Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Model-Reference-Based Intelligent Active Disturbance Rejection Attitude Control for Morphing Aircraft

Haipeng Chen
China Academy of Launch Vehicle Technology
Fanyi Meng
Xi'an Jiaotong University
Yanru Huang
China Academy of Launch Vehicle Technology
Gang Chen
aachengang@xjtu.edu.cn

Xi'an Jiaotong University

Research Article

Keywords: Morphing aircraft, Model reference adaptive control (MRAC), Deep reinforcement learning, Twin Delayed Deep Deterministic Policy Gradient (TD3), Attitude control

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11549312/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Model-Reference-Based Intelligent Active Disturbance Rejection Attitude Control for Morphing Aircraft

Haipeng Chen[1], Fanyi Meng[2], Yanru Huang[1], Gang Chen[2*]

[1]The Research & Development Center, China Academy of Launch Vehicle Technology, Beijing, 100076, China.

[2]Shaanxi Key Laboratory of Environment and Control for Flight Vehicle, State Key Laboratory for Strength and Vibration of Mechanical Structures, School of Aerospace Engineering, Xi'an Jiaotong University, Xi'an, 710049, China.

*Corresponding author(s). E-mail(s): aachengang@xjtu.edu.cn;

Abstract

Attitude stabilization of hypersonic morphing vehicles is challenging because geometric reconfiguration continuously alters aerodynamic characteristics, intensifies cross-axis interactions, and changes available stability margins. Parameterizing these effects across the morphing envelope is impractical, limiting conventional model-based adaptive control. This paper presents a learning-augmented model reference adaptive control scheme, TD3-MRAC, that compensates for reconfiguration-induced uncertainties without explicit aerodynamic representation. A fixed-geometry reference model defines the desired closed-loop response, while a backstepping controller provides nominal stabilization. To use only low-dimensional output measurements, a linear extended state observer reconstructs the system states and estimates the aggregate influence of unmodeled dynamics and external disturbances. Instead of a prescribed analytical adaptation law, a Twin Delayed Deep Deterministic Policy Gradient agent generates bounded incremental control commands to reduce mismatch between the morphing vehicle and reference dynamics. Stability is established independently of reinforcement-learning convergence: the reference closed loop is globally exponentially stable, the observer error dynamics are input-to-state stable with an adjustable estimation bound, and the complete closed-loop system is uniformly ultimately bounded with an ultimate bound explicitly characterized by the design parameters. Numerical evaluations under combined aerodynamic, parametric, and external

perturbations show improved tracking over a conventional robust adaptive controller, especially in roll and yaw, where integral absolute error is reduced by up to 46%, while pitch performance remains comparable. The trained policy remains effective on unseen perturbation combinations without additional training. These results indicate that TD3-MRAC offers a theoretically supported and practically implementable approach for morphing-flight control while substantially reducing dependence on labor-intensive aerodynamic uncertainty modeling.

Keywords: Morphing aircraft, Model reference adaptive control (MRAC), Deep reinforcement learning, Twin Delayed Deep Deterministic Policy Gradient (TD3), Attitude control

1 Introduction

The morphing aircraft, due to its ability to adjust its aerodynamic shape according to the environment and mission requirements, has emerged as a prominent direction in aircraft development [1, 2]. Common deformable-wing aircraft include those with variable trailing edge, variable span length, and variable camber [3]. However, during the transformation process of such deformable aircraft, problems such as strong time-varying and strong uncertainty in the system often occur, and the stability boundary of the aircraft changes, which brings certain difficulties to the stable control of the aircraft. In particular, the aerodynamic and structural parameters of hypersonic morphing vehicle (HMV) exhibit significant nonlinear changes during deformation. At the same time, the inertial additional forces and additional moment effects cannot be ignored, and the poor structural stiffness leads to significant aerodynamic and structural coupling phenomena. These characteristics make the morphing aircraft model highly uncertain and subject to complex external disturbances, posing severe challenges to the design of control systems.

