plant dynamics in Equation (1) and the closed-loop reference model dynamics in Equation (12). Given the update laws in Equation (17) and the control law in Equation (18), the tracking error dynamics are globally and asymptotically stable, leading to $\mathbf{x}_m(t) \to \mathbf{x}_m^o(t)$ as $t \to \infty$.

Proof. Stability properties of the CRM-adaptive system can be investigated using a Lyapunov function candidate, as given in Equation 19.

$$V(e, \Delta\mathbf{K}_x, \Delta\mathbf{K}_r, \Delta\mathbf{\Theta}) = \mathbf{e}^T \tilde{\mathbf{P}}_\upsilon \mathbf{e} + tr(\Lambda \Delta\mathbf{K}_x^T \mathbf{\Gamma}_x^{-1} \Delta\mathbf{K}_x) + tr(\Lambda \Delta\mathbf{K}_r^T \mathbf{\Gamma}_r^{-1} \Delta\mathbf{K}_r) + tr(\Lambda \Delta\mathbf{\Theta}^T \mathbf{\Gamma}_\Theta^{-1} \Delta\mathbf{\Theta}) \tag{19}$$

where *tr* refers to the trace operator. By using the adaptive laws given in Equation 17, the time derivative of the Lyapunov function candidate can be calculated as given in Equation 20:

$$\dot{V}(e, \Delta\mathbf{K}_x, \Delta\mathbf{K}_r, \Delta\mathbf{\Theta}) = -\mathbf{e}^T (\mathbf{R}_\upsilon^{-1} + \tilde{\mathbf{P}}_\upsilon \mathbf{Q}_\upsilon \tilde{\mathbf{P}}_\upsilon)\, \mathbf{e} \leq 0 \tag{20}$$

Because $\dot{V} \leq 0$ in Equation 20, the state tracking error and adaptation parameters are bounded in time. Additionally, the second derivative of the Lyapunov function is;

$$\ddot{V}(e, \Delta\mathbf{K}_x, \Delta\mathbf{K}_r, \Delta\mathbf{\Theta}) = -2\mathbf{e}^T (\mathbf{R}_\upsilon^{-1} + \tilde{\mathbf{P}}_\upsilon \mathbf{Q}_\upsilon \tilde{\mathbf{P}}_\upsilon)\, \dot{\mathbf{e}} \in L_\infty \tag{21}$$

this implies that the second derivative of the Lyapunov function is bounded. According to these observations, the Lyapunov function is lower bounded and has a non-increasing time derivative. Hence, the Lyapunov function tends to a limit as time reaches infinity. Additionally, the second time derivative of the Lyapunov function is uniformly bounded. Hence, $\dot{V}$ is a uniformly continuous function of time, according to Barbalat's Lemma[4], and it tends to zero as time reaches infinity. If $\dot{V}$ tends to zero, the state tracking error tends to zero according to Equation 20; thus,

$$\lim_{t \to \infty} \|\mathbf{e}(t)\| = 0 \tag{22}$$

which shows the global asymptotic stability of the state tracking error $\mathbf{e}$ in the CRM-adaptive control system. This also implies that $\mathbf{x}_m(t) \to \mathbf{x}_m^o(t)$ as $t \to \infty$. According to this result, it is important to state that the closed-loop reference model asymptotically converges to the stabilized reference model, which is the true mathematical model of the desired response[8,4]. □

2.3.1 | Transient Response Characteristics

The main improvement of the CRM and RL-CRM adaptive control systems is observed on the transient response performance of the controlled system. The high-frequency oscillations on the adaptive parameters, control signal and system response are suppressed in the transient phase using these control algorithms. However, it becomes an important issue to define the adequate, tolerable, and unacceptable behaviors. Hence, the characterization of the transient response has a crucial role in the controller design and evaluation processes.

Several performance metrics were introduced by Gibson et al.[15,8], such as $\|\mathbf{e}(t)\|$, $\|\mathbf{e}^o(t)\|$, $\|\dot{u}(t)\|$, $\|\dot{\hat{\mathbf{\Theta}}}(t)\|$, and $\|\mathbf{y}_m(t)\|_\infty$. Here, $\mathbf{e}(t) \in \mathbb{R}^{n_p}$ is the state tracking error between the system response and closed-loop reference model response, as given in Equation 13. In addition, $\mathbf{e}^o(t) \in \mathbb{R}^{n_p}$ is the true state tracking error between the system response and stabilized reference model response, as given in Equation 7, $\dot{u}(t) \in \mathbb{R}^{m_p}$ is the time derivative of the control signal, $\dot{\hat{\mathbf{\Theta}}}(t) \in \mathbb{R}^{N \times m_p}$ is the time derivative of the adaptive parameter matrix and $\mathbf{y}_m(t) \in \mathbb{R}^{l_p}$ is the closed-loop reference model output vector. These metrics directly provide quantitative information about the oscillatory transient behavior and peak response of the system and the reference model. They are used for the optimal selection of the reference model tracking error feedback gain matrix ($\mathbf{L}_\upsilon$) in the CRM-adaptive control design process and evaluating the transient response performance of the MRAC, CRM, and RL-CRM adaptive control systems.

2.3.2 | Selection of Design Parameter υ

CRM-adaptive control systems provide improved transient response performance in comparison to MRAC. This improvement is directly observed in the L_2 norm of the derivatives of the adaptation parameters, the reference tracking error, and the derivative

of the control signal. Smaller values of υ damp the oscillations on the adaptive parameters and control signal. However, they increase the open-loop reference model tracking error and peak value of the system response. In other words, there is a trade-off between the improved transient response and open-loop reference model tracking performance.

This trade-off is explained by applying the CRM-adaptive controller design process on the longitudinal dynamics of the quadrotor platform, which is described in Section 4. A sweep analysis is performed to evaluate several performance metrics for the interval $\upsilon \in [\upsilon_{min}, \upsilon_{max}]$; the results are illustrated in Figure 5. Here, υ_{min} and υ_{max} are used to denote the lower and upper limits of the design parameter υ, respectively. According to these results, it can be observed that the L_2 norm of most of the adaptation parameters and the control signal derivative decrease as υ decreases; this implies the transient response oscillations are damped. However, the L_∞ norm of the system response and L_2 norm of the true output error increase at the cost of the damped transient oscillations. It should be noted that as the design parameter υ tends to infinity, the effects of the feedback gain matrix $\mathbf{L}_\upsilon$ disappear, oscillations increase, and the CRM-adaptive controller behaves like MRAC.

The water-bed effect, which is a result of the inappropriate selection of the adaptation parameters ($\mathbf{\Gamma}_x, \mathbf{\Gamma}_r, \mathbf{\Gamma}_\theta$) and observer gain ($\mathbf{L}_\upsilon$), is another phenomenon that should be considered in the design phase [15]. It can be defined as undesirable jumps in several signals such as $|\dot{u}(t)|$ and $|e^o(t)|$, which may result in actuator rate and actuator saturation. Hence, the water-bed effect degrades the system performance and poses a considerable threat to the acceptable flight envelope and operational safety.