Traditional linear control methods such as LQR are limited to handling simple deformations and small disturbances during flight. Consequently, nonlinear methods such as backstepping, model predictive control (MPC), adaptive control, sliding mode control (SMC), and active disturbance rejection control (ADRC), which are applicable to nonlinear control, have been proposed and widely applied in aircraft control systems [4–7]. However, these nonlinear control methods also face problems such as complex parameter tuning, poor adaptability, and dependence on precise models. Therefore, improving the adaptive ability of morphing aircraft control systems and reducing the reliance on precise models has become an important issue at present. Fortunately, reinforcement learning (RL) has brought positive effects in enhancing the robustness of control systems. In the design of control systems, common approaches to applying reinforcement learning include adaptive tuning of controller parameters, end-to-end control, and auxiliary control methods combined with existing controllers.

In early research efforts, reinforcement learning was employed for the adaptive tuning of the parameters of the controller. Sönmez et al. [8] propose a deep reinforcement learning (DDPG) approach [9] to online fine-tune PID gains of a quadrotor UAV during

flight, showing that the RL-tuned controller yields significantly smaller attitude tracking errors compared with hand-tuned gains. Zhang et al. [10] integrate reinforcement learning (A2C) with a fuzzy PID framework for UAV altitude control, enabling automatic self-tuning of controller parameters and improving anti-disturbance performance over fixed rule-based settings. Hezaveh et al. [11] explore reinforcement learning combined with genetic algorithms to optimize the PID gains of a UAV altitude controller, leading to more precise tracking of setpoints and trajectories.

Furthermore, deep reinforcement learning (DRL) enables adaptive learning based on high-dimensional data, providing a new solution for optimizing the parameters of backstepping control [12]. Specifically, DRL dynamically adjusts the size of the control switching gain parameters according to the current system state. In complex operating conditions such as aerodynamic parameter perturbations and strong disturbances, it automatically increases the parameters to enhance the robustness against disturbances; while when the system approaches a stable state, it adaptively reduces the parameter size to avoid the risk of oscillation. DRL can effectively balance the relationship between control performance and system stability through adaptive controller gain. This parameter adaptive optimization method based on DRL can adapt to complex and variable flight environments, fundamentally differentiating itself from traditional methods that rely on manual experience or offline optimization.

However, even with reinforcement learning for parameter tuning, the inherent limitations of the original controller structure may persist. This is because RL only optimizes parameters without altering the mathematical structure or logic of the controller. For the fundamental limitations of structures like PID in terms of non-linearity, disturbance resistance, or dynamic changes, RL parameter tuning can only improve performance within its inherent framework and cannot resolve these structural limitations. Therefore, for HMV with stronger time-varying and disturbance conditions, such methods may no longer be applicable.

Beyond parameter tuning, reinforcement learning can also directly generate control instructions for the morphing aircraft in an end-to-end manner. Lin et al. [13] develop an end-to-end neural policy that maps raw environment observations to thrust and attitude commands for a quadrotor and show that their learned controller can successfully fly through a tilted narrow gap, demonstrating competitive dynamic control performance compared to traditional pipelines. Zhao et al. [14] applies deep reinforcement learning to build an end-to-end control network mapping raw image observations directly to UAV control actions, effectively simplifying traditional perception and planning pipelines and achieving autonomous flight behaviors. However, despite the appealing end-to-end paradigm, directly deploying DRL policies in safety-critical flight control remains challenging. The exploration mechanism in RL may lead to unsafe behaviors during training, and the learned policy is often difficult to interpret and verify in terms of stability margins under time-varying aerodynamics, inertia-coupling effects, and external disturbances. To address these concerns, recent safe RL studies introduce stability- or safety-related constraints (e.g., Lyapunov-based constraints or control-barrier-function layers) to enforce safety during learning and execution, but their practical integration with highly uncertain flight dynamics is still nontrivial [15]

Motivated by these limitations, a promising direction is to combine RL with structured controllers rather than to fully replace them. A representative idea is residual reinforcement learning, where a baseline controller guarantees fundamental stability while RL learns an additive residual term to compensate for unmodeled dynamics and complex disturbances [16]. In aerospace applications, such hybrid paradigms can be further integrated with established nonlinear flight-control frameworks (e.g., dynamic inversion or incremental control) to retain interpretable control structure while enhancing adaptivity [17]. In addition, disturbance-observer-based designs (e.g., ESO/DOB) provide a natural interface for learning-based disturbance compensation, enabling RL to improve robustness with reduced model reliance [18]. Johannink et al. [16] proposed Residual Reinforcement Learning, where the final control input is a superposition of a conventional feedback controller and an RL-learned residual policy, enabling improved real-world performance on contact-rich manipulation tasks without discarding the baseline control structure. Alakuijala et al. extended the residual formulation to leverage demonstrations and high-dimensional observations, showing that residual policies can be trained with sparse rewards while maintaining the benefits of structured baseline controllers [19]