Based on these observations about the water-bed effect and the trade-off between the improved transient dynamics and model following error, formulating a proper optimization problem to obtain the optimal value of the design parameter υ becomes critical for the CRM-adaptive control system. For example, consider the sweep analysis of the transient response characteristic norms of the example model from Section 4. Figure 5 indicates that the optimization problem should include several performance metrics such as $\|\mathbf{e}^o(t)\|$, $\|\dot{u}(t)\|$ and $\|\mathbf{y}_m(t)\|_\infty$, which reflect the fundamental characteristics of the trade-off between the fast convergence in the tracking error ($\mathbf{e}$) and the true tracking error, ($\mathbf{e}^o$). Owing to these requirements, the optimization problem can be formulated as given in Equation 23.

$$\begin{aligned} &\min \ \ J(\mathbf{e}^o, \mathbf{y}_m, \dot{u}) \\ &\text{s.t.} \ \ t \in [t_0, t_f] \\ &\qquad \upsilon \in [\upsilon_{min}, \upsilon_{max}] \end{aligned} \tag{23}$$

where J is the cost function, given by Equation 24.

$$J(\mathbf{e}^o, y_m, \dot{u}) = w_1\|\mathbf{e}^o(t,\upsilon)\| + w_2\|\mathbf{y}_m(t,\upsilon)\|_\infty + w_3\|\dot{u}(t,\upsilon)\| \tag{24}$$

where $t \in [t_0, t_f]$ defines the simulation time interval that includes the transient response dynamics and $w_i \in \mathbb{R}^+, i = \{1,2,3\}$ correspond to the weights spanning the Pareto-optimal frontier.

After defining the cost function, constraints, and bounds, the optimization problem is solved using an interior-point algorithm in MATLAB®. Owing to the optimization process, the optimal value of the design parameter (υ_{opt}) is obtained as 0.5 (also illustrated in Figure 5) for the example model from Section 4 for an initial selection of unity weights.

As discussed in Section 4, introducing a variable parameter υ minimizes the water-bed effect observed in the $|\dot{u}(t)|$ and $|e^o(t)|$ signals. For more information about the water-bed effect in the CRM-adaptive systems, readers may refer to Gibson et al. [15].

3 | RL-CRM CONTROL SYSTEM DESIGN

As detailed in Section 1, the adaptation time of the controlled system can be decreased by increasing the learning rates of MRAC. However, this increase is observed at the expense of increasing oscillations in the system response and the control signal. This is a particularly undesirable situation in aerospace applications because such oscillations can result in operationally unsafe inertial and aerodynamic configurations with high angle of attack and roll angles. To handle this problem, the CRM-adaptive control algorithm is developed with the ability to improve the transient performance of the closed-loop adaptive systems by utilizing a feedback loop in the reference model. The value of the design parameter υ_{opt} is selected by following an optimization process in which the key performance metrics are included in the cost function to capture the transient behavior. Although this approach provides an effective solution for damping the transient oscillations in the system response and control signal, this is achieved at the expense of increased true output tracking error and peak response magnitude despite the use of the optimal design parameter. This situation and the associated trade-off is also illustrated in $\|\mathbf{e}^o(t)\|$ and $\|y_p(t)\|_\infty$ in Figure 5.

In this section, to overcome this problem, we introduce a novel adaptive control algorithm by designing and embedding a variable parameter within the CRM-adaptive control system through machine learning. Specifically, the approach leads to an

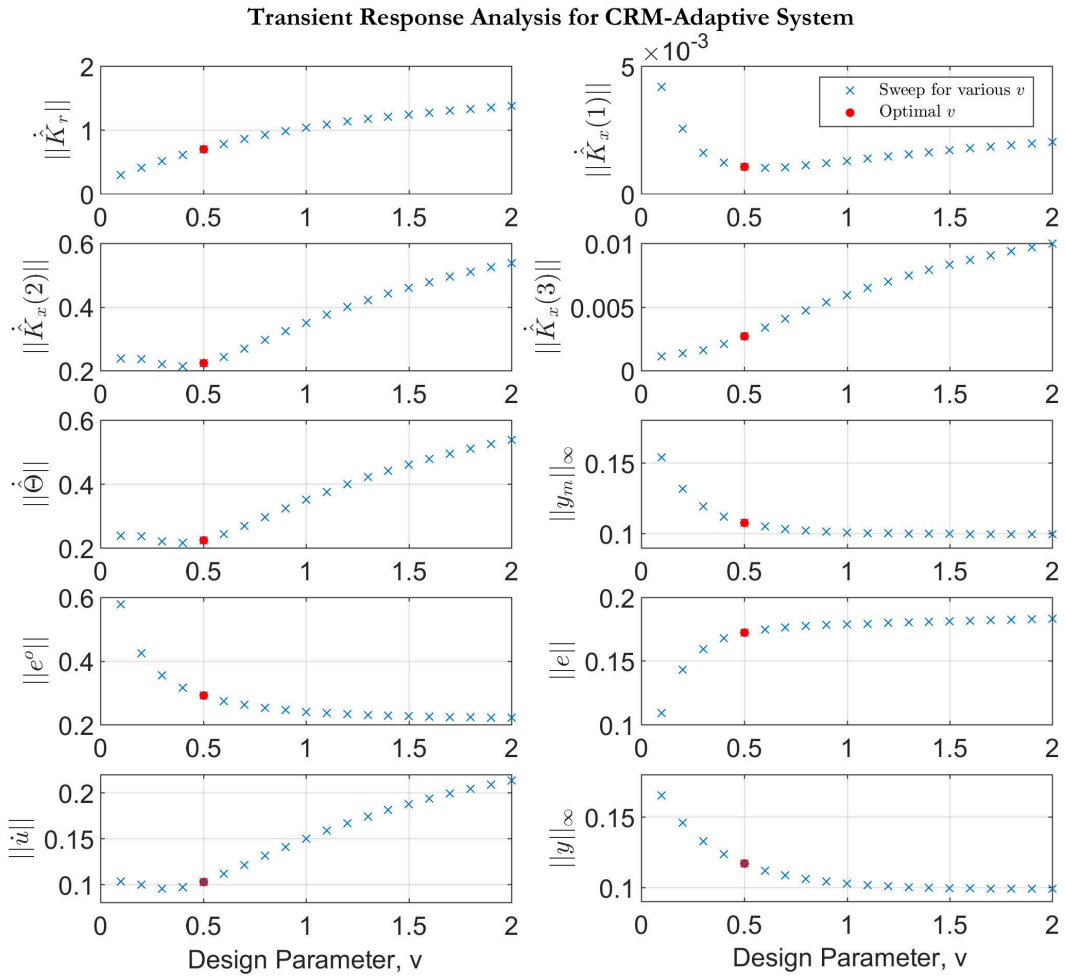

FIGURE 5 Transient response analysis for different υ.

intelligent method to combine the MRAC and CRM-adaptive systems using the variable $\upsilon_{opt}k(t)$. Higher values of $\upsilon_{opt}k(t)$ decrease the magnitude of the reference model feedback matrix $\mathbf{L}_\upsilon$ and the CRM-adaptive controller behaves like MRAC. The design parameter is scaled up or down by utilizing an actor-critic RL agent. Hence, the proposed algorithm is called RL-CRM adaptive control system. In this control system, a training process must be performed using a specific learning algorithm to train the RL agent. Observation vector and reward from the environment are used in the learning process to optimize the weights and biases in the actor-critic structure.

Mathematical description of the proposed closed-loop reference model with the variable feedback gain matrix $\mathbf{L}'_\upsilon(t)$ is given in Equation 25:

$$\dot{\mathbf{x}}_m(t) = \mathbf{A}_m\mathbf{x}_m(t) + \mathbf{L}'_\upsilon(t)\big(\mathbf{x}_p(t) - \mathbf{x}_m(t)\big) + \mathbf{B}_m r(t) \tag{25}$$

where the variable observer gain matrix $\mathbf{L}'_\upsilon(t)$ and weight matrix $\mathbf{R}'_\upsilon(t)$ correspond to

$$\mathbf{L}'_\upsilon(t) = \mathbf{P}_\upsilon\big(\mathbf{R}'_\upsilon(t)\big)^{-1} \tag{26}$$

$$\mathbf{R}'_\upsilon(t) = \frac{\upsilon'(t)}{\upsilon'(t)+1} I_{n_p \times n_p} \tag{27}$$

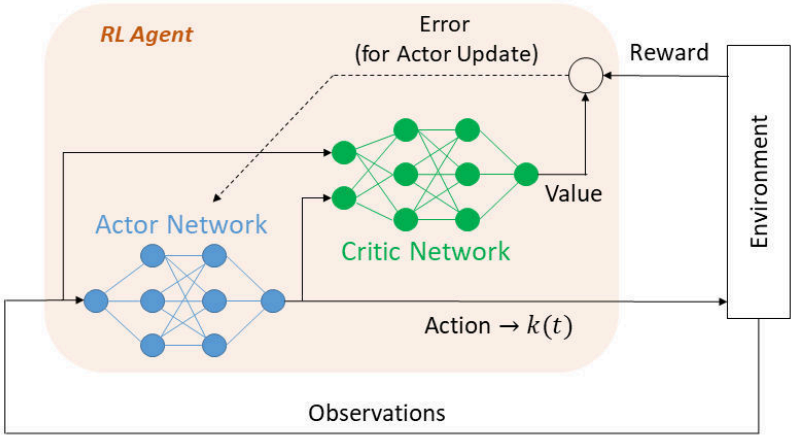

FIGURE 6 Actor-Critic agent structure.

respectively. Here, $v'(t)$ is the variable design parameter, which is given in Equation 28. v_{opt} is the CRM optimized design parameter and $k(t)$ is a variable scaling factor determined by the RL agent. Thus, $(')$ is used to indicate the time-varying properties of the design parameter and matrices, leading to

$$v'(t) = v_{opt}\, k(t) \tag{28}$$

The general structure of the proposed RL-CRM adaptive control system is illustrated in Figure 1. In this algorithm, an RL agent with an actor-critic structure is used to determine the scaling factor $k(t)$ to increase or decrease the magnitude of the design parameter. The RL agent is trained using the DDPG algorithm proposed by Lillicrap et al[20]. DDPG is a model-free, off-policy actor-critic algorithm that uses function approximators to learn policies in high-dimensional and continuous action spaces.

The actor-critic agent structure includes two neural networks called *actor* and *critic*. The actor component generates action (i.e., control policy) using observation signals and applies it to the system. Then, the critic component compares the actual and estimated values of the reward and quantifies the optimality of the action. A general actor-critic structure is illustrated in Figure 6. The hyperparameters of the actor-critic agent include the number of layers, number of nodes, gradient threshold, and leaning rate, which define the neural network architecture. The hyperparameters of the actor and critic structures are selected by evaluating several training results; they are summarized in Table 2. In addition to the progressive tuning that we used, several methods could be used to optimize the hyperparameters of the RL agent such as random search[21] and Bayesian optimization[22].

After generating a proper actor-critic agent structure, it is important to define the training process parameters for the DDPG algorithm, such as the learning rate, target smooth factor, discount factor, mini batch size and buffer length. Similar to the selection of the hyperparameters of the actor-critic structure, the training parameter values are adjusted progressively in which the performance of numerous training results across a wide parameter range. Adequate values of the training parameters are set after several training sessions by evaluating the average reward and expected long-term reward which is estimated by the critic network. Also, it is important to note that the training performance is quite sensitive to the training parameters. For example, if a large sample time is selected to accelerate the simulation, there may not be enough data to train the agent and fast system dynamics may not be captured in the simulation environment. This will result in an poorly trained RL-agent with improper weights in the actor-critic networks. Similar issue could be observed in selection of the buffer length which is directly related with the experience replay. The experience replay could improve the stability and efficiency of the training process by providing uncorrelated data. It seems that the learning process is robust against different size of replay buffer capacity. However, recent studies show that the agent is quite sensitive to the size of the replay buffer length and novel methods are developed to suppress the negative effects of the large replay buffer[23]. A parameter set suitable for our application is presented in Table 3; specific values are listed to ensure reproducibility of the results.

In the training phase, it is critical to define a suitable reward function ($R(t)$). The RL agent learns the action policy to increase the reward value, which directly represents the performance of the system based on the tracking error, peak value of the system response, and other performance metrics. The reward function design is related to the desired response characteristics and this function directly affects the system performance.

TABLE 2 Hyperparameters of Actor-Critic agent.

Network	Parameter	Value
Actor	Number of Hidden Layers	1
	Number of Nodes in Hidden Layers	10
	Activation Functions	Tanh
	Learning Rate	0.002
	Gradient Threshold	1
Critic	Number of Obs. Path Hidden Layers	2
	Number of Nodes in Obs. Path Hidden Layers	10
	Number of Action Path Hidden Layers	1
	Number of Nodes in Action Path Hidden Layers	10
	Activation Functions	Tanh
	Learning Rate	0.002
	Gradient Threshold	1

TABLE 3 Training parameters for DDPG algorithm.

Parameter	Value
Sample Time	0.005
Target Smooth Factor	0.001
Discount Factor	0.99
Mini-Batch Size	1024
Buffer Length	1E6

As discussed in Section 1, the main purpose of the RL-CRM adaptive control algorithm is to further improve the transient response performance of the dynamical system. To achieve this, several performance metrics, based on the system response, tracking error and control signal, are introduced to quantify the transient response. Subsequently, the optimal value of the design parameter is obtained using an optimization process in which the aforementioned signals are used in the cost function. These signals are also used in the reward function to evaluate and improve the system performance following a similar optimization process as that discussed in Section 2. Accordingly, the reward function is designed as given in Equation 29. The reward function includes the peak response, output tracking error, derivative of the control signal, command tracking error, and true output tracking error.