In addition, several studies have introduced barrier functions in reinforcement learning to enforce system safety. Luo and Ma [20] studied Learning Barrier Certificates toward safe RL with zero training-time violations, exploring certificate-based mechanisms to maintain safety throughout training rather than only at deployment. Hou et al. [21] proposed a safe and efficient RL method using disturbance observers (DOBs) and CBFs as a safety filter, where DOB-estimated uncertainty is incorporated into a robust CBF condition so that RL actions are modified only when needed to ensure safety. Kalaria et al. combined DOB-based CBFs with residual model learning to quantify discrepancies between nominal and true dynamics, demonstrating improved safe-learning behavior in benchmark tasks and physical platform validation [22]

Beyond the aforementioned approaches, model reference adaptive control (MRAC) provides a structured paradigm that is particularly attractive for safety-critical flight control [23–25]. In MRAC, a desired reference model specifies the target closed-loop dynamics, and the controller adapts online to drive the plant to follow the reference model, offering an interpretable control architecture with analyzable stability properties. However, conventional MRAC designs may still face difficulties under severe nonlinear, time-varying uncertainties, actuator constraints, and strong disturbances, where selecting adaptation gains and robustness modifications becomes nontrivial. Recent variants for aircraft further enhance MRAC by combining fast adaptation with function-approximation terms to improve disturbance tolerance [26], and adaptive compensation has also been applied to morphing-actuator failure cases [27]

Motivated by the above considerations, integrating reinforcement learning with MRAC has emerged as a promising direction to enhance adaptivity while retaining a structured control backbone. One representative line of research formulates MRAC in an optimal control framework and adopts adaptive dynamic programming /actor–critic learning to approximate the optimal policy and value function

for unknown nonlinear systems, enabling robustness-aware adaptation under constraints. [28] Another line develops deep/neuro-adaptive MRAC architectures and demonstrates their effectiveness in real flight tests under wind disturbances and system faults, indicating that learning-enhanced MRAC can maintain high-bandwidth attitude control in challenging environments [29]. In addition, adaptive robustification methods can be combined with RL to mitigate dynamic variations and improve the out-of-distribution robustness of learned policies, which is particularly relevant for highly uncertain aerospace systems. For hypersonic morphing vehicles, the strong coupling between aerodynamics, structural flexibility, and inertial effects during morphing imposes stringent requirements on both robustness and stability guarantees. Therefore, a learning-enhanced MRAC framework that leverages RL for adaptation/compensation while preserving a reference-model-guided control structure becomes a compelling candidate to address time-varying uncertainties and complex disturbances with reduced reliance on precise models [30]

Although MRAC provides an attractive structured paradigm for safety-critical flight control, conventional MRAC is inadequate for morphing vehicles for the following concrete reasons. First, the linear-in-parameters assumption breaks down: classical MRAC represents the uncertain dynamics as $\mathbf{f}(\cdot) = \mathbf{\Phi}^T(\mathbf{x})\,\mathbf{\Theta}^*$ with a constant parameter vector, whereas for a morphing vehicle the aerodynamic and inertial maps depend on the morphing state $\mathbf{\Lambda}(t)$, so the "true parameters" are configuration-dependent, $\mathbf{\Theta}^* = \mathbf{\Theta}^*(\mathbf{\Lambda}(t))$, and a constant-parameter adaptation law persistently tracks a moving target. Second, morphing-induced inertial coupling enters the dynamics through non-matched channels ((17)), violating the matched-uncertainty condition on which the stability proofs of classical MRAC rely. Third, long steady-state command segments degrade the persistence of excitation of the regressor, causing parameter drift and intermittent bursting [31]; the standard remedy, σ-modification, introduces a non-vanishing steady-state bias that itself varies with the configuration. Fourth, during fast reconfiguration the rate of change of the true parameters exceeds the adaptation bandwidth, so the parameter estimates lag the morphing dynamics, while increasing the adaptation gain re-triggers the drift and robustness problems above. Fifth, high-gain adaptation in the presence of low-frequency structural modes and actuator dynamics violates the robustness small-gain condition of [32], again yielding intermittent bursting [31]. Finally, actuator saturation under combined maneuvering and morphing causes adaptation windup unless additional anti-windup structure is added. These mechanisms motivate replacing the analytic adaptation law with a bounded, offline-trained compensation policy embedded in a structured baseline, as developed in Section 4.

A central design choice of TD3-MRAC is the adoption of a fixed-wing aircraft as the reference model. This choice is deliberate and rests on the following engineering rationale. First, at any instant a morphing vehicle is physically a fixed-geometry aircraft at some configuration; adopting a fixed-wing reference model therefore amounts to prescribing, as the tracking objective, a closed-loop response that satisfies conventional flying-qualities requirements (e.g., MIL-STD-1797B short-period and Dutch-roll criteria [33]), which provides a verifiable and certification-oriented performance specification that a purely learned objective cannot offer. Second, the true morphing dynamics

decompose into the nominal fixed-wing model plus configuration-dependent deviations; as long as the morphing state evolves within a bounded envelope (Assumption A4), the deviations are uniformly bounded, reducing the control problem to tracking a known nominal object while compensating bounded mismatch—the classical MRAC setting. Third, morphing evolves on a time scale (the sweep schedule of Section 5 has a period of 24 s) well separated from the attitude-loop and observer bandwidths, so a quasi-static frozen-configuration rationale applies; the limits of this assumption are quantified by the perturbation sensitivity analysis of Section 5.3. Fourth, the fixed-wing anchor enables a three-layer responsibility decomposition—baseline for flying qualities, ESO for lumped-disturbance compensation, RL for residual mismatch—each layer being independently analyzable (Theorems 1–3) and diagnosable, a property that end-to-end policies do not possess. Finally, for large configuration deviations, the architecture supports a three-level strategy: perturbation-rich training over the envelope, gain-scheduled reference models with hysteresis switching [34], and a deterministic safety monitor that disables the learned increment outside the certified envelope (Section 4.3).

In this paper, a learning-enhanced attitude control framework is developed for hypersonic morphing vehicles (HMVs) by introducing a fixed-wing aircraft as the nominal (baseline) model. Based on the nominal fixed-wing dynamics, a standard attitude controller is first designed to provide a reliable baseline stabilization capability. Then, reinforcement learning (RL) is employed as an auxiliary compensation module to optimize and augment the baseline controller online, so that the final control input can be interpreted as the superposition of a structured nominal control law and an RL-generated correction term—an idea that is closely aligned with the "residual/auxiliary RL" paradigm, where RL learns the unmodeled residual while the baseline controller preserves fundamental stability and interpretability. The remainder of this paper is organized as follows: Section 2 reviews the preliminaries and theoretical foundations used in this study; Section 3 establishes the nonlinear dynamics model of the hypersonic morphing vehicle; Section 4 presents the proposed controller design, including the nominal fixed-wing attitude controller and the RL-based compensation strategy, together with the closed-loop stability analysis; Section 5 provides simulation studies to validate the effectiveness and robustness of the proposed method under morphing-induced uncertainties and disturbances; and Section 6 concludes the paper and discusses future work.

The innovations of this paper are as follows.

i) A flying-qualities-oriented modeling paradigm for morphing vehicles. A fixed-wing aircraft is adopted as the reference model, decomposing the morphing dynamics into a nominal fixed-wing model plus bounded configuration-dependent deviations; the engineering rationale and the standing assumptions are established in the Introduction and Section 4.3.

ii) A stability-certified learning-enhanced MRAC architecture. The baseline backstepping controller, the linear extended state observer with explicit feedforward compensation, and the bounded TD3-learned residual are integrated with a complete layered stability analysis—global exponential stability of the reference loop,

input-to-state stability of the observer, and uniform ultimate boundedness of the composite loop with an explicit ultimate bound (Theorems 1–3, Corollary 1, Lemma 1, Proposition 1).

iii) A validated demonstration of the learning benefit. Against a robust adaptive baseline under an identical 500-scenario multi-source perturbation ensemble, the framework achieves up to 46% IAE reduction in roll and 13–20% in yaw with competitive pitch performance, and the theoretical guarantees of Section 4.3 quantify the worst-case tracking tube within which the learned policy operates.