$$R(t) = w_1 R_p(t) + w_2 R_{e_y}(t) + w_3 R_u(t) + w_4 R_{e_{cmd}}(t) + w_5 R_o(t) \tag{29}$$

where $R_p(t)$, $R_e(t)$, $R_u(t)$, $R_{e_{cmd}}(t)$, $R_o(t)$, and w_i are defined below for our specific implementation;

$$R_p(t) = \begin{cases} -1, & \text{if } \|y_p(t)\|_\infty \geq 0.105 \\ 0, & \text{otherwise} \end{cases} \tag{30}$$

$$R_{e_y}(t) = \begin{cases} 4, & \text{if } |e_y(t)| \leq 0.0005 \\ 0, & \text{otherwise} \end{cases} \tag{31}$$

$$R_u(t) = \begin{cases} 2, & \text{if } |\dot{u}(t)| \leq 0.02 \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

$$R_{e_{cmd}}(t) = \begin{cases} 2, & \text{if } |e_{y_{cmd}}(t)| \leq 0.01 \text{ and } t \geq 0.3 \text{ sec} \\ 0, & \text{otherwise} \end{cases} \tag{33}$$

$$R_o(t) = \begin{cases} 1, & \text{if } |e_y^o(t)| \leq 0.02 \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

$$w_i = 1 \quad \forall i \in \{1, 2, 3, 4, 5\} \tag{35}$$

for our specific implementation. Here, $w_i \in \mathbb{R}^+$ corresponds to the weights spanning the Pareto-optimal frontier. The specific function forms and the bounding values are listed to ensure reproducibility of the results. Here, $e_y(t) \in \mathbb{R}$, $e_y^o(t) \in \mathbb{R}$, and $e_{y_{cmd}}(t) \in \mathbb{R}$ correspond to the reference model output tracking error, true output tracking error, and command signal tracking error, denoted by Equation 36, 37 and 38, respectively.

$$e_y(t) = y_p(t) - y_m(t) \tag{36}$$

$$e_y^o(t) = y_p(t) - y_m^o(t) \tag{37}$$

$$e_{y_{cmd}}(t) = y_p(t) - r(t) \tag{38}$$

where $y_m(t) \in \mathbb{R}$ is the closed-loop reference model output, $y_m^o(t) \in \mathbb{R}$ is the open-loop reference model output, and $r(t) \in \mathbb{R}$ is the command signal. In the reward function, $R_p(t)$ is used to minimize the peak response of the system which is related with L_∞ norm of the output signal, $R_{e_y}(t)$ and $R_u(t)$ are used to bound the reference model tracking error and time derivative of the control signal, respectively, $R_{e_{cmd}}(t)$ is used to specify the settling time boundary for the step response, and $R_o(t)$ is used to minimize the true output tracking error ($e_y^o(t)$). Here, these reward functions are selected in such structures (i.e. band pass functions centered around zero with absolute value boundaries) not only to shape the step response (in terms of output peak value and settling time) of the closed-loop system but also to suppress the water-bed effect that occurs in time-derivative of the control signal. The boundaries in the reward functions correspond to the application-specific soft limits on peaks associated with output errors and input rates acceptable as air vehicle's dynamic behaviour. Also, the weights (w_i) define the relative importance of each separate reward in the total reward function $R(t)$. Selection of these weights is directly related to desired performance of the closed-loop system. For example, if it is required to minimize the peak amplitude of the output signal and slow down the system response, this can be achieved by increasing the weights of the $R_p(t)$ and $R_{e_{cmd}}(t)$ after setting the adequate time boundary in $R_{e_{cmd}}(t)$. Conditional reward values in Equations 31-34 are also related with the relative importance of the each reward function. Although they can be adjusted according to the desired performance of the closed-loop system, we fixed them and used the weights within the reward functions to define the relative importance of each reward and, consequently, to achieve the required closed-loop system behavior. In RL applications, designing an observation vector is a mission specific task. It is important to define the required signals that contain the key observations about the signal tracking performance of the closed-loop system. In addition, the system performance is directly related to the data included in the observation vector. If the agent can correctly observe the key characteristics of the closed-loop system, it can generate a suitable and desired action signal.

Several simulation studies on the MRAC algorithms elucidate that the magnitude of the true output tracking error ($e_y^o(t)$) is one of the main dominant parameters affecting the transient behavior of the adaptive control system. The reason behind that relationship directly originates from the adaptation laws of the MRAC given in Equation 10. Here, it is evident that the high values of ($e_y^o(t)$) result in high time derivatives of the adaptation parameters, which cause oscillations in the control signal and system response. Hence, the true output tracking error, its derivative, and its integral are included in the observation vector to represent the true error dynamics. Thus, the observation vector is formulated as

$$\mathbf{O}\left(e_y^o(t)\right) = \left[e_y^o(t),\ \dot{e}_y^o(t),\ \int e_y^o(t)\, dt\right]^T \tag{39}$$

The observation vector has a critical role in the RL-CRM algorithm because it provides information about how far the dynamical system response is from the stabilized reference model (i.e., the open-loop reference model) response which is the actual desired dynamical behavior. The RL-agent uses these data to evaluate and create suitable action signals in the actor-critic network structure. In the next section, we discuss the simulation results on benchmark examples, by comparing the performance of the MRAC, CRM, and RL-CRM adaptive control systems.

4 | SIMULATION RESULTS

In this section, we provide a high-fidelity agile maneuvering quadrotor model and a benchmark example helicopter model to compare the performances of MRAC, CRM, and RL-CRM. Specifically, transient response performance of each control algorithm is evaluated using the Monte-Carlo and worst-case analyses in the presence of parametric uncertainties.

Linearized equations of motion for the pitch dynamics of the agile maneuvering quadrotor are given below;

$$\begin{aligned}\dot{u}(t) &= X_u\, u(t) + X_q\, q(t) + X_{\delta_e}\delta_e(t)\\ \dot{q}(t) &= M_u\, u(t) + M_q\, q(t) + M_{\delta_e}\delta_e(t)\\ \dot{\delta}_e(t) &= -\omega_e\, \delta_e(t) + \omega_e\, \delta_{e_{cmd}}(t)\end{aligned} \tag{40}$$

where $u(t)$, $q(t)$ and $\delta_e(t)$ are the X_b component of the body axis velocity, pitch rate, and mixer input on pitch axis, respectively. $\delta_{e_{cmd}}(t)$ is the mixer input, and X_u, X_q, X_{δ_e}, M_u, M_q and M_{δ_e} are the aerodynamic control and stability derivatives. ω_e corresponds to the natural frequency of the BLDC motor. Equation 40 can be rewritten in the state-space form as

$$\begin{aligned}\begin{bmatrix}\dot{u}\\ \dot{q}\\ \dot{\delta}_e\end{bmatrix} &= \begin{bmatrix}X_u & X_q & 0\\ M_u & M_q & M_{\delta_e}\\ 0 & 0 & -\omega_e\end{bmatrix}\begin{bmatrix}u\\ q\\ \delta_e\end{bmatrix} + \begin{bmatrix}0\\ 0\\ \omega_e\end{bmatrix}\Lambda(\delta_{e_{cmd}} + f(\mathbf{x}_p))\\ y_p &= \begin{bmatrix}0 & 1 & 0\end{bmatrix}\begin{bmatrix}u\\ q\\ \delta_e\end{bmatrix}\end{aligned} \tag{41}$$

Here, the parametric uncertainties are modelled as matched uncertainties and integrated in the system model as $f(\mathbf{x}_p)$. In this study, the time-delay of the pitch dynamics (τ_{δ_e}), which represents the high-frequency unmodeled dynamics, is neglected to simplify the problem.