2 Preliminaries and problem background

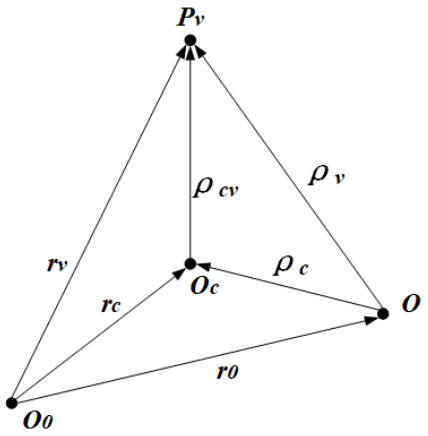


Fig. 1: Conceptual diagram of the Newton-Euler formulation

2.1 Trajectory tracking cast as a reinforcement learning task

The trajectory-following problem is formalized within the Markov decision process (MDP) framework. Let x_t represent the instantaneous plant state vector and x_t^{ref} denote the prescribed reference trajectory, which may vary with time. The instantaneous discrepancy between the actual and commanded configurations is captured by the tracking deviation

$$\varepsilon(t) = x(t) - x^{ref}(t) \tag{1}$$

When the reference signal varies over time, the raw tracking error alone is generally insufficient to render the decision process Markovian. The standard remedy is to augment the observation vector with adequate summary statistics of the reference dynamics—for instance, by appending the current reference value together with its higher-order time derivatives, or by introducing an auxiliary phase or

clock variable. This augmentation produces a composite state of the form $s_t = \left[x(t),\ x^{ref}(t), \dot{x}^{ref}(t)\right]$, or, more compactly,

$$s_t = [\varepsilon(t), \dot{\varepsilon}(t), x^{ref}(t)]$$

and the action $a_t \in \mathbf{A}$ is identified with the control effort $u(t)$. The environment then transitions according to the underlying dynamics $s_{t+1} \sim p(\bullet|s_t,\ a_t)$, while a scalar reward $r_t = r(s_t, a_t)$ provides immediate feedback on the quality of the current action. Augmenting the state with reference-related information is a standard and well-established technique in RL-based trajectory-tracking control.

The learning objective is to synthesize a deterministic feedback policy that maximizes the expected discounted sum of future rewards over an infinite planning horizon

$$I(\pi) = E\left[\sum_{t=0}^{\infty} \gamma^t r\left(s_t, a_t\right)\right], \qquad \gamma \in (0,1) \tag{2}$$

For trajectory-tracking applications, a natural choice for the stage cost is a quadratic penalty that jointly weights the instantaneous tracking residual and the magnitude of the control effort. An additional penalty on control increments may also be incorporated to encourage smooth actuation and temporal regularity in the control signal:

$$r_t = -\varepsilon^T(t) Q \varepsilon(t) - a_t^T V a_t - \lambda \left\| a_t - a_{t-1} \right\|^2$$

with weighting matrices $Q \geq 0, V \geq 0$, $\lambda \geq 0$.

Deterministic actor–critic design and the TD3 algorithm. For continuous action domains, a deterministic actor–critic architecture is adopted. The actor implements a deterministic state-to-action mapping $a = \mu_\varphi(s)$, while the critic $Q_\vartheta(s, a)$ learns to estimate the associated action-value function. Following the deterministic policy gradient (DPG) theorem [35], the actor is improved by ascending the gradient of the critic's prediction with respect to the action—quantities that are readily obtained by backpropagating through the critic network.

We instantiate the proposed framework using the Twin Delayed Deep Deterministic Policy Gradient (TD3) method [36], which integrates three complementary mechanisms to counteract the overestimation bias and training instability commonly encountered under nonlinear function approximation: (i) clipped double-Q learning, in which two independent critic estimates are maintained and the more conservative of the two is used when constructing bootstrap targets; (ii) target policy smoothing, whereby a small random perturbation is added to the target action to discourage the policy from exploiting sharp peaks in the value estimate; and (iii) delayed policy updates, meaning the actor and target networks are updated less frequently than the critics, so that the value function can partially converge before it is used to guide policy improvement. In formal terms, TD3 maintains a pair of critics $Q_{\vartheta 1}, Q_{\vartheta_2}$ with corresponding target networks $Q_{\vartheta'_1}, Q_{\vartheta'_2}$, together with a deterministic actor μ_φ and its delayed target copy μ'_φ. For a sampled transition (s, a, r, s'), a perturbed target action is first formed as

$$\widetilde{a} = \mu'_{\varphi}(s') + \epsilon, \qquad \epsilon \sim clip(N(0,\sigma), -c, c)$$

from which the temporal-difference target is formed as

$$y = r + \gamma \min_{i \in \{1,2\}} Q_{\vartheta'_i}(s', \widetilde{a})$$

Each critic is optimized by minimizing the mean-squared Bellman residual

$$H(\vartheta_i) = \mathbf{E}\left[\left(Q_{\vartheta_i}(s,a) - y\right)^2\right], \qquad i \in \{1,2\} \tag{3}$$

The actor is updated once every d critic iterations, by maximizing the output of the first critic (typically Q_{ϑ_1} serves as the optimization surrogate):

$$\max_{\varphi} \mathbf{E}[Q_{\vartheta_1}(s, \mu_{\varphi}(s))]$$

The target networks are adjusted via Polyak (exponential moving) averaging:

$$\vartheta'_i \leftarrow \tau\vartheta_i + (1-\tau)\vartheta'_i, \qquad \varphi' \leftarrow \tau\varphi + (1-\tau)\varphi', \qquad \tau \in (0,1)$$

Taken together, these design choices effectively suppress value overestimation and limit the per-update accumulation of policy errors, yielding a more stable and sample-efficient learning procedure well suited to high-dimensional continuous control problems.

2.2 Fundamentals of model-reference control

Model Reference Control (MRC) is a systematic methodology for steering a dynamical plant so that its output faithfully reproduces the response prescribed by a user-specified reference model. The reference model encapsulates the desired closed-loop characteristics, including transient response shaping and steady-state accuracy targets.

In state-space form, the reference dynamics are characterized by the following linear representation:

$$\begin{aligned} \dot{\mathbf{x}}_{\mathbf{m}}(t) &= \mathbf{A}_{\mathbf{m}}\mathbf{x}_{\mathbf{m}}(t) + \mathbf{B}_{\mathbf{m}}\mathbf{r}(t) \\ \mathbf{y}_{\mathbf{m}}(t) &= \mathbf{C}_{\mathbf{m}}\mathbf{x}_{\mathbf{m}}(t) \end{aligned} \tag{4}$$

In the description above, $\mathbf{x_m(t)}$ denotes the state vector of the reference model, $\mathbf{r(t)}$ is a bounded external command input, and $\mathbf{y_m(t)}$ represents the corresponding desired output trajectory. The system matrices $\mathbf{A_m}$, $\mathbf{B_m}$, and $\mathbf{C_m}$ fully specify the target closed-loop dynamics.

The actual plant subject to control is governed by

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) \end{aligned} \tag{5}$$

with state vector $\mathbf{x(t)}$, control input $\mathbf{u(t)}$, and measured output $\mathbf{y(t)}$.

The core objective is to synthesize a control policy $\mathbf{u(t)}$ that drives the plant output $\mathbf{y(t)}$ to track the model reference $\mathbf{y_m(t)}$ as faithfully as possible. A natural performance metric is the instantaneous output mismatch $\mathbf{e}(t) = \mathbf{y}(t) - \mathbf{y_m}(t)$, which the controller should drive toward zero.

In the adaptive extension of MRC—MRAC—the controller gains are continuously tuned online via a parameter adaptation law, typically derived from Lyapunov stability arguments or gradient-based optimization of the tracking error. This provides the closed-loop system with robustness against parametric uncertainties and gradual variations in the plant dynamics.