Aerodynamic control and stability derivatives, and BLDC motor dynamics of the quadrotor test platform are obtained using the frequency-domain system identification method in the CIFER tool[18]. In this process, the frequency responses of the aerial vehicle are obtained via frequency sweep tests in which the variable frequency *sine* signal is applied to the related control channel. Then, the parametric models are fitted on the frequency responses. The identified mathematical models are verified in the time-domain using doublet signals. The pitch axis system identification results of the quadrotor platform are presented in Table 4. Here, the Cramer-Rao (CR) Bound defines the uncertainty of the aerodynamic parameter; as a design rule-of-thumb the percentage of the CR-Bound should be less than 25%. Insensitivity gives insight regarding the dominance of the related parameter on the system dynamics, which should be less than 10% as a design rule-of-thumb. High insensitivity means that the related aerodynamic parameter has minimal effect on the system dynamics and it can be neglected. While verifying the identified model, a doublet attitude command signal is applied to the actual system and the mixer signal is logged onboard. Then, the logged mixer signal is applied to the identified linear model and the responses are compared, as shown in Figure 7. It is evident that the identified linear model can capture the pitch dynamics of the actual system. For more information about system identification, verification, and classical control system design process of the quadrotor platform, readers may refer to Yuksek et al.[19]. According to Table 4, the pitch damping derivative M_q, which is one of the most effective aerodynamic parameter in flight dynamics, has a CR-Bound and insensitivity level of 20.09 % and 5.825 %, respectively. Accordingly, the matched uncertainty is defined for this parameter and modeled in the $f(\mathbf{x}_p)$ vector. From the perspective of resilience, this is important for evaluating the adaptation and reconfiguration capability (i.e., resilience capability) of the closed-loop system in the presence of variations in the mass and the center of gravity of the vehicle.

The step responses of the closed-loop systems with MRAC, CRM, and RL-CRM adaptive controllers are illustrated in Figure 8. Here, it can be observed that MRAC results in a high-frequency oscillatory system response in the transient phase, which affects the flight experience and operational safety. This transient response oscillation is damped by utilizing the fixed-gain CRM-adaptive controller. A sweep analysis is performed for the interval $\upsilon \in [0.2, 1.2]$ and the transient responses for these cases are illustrated in Figure 8. The system response corresponding to the optimal design parameter υ_{opt} for the CRM-adaptive system, which was evaluated to be 0.5 in Section 2.3, is represented by blue solid line. It is evident that the CRM-adaptive controller with the optimal design parameter provides better transient response in comparison to MRAC. The step response of the closed-loop system with the proposed RL-CRM adaptive controller is also depicted in Figure 8 as a red solid line. As shown in this figure, the RL-CRM adaptive control system exhibits superior performance in comparison to MRAC and fixed-gain CRM adaptive systems in terms of the settling time and overshoot. This improvement is obtained by utilizing the variable scaling factor (action signal) $k(t)$ in the closed-loop reference model. The time history of the action signal, generated by the RL agent, is illustrated in Figure 9. Here, the time-varying adaptation strategy switches the system focus from overshoot minimization to error minimization as the state evolves.

As detailed in Sections 1 and 3, the water-bed effect is an important phenomenon that may occur in CRM-adaptive control systems if the adaptation gains and closed-loop model feedback matrix are selected improperly. Instantaneous jumps may occur

TABLE 4 Identified model parameters of the racer quadrotor test platform.

Aerodynamic Derivative	Value	CR Bound	CR Bound (%)	Insensitivity (%)
X_u	-0.2586	0.01812	7.007	2.274
X_q	-0.07132	9.992E-03	14.01	5.033
M_u	5.688	0.3699	6.503	1.724
M_q	1.958	0.3935	20.09	5.825
X_{δ_e}	-9.124	0.4095	4.488	1.6
M_{δ_e}	765.7	26.51	3.462	0.9258
ω_e^*	29	-	-	-
τ_{δ_e}	0.03368	2.68E-03	7.958	3.218

*Fixed in the state-space model identification process.

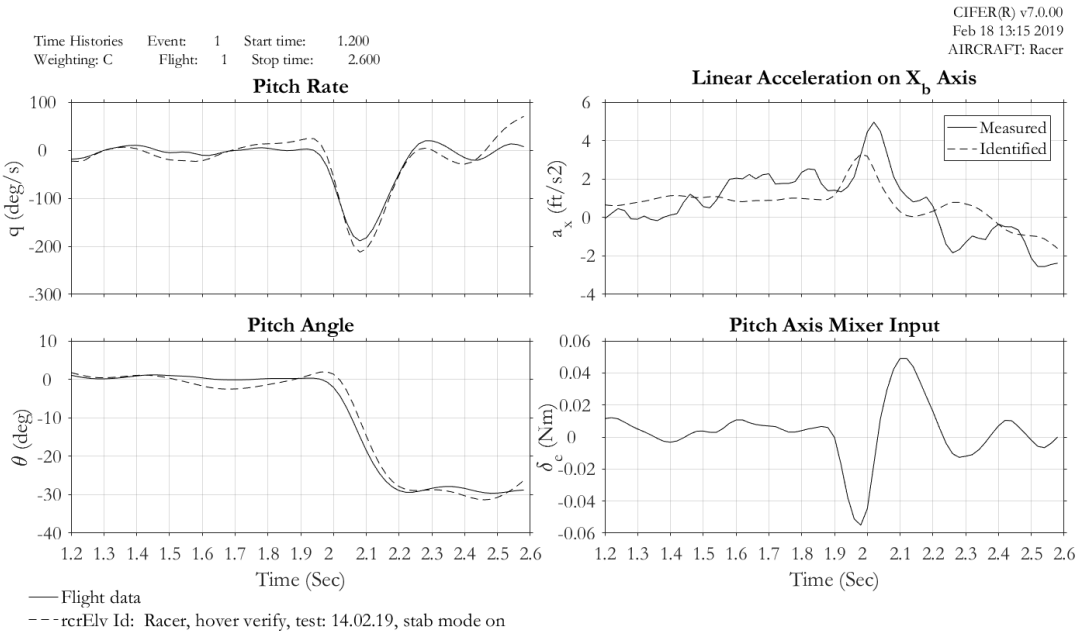

FIGURE 7 Longitudinal model verification flight test results.

in several signals such as $|e_y^o(t)|$ and $|\dot{u}(t)|$. Figure 10, demonstrates the time history of the $|e_y^o(t)|$ and $|\dot{u}(t)|$ signals of the CRM and RL-CRM systems, which provides insight on the water-bed effect. In this figure, the response of the CRM-adaptive system with the optimal reference model feedback gain matrix is indicated by the blue solid line. The response of the RL-CRM adaptive system is represented by red solid line. Here, it is evident that the RL-CRM adaptive system minimizes the water-bed effect and damps the undesirable jumps in the $|e_y^o(t)|$ and $|\dot{u}(t)|$ signals. Absorbing the water-bed effect in the $\dot{u}(t)$ signal is especially important to avoid the actuator rate and position limits. In a multicopter platform, these limits refer to the time constant and maximum angular velocity of the BLDC motor.