2.3 Newton–Euler equations for multibody systems

With reference to Fig. 1, consider a mechanical assembly composed of an arbitrary collection of material particles G_v $(v = 1, 2, \ldots)$. Both an isolated rigid body and an articulated chain of multiple rigid bodies constitute special instances of such a particulate system. Select an inertially fixed reference point O_0 as the base measurement origin. For any constituent particle in the assembly, let $\mathbf{r_v}$ denote its position vector relative to O_0, and let m_v be its mass. The total mass accumulated over the entire system is $m = \sum_v m_v$. The linear momentum of the particle assembly is then defined as

$$\mathbf{p} = \sum_v m_v \dot{\mathbf{r}}_v = m\dot{\mathbf{r}}_c \tag{6}$$

Here $\mathbf{r_c}$ is the position vector of the system centroid O_c with respect to the base point O_0. Equation (6) reveals a fundamental fact: the total linear momentum of a distributed system is identical to that of a notional point mass—whose entire mass is concentrated at the center of mass $\mathbf{O_c}$—moving with the centroid velocity. Now let r_0 be an arbitrarily chosen moving reference point, with position $\mathbf{O_c}$ measured from O_0, and let $\rho_\mathbf{v}$ be the position of particle G_v relative to $\mathbf{O_c}$, so that $\mathbf{r_v} = \mathbf{r_0}+\rho_\mathbf{v}$. The moment of momentum (angular momentum) $\mathbf{H}$ of the particle system about $\mathbf{O_c}$ is then given by (7).

$$\mathbf{H} = \sum_v \rho_\mathbf{v} \times m_v \dot{\mathbf{r}}_v \tag{7}$$

The angular momentum $\mathbf{H'_c} = \sum_v \rho_\mathbf{cv} \times m_v \dot{\rho}_{cv} = \mathbf{I_c}\omega$ of the system computed about O in a translating reference frame attached to the base point O is introduced into (7), yielding the decomposition shown below.

$$\mathbf{H} = \mathbf{H'_c} + m\rho_\mathbf{c} \times (\dot{\mathbf{r}}_\mathbf{0} + \dot{\rho}_\mathbf{c}) = \mathbf{I_c}\omega + m\rho_\mathbf{c} \times (\dot{\mathbf{r}}_\mathbf{0} + \dot{\rho}_\mathbf{c}) \tag{8}$$

$\rho_\mathbf{c}$ is the position vector of the center of mass $\mathbf{O_c}$ relative to $\mathbf{O}$, and $\mathbf{v_0} = \dot{\mathbf{r}}_\mathbf{0}$ is the translational velocity of $\mathbf{O}$. The term $\mathbf{H}$ denotes the absolute angular momentum and $\mathbf{H'_c}$ the relative one. When the reference point $\mathbf{O}$ coincides with the system center of mass $\mathbf{O_c}$, the coupling terms vanish and the relative angular momentum reduces to $\mathbf{H'_c} = \mathbf{I_c}\omega$, where the inertia tensor is evaluated about $\mathbf{O_c}$. Euler's rotational equation then takes the compact form

$$\begin{aligned} \frac{d\mathbf{H}}{dt} = \frac{\partial\left(\mathbf{I_c}\omega\right)}{\partial t} + \omega \times \mathbf{I_c}\omega + \frac{d}{dt}\left[m\rho_\mathbf{c} \times (\dot{\mathbf{r}}_\mathbf{0} + \dot{\rho}_\mathbf{c})\right] \\ + \mathbf{v_0} \times \mathbf{p} = \mathbf{M} \end{aligned} \tag{9}$$

In (9), $\mathbf{v_0} = \dot{\mathbf{r}}_\mathbf{0}$ and $\mathbf{p} = m\dot{\mathbf{r}}_c$; $\mathbf{M}$ represents the net external torque applied to the body.

3 Dynamic model of the hypersonic morphing vehicle

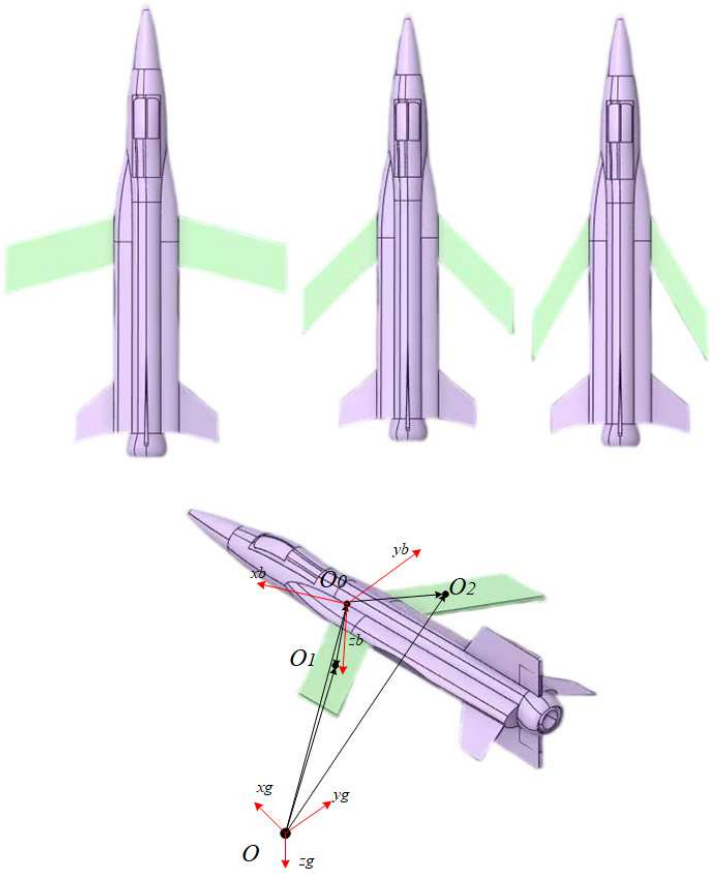


Fig. 2: Configuration of the morphing aircraft with variable-geometry wings

This section develops a general framework for deriving the attitude dynamics of a morphing vehicle. To render the analysis tractable while preserving the essential coupling between shape reconfiguration and rigid-body motion, the following simplifying hypotheses are adopted throughout the derivation.

Assumption 1: The airframe and both wings are treated as rigid bodies. Elastic deformations are neglected, so that structural flexibility does not appear in the equations of motion.

Assumption 2: The magnitude of the aircraft's translational velocity vector expressed in the inertial frame remains constant over the timescale of interest. Consequently, the analysis focuses exclusively on the rotational (attitude) degrees of freedom; the translational dynamics are decoupled and need not be treated here.

3.1 Multibody formulation of the morphing aircraft dynamics

Fig. 2 depicts a representative morphing air vehicle equipped with two independently reconfigurable wings. In the diagram, O_i, $(i = 0, 1, 2)$ identify the centroids of the fuselage, the left wing, and the right wing, respectively; $(\varphi,\ \vartheta,\ \psi)$ are the three Euler angles—roll, pitch, and yaw—that parameterize the orientation of the fuselage body frame; and the hinge points connecting each wing to the fuselage have coordinates l_1 and l_2 in the fuselage-fixed reference frame. The geometric deformation states of the left and right wings are encoded by the parameters ξ_1 and ξ_2, respectively.

3.1.1 Newton–Euler derivation of the coupled equations of motion

To faithfully capture how wing reconfiguration influences the overall rotational motion, the fuselage and each wing must be treated as distinct dynamical entities whose mutual interactions are transmitted through the hinge constraints. The derivation proceeds by first writing the angular momentum balance for the fuselage about its own center of mass, invoking the moment-of-momentum (angular momentum) theorem.

$$\mathbf{L_0} = \mathbf{I_{c0}}\boldsymbol{\Omega_0} \tag{10}$$

In the above expression, $\mathbf{I_{c0}}$ is the inertia dyadic of the fuselage referenced to its own centroid, and $\boldsymbol{\Omega_0}$ denotes the instantaneous angular rate of the fuselage resolved in body-fixed axes. Taking the material time derivative of (10) yields the resultant external torque acting on the fuselage:

$$\dot{\mathbf{L}}_\mathbf{0} = \mathbf{I_{c0}}\frac{\partial \boldsymbol{\Omega_0}}{\partial t} + \boldsymbol{\Omega_0} \times \mathbf{I_{c0}}\boldsymbol{\Omega_0} = \mathbf{M_{a0}} + \mathbf{M_{10}} + \mathbf{M_{20}} \tag{11}$$

Turning to (11), the terms $\mathbf{M_{a0}}$, $\mathbf{M_{10}}$, and $\mathbf{M_{20}}$ designate, respectively, the aerodynamic torque about the fuselage center of mass and the constraint torques transmitted from the left and right wings to the fuselage through the hinge joints. For the angular momentum analysis of the left wing, the fuselage center of mass is retained as the moment reference point for convenience in the subsequent algebraic manipulation. Accordingly, the angular momentum of the left wing about this reference