Monte-Carlo Analysis

To perform comparative benchmarking over the worst cases, we utilized the pitch dynamics of a generic helicopter mathematical model used by Lavretsky et al.[4]. The benchmark example corresponds to the simplified pitch dynamics of a helicopter in hover flight, captured using the pitch rate ($q(t)$) and longitudinal control input ($\delta_e(t)$). We assume that the forward and vertical speed components of the helicopter are significantly small in the hover flight phase. Hence, the aerodynamic speed derivatives such as

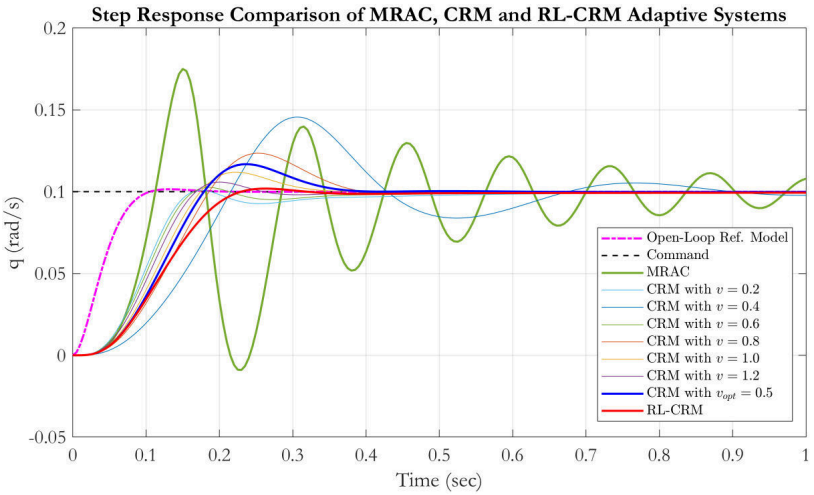

FIGURE 8 Comparison of state-space model responses with MRAC, CRM, and RL-CRM adaptive controllers.

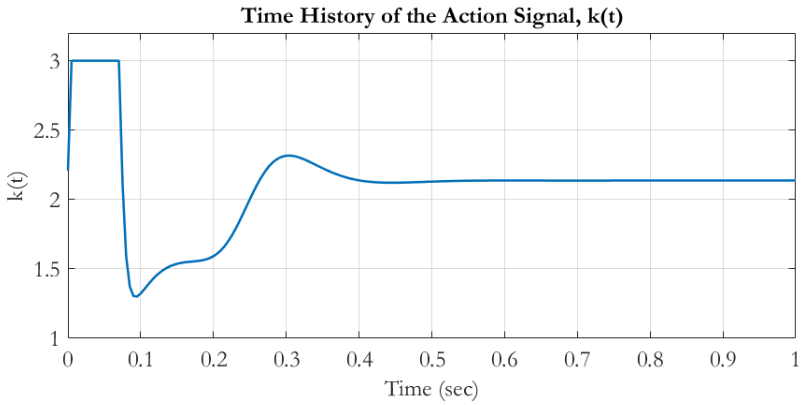

FIGURE 9 Action signal of the RL-CRM adaptive controller.

M_u, M_w, X_u, and X_w are neglected to simplify the mathematical model. Correspondingly, the pitch dynamics of the helicopter can be modelled as a scalar system, given by

$$\dot{q}(t) = M_q\, q(t) + M_{\delta_e}\left(\delta_e(t) + f(q)\right) \tag{42}$$

where M_q is the pitch damping derivative and M_{δ_e} is the elevator effectiveness (i.e., the longitudinal control power). Pitch dynamics also includes $f(q)$ (given in Equation 43), which represents the matched system uncertainties as a function of pitch rate and introduces instability into the open-loop dynamics. This is further illustrated by

$$f(q) = -0.01\, tanh\left(\frac{360}{\pi}\, q(t)\right) = \theta^T\, \Phi(q) \tag{43}$$

where $\theta \in \mathbb{R}^N$ is an unknown parameter and $\Phi(q) \in \mathbb{R}^N$ is the known regressor. This mathematical model represents a locally unstable open-loop dynamics at the origin (i.e., $q = 0$).

For the transport class helicopter, the aerodynamic parameters are set to be $M_q = -0.61\, rad/s$ and $M_{\delta_e} = -6.65\, rad/s^2$ according to Lavretsky et al.[4]. To evaluate the fragility of the control systems, the worst case scenario is evaluated with -35% parametric uncertainties on both M_q and M_{δ_e}. The $|e_y^o(t)|$ and $|\dot{u}(t)|$ responses of the CRM and RL-CRM sytems are shown in Figure 11 for this case. Here, it is evident that the proposed RL-CRM adaptive system exhibits better performance than the CRM-adaptive system even in the worst case scenario with -35% parametric uncertainty. However, when the system is forced above an uncertainty level of -35%, the level of oscillations on $|\dot{u}(t)|$ response increase which increases the L_2 norm of this

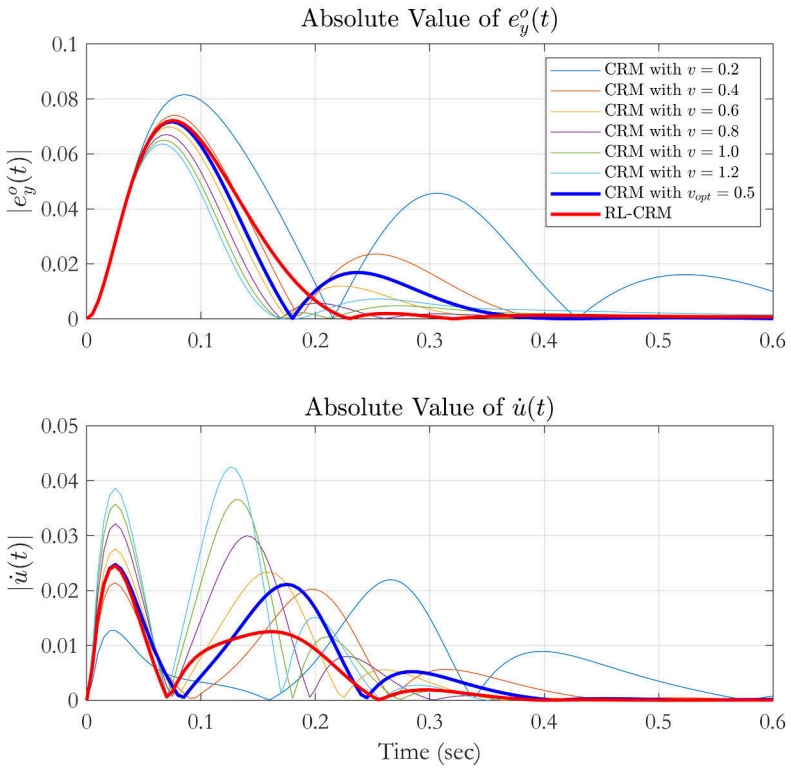

FIGURE 10 Water-bed effect comparison of CRM and RL-CRM adaptive systems.

TABLE 5 500-Run Monte-Carlo analysis results of the MRAC, CRM and RL-CRM adaptive systems.

Performance Metrics	MRAC	CRM	Improvement (%)	RL-CRM	Improvement (%)
$\overline{\|\dot{\hat{K}}_x\|}$	15.2114	3.7341	75.4520	2.4489	83.9008
$\overline{\|\dot{\hat{K}}_r\|}$	18.4647	7.8298	57.5958	5.5146	70.1344
$\overline{\|\dot{\hat{\theta}}\|}$	0.0888	0.0338	61.9369	0.0207	76.6892
$\overline{\|y_m\|_\infty}$	0.2	0.2064	-3.2	0.2	-
$\overline{\|e_y\|}$	0.4616	0.1957	57.6039	0.1379	70.1256
$\overline{\|e_y^o\|}$	0.4616	0.3928	14.9047	0.3886	15.8145
$\overline{\|\dot{u}\|}$	6.5704	2.0811	68.3262	1.4163	78.4290

signal. Beyond this uncertainty level, the system performance decreases and additional precautions should be considered to ensure flight safety owing to the actuator rate and actuator limitations.

The step responses of the CRM and RL-CRM adaptive systems for the worst case scenario are compared in Figure 12. In this figure, it is evident that the proposed RL-CRM system has a lower peak response than the CRM-adaptive system even in the worst case situation. Additionally, the difference between the peak responses in the nominal and worst cases is significantly lower in the RL-CRM adaptive system in comparison to that in the CRM adaptive system. However, to precisely quantify the performance improvements, we performed 500-run Monte-Carlo simulation of the proposed RL-CRM adaptive control system on the simplified helicopter model in a parametric uncertainty range of $\pm 35\%$. The transient response characteristics are evaluated based on the selected performance metrics and the results are summarized in Table 5. In this table, $\bar{(\cdot)}$ indicates the mean value of the related performance metric. According to the Monte-Carlo analysis, the RL-CRM adaptive control system provides an improvement of approximately $10-15\%$ in most of the performance metrics in comparison to the CRM-adaptive control system

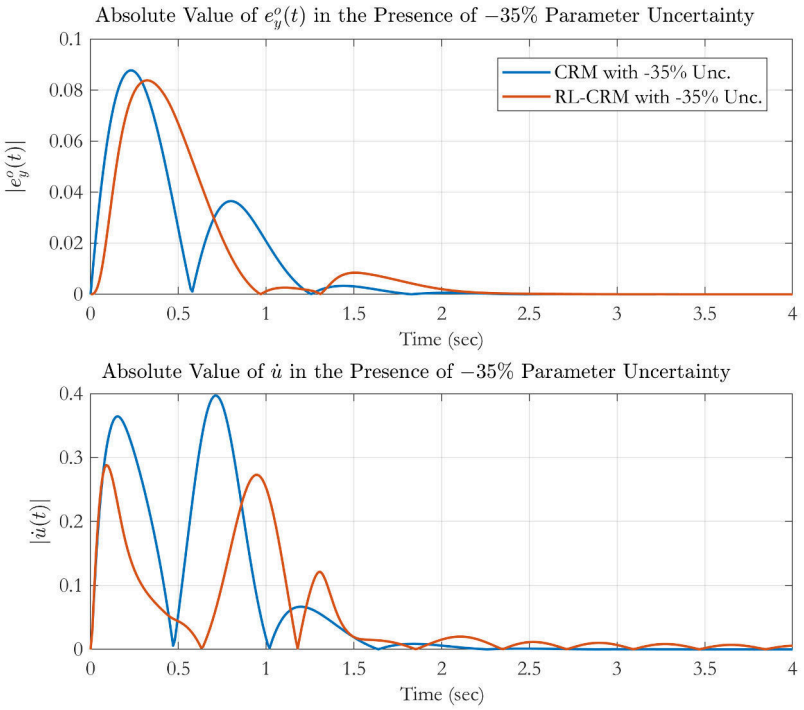

FIGURE 11 Comparison of $|e_y^o(t)|$ and $|\dot{u}(t)|$ time history of the CRM and RL-CRM systems in the presence of −35% parametric uncertainty.

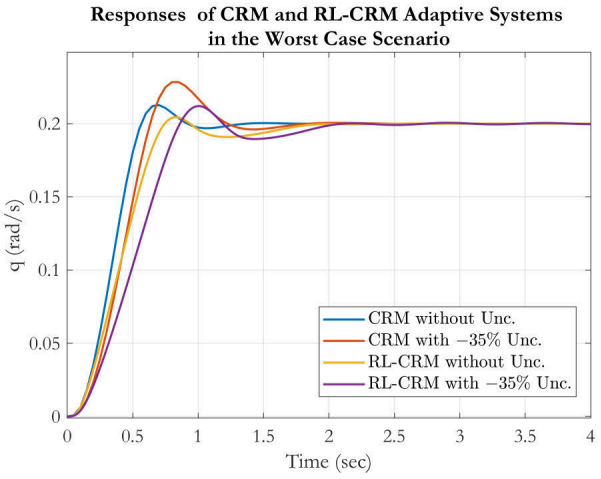

FIGURE 12 Step responses of the CRM and RL-CRM systems in the presence of −35% parametric uncertainty.

with the optimized design parameter; furthermore it yields an improvement of approximately 70 − 80% in almost all categories in comparison to MRAC.

The performance analysis of the MRAC, CRM, and RL-CRM adaptive systems in the worst case design scenario is presented in Table 6. Here, it is evident that the proposed RL-CRM adaptive system provides almost the same level of improvements on the key signal norms in comparison to the Monte-Carlo analysis presented in Table 5. In other words, the RL-CRM adaptive control system is robust against the parametric uncertainties and it exhibits similar transient response performance in terms of key metrics even in the worst design cases.

TABLE 6 Robustness analysis results of the MRAC, CRM, and RL-CRM adaptive systems in the worst case scenario.

Performance Metrics	MRAC	CRM	Improvement (%)	RL-CRM	Improvement (%)
$\|\dot{\hat{K}}_x\|$	19.7655	4.9225	75.0955	3.4801	82.3931
$\|\dot{\hat{K}}_r\|$	22.9284	9.4137	58.9431	6.4318	71.9483
$\|\dot{\hat{\theta}}\|$	0.1103	0.0407	63.1010	0.0246	77.6972
$\|y_m\|_\infty$	0.2	0.2171	-8.5500	0.2005	-0.2500
$\|e_y\|$	0.5732	0.2353	58.9498	0.1608	71.9470
$\|e_y^o\|$	0.5732	0.5101	11.0084	0.5214	9.0370
$\|\dot{u}\|$	8.5403	2.6274	69.2353	1.8001	78.9223

5 | CONCLUSIONS

The concepts of UAM and air cargo delivery heavily rely on the technological advances in autonomy, as well as the U-space and air traffic management solutions. Therefore, it is important to create safe, sustainable, cost-effective, and high-quality air transportation solutions for such urban airspace concepts. In aerial vehicles, specifically, the ability to precisely track the designed trajectories or flight corridors is a crucial requirement for the feasibility of such concepts. The aforementioned requirements become further complicated owing to the real-life variations in the mass, moment of inertia, aerodynamic properties, and power system properties of the aerial vehicles due to the changes observed in a wide range of operating conditions and payload weights. Additionally, flight control systems should have a certain level of resilience thorough adaptation, fault tolerance, and robustness to avoid catastrophic accidents in the presence of power system and/or actuation system anomalies. Within the context of air vehicles, as a part of designing resilient controllers, the transient behaviour observed during reconfiguration and adaptation is a very important part of the flight experience. Therefore, if the error magnitude between the system states and reference model states is high, the oscillatory transient response can lead to uncontrolled and undesirable flight states that can risk the passengers and the operational safety of the platform. Thus, in this study, we presented a novel RL-based approach for the CRM-adaptive flight control system design as to improve the transient response performance of the optimized fixed-gain CRM-adaptive control system. The proposed approach is based on the real-time trade-off between the improved transient dynamics and the model following error through a variable adaptation factor $\upsilon_{opt}k(t)$. This variable is introduced to dynamically update the magnitude of the design parameter υ of the closed-loop reference model feedback gain matrix $\mathbf{L}_\upsilon$. The adaptation strategy of $k(t)$ is designed through RL wherein an actor-critic agent structure is designed and trained by utilizing the DDPG algorithm to learn the optimal scaling policy of the design parameter. Simulation studies were performed on the identified pitch dynamics of an agile quadrotor platform and a benchmark example helicopter model. The results indicate that the proposed RL-CRM adaptive control system exhibits superior transient response performance in comparison to the MRAC and classical CRM-adaptive systems in terms of the selected key performance metrics. In addition, the proposed control structure provides the possibility to learn numerous adaptation strategies across various flight and vehicle conditions instead of being driven by high-fidelity simulators or through flight testing and real flight scenarios, as experienced and recorded by the business operations of such vehicles. Our current research focuses on (a) the stability proofs of the proposed RL-CRM adaptive controller with variable design parameter $\upsilon_{opt}k(t)$ and (b) the design engineering related to the reward functions of the RL agent to further enhance the transient performance while addressing the issues of fragility and time-delay effects. As such, we are further enhancing the robustness of the design methodology toward real-life effects such as the measurement noise, while designing projection operators and filtering methodologies for the observation vector.

References

1. Edwards C, Lombaerts T, Smaili H. *Fault Tolerant Flight Control*. Springer . 2010.

2. Zolghadri A, Henry D, Cieslak J, Efimov D, Goupil P. *Fault Diagnosis and Fault Tolerant Control and Guidance for Aerospace Vehicles*. Springer . 2014.

3. Narendra K, Annaswamy A. *Stable Adaptive Systems*. Courier Corporation . 2012.

4. Lavretsky E, Wise K. *Robust and Adaptive Control*. Springer . 2013.

5. Lavretsky E. Combined/composite model reference adaptive control. *IEEE Transactions on Automatic Control* 2009; 54(11): 2692–2697.

6. Gregory I, Gadient R, Lavretsky E. Flight test of composite model reference adaptive control (CMRAC) augmentation using NASA AirSTAR infrastructure. In: AIAA guidance, navigation, and control conference. ; August 8-11, 2011; Portland, Oregon: 6452.

7. Gibson T, Annaswamy A, Lavretsky E. Improved transient response in adaptive control using projection algorithms and closed loop reference models. In: AIAA Guidance, Navigation, and Control Conference. ; August 13-16, 2012; Minneapolis, Mennesota: 4775.

8. Gibson T, Annaswamy A, Lavretsky E. On adaptive control with closed-loop reference models: transients, oscillations, and peaking. *IEEE Access* 2013; 1: 703–717.

9. Dydek Z, Annaswamy A, Lavretsky E. Adaptive control of quadrotor UAVs: A design trade study with flight evaluations. *IEEE Transactions on control systems technology* 2012; 21(4): 1400–1406.

10. Cho N, Shin HS, Kim Y, Tsourdos A. Composite model reference adaptive control with parameter convergence under finite excitation. *IEEE Transactions on Automatic Control* 2017; 63(3): 811–818.

11. Wiese D, Annaswamy A, Muse J, Bolender M, Lavretsky E. Adaptive output feedback based on closed-loop reference models for hypersonic vehicles. *Journal of Guidance, Control, and Dynamics* 2015; 38(12): 2429–2440.

12. Zollitsch A, Holzapfel F, Annaswamy A. Application of adaptive control with closed-loop reference models to a model aircraft with actuator dynamics and input uncertainty. In: 2015 American Control Conference (ACC). IEEE. ; July 1-3, 2015; Chicago, IL: 3848–3853.

13. Gibson T, Annaswamy A, Lavretsky E. Adaptive Systems with Closed–loop Reference Models: Composite control and observer feedback. *IFAC Proceedings Volumes* 2013; 46(11): 440–445.

14. Gibson T, Annaswamy A, Lavretsky E. Closed-loop reference models for output-feedback adaptive systems. In: 2013 European Control Conference (ECC). IEEE. ; July 17-19, 2013; Zurich, Switzerland: 365–370.

15. Gibson T, Annaswamy A, Lavretsky E. Adaptive systems with closed-loop reference-models, part I: Transient performance. In: 2013 American Control Conference. IEEE. ; June 17-19, 2013; Washington, DC: 3376–3383.

16. Yuksek B. *A model based flight control system design approach for micro aerial vehicles using integrated flight testing and HIL simulations*. PhD thesis. Istanbul Technical University, 2019.

17. Tischler M, Berger T, Ivler C, Mansur M, Cheung K, Soong J. *Practical methods for aircraft and rotorcraft flight control design: an optimization-based approach*. American Institute of Aeronautics and Astronautics . 2017.

18. Tischler M, Remple R. *Aircraft and Rotorcraft System Identification*. AIAA education series. 2 ed. 2012.

19. Yuksek B, Saldiran E, Cetin A, Yeniceri R, Inalhan G. System identification and model-based flight control system design for an agile maneuvering quadrotor platform. In: AIAA 2020 SciTech Forum. AIAA. ; January 6-10, 2020; Orlando, FL.

20. Lillicrap T, Hunt J, Pritzel A, et al. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* 2015.

21. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research* 2012; 13(1): 281–305.

22. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. In: ; 2012: 2951–2959.

23. Zhang S, Sutton RS. A deeper look at experience replay. *arXiv preprint arXiv:1712.01275* 2017.

AUTHOR BIOGRAPHY

Burak Yuksek received his B.Sc. degree in Mechanical Engineering from Yildiz Technical University in 2010 and the M.Sc. and Ph.D. degrees in Mechatronics Engineering from Istanbul Technical University (ITU) in 2013 and 2019, respectively. He has been working as a research assistant in the ITU Aerospace Research Center, Control and Avionics Laboratory since 2013. His research interests include flight control system design, adaptive control theory, reinforcement learning, mathematical modeling, and system identification.

Gokhan Inalhan received his B.Sc. degree in Aeronautical Engineering from Istanbul Technical University in 1997, and M.Sc. and Ph.D. degrees in Aeronautics and Astronautics from Stanford University in 1998 and 2004, respectively. In 2003, he received a Ph.D. Minor from Stanford University in Engineering Economics and Operations Research (currently, Management Science and Engineering). Between 2004 and 2006, he worked as a Postdoctoral Associate at Massachusetts Institute of Technology. During this period, he led the Communication and Navigation group in the MIT-Draper Laboratory NASA CER project. He has served as the Director of the Controls and Avionics Laboratory (2006-2016) and Director General of the Aerospace Research Centre (2016-2019) at Istanbul Technical University. Gokhan is currently the BAE Systems Chair, Professor of Autonomous Systems and Artificial Intelligence and Deputy Head of the Centre for Autonomous and Cyber-Physical Systems at Cranfield University. He and his research group focus on design, modeling, GNC, resilience, and security aspects of autonomy and artificial intelligence for air, defense, transportation and space systems.

How to cite this article: Yuksek B., Inalhan G. (2020), Reinforcement Learning-Based Closed-Loop Reference Model Adaptive Flight Control System Design, *International Journal of Adaptive Control and Signal Processing*, *2020;XX:X–X*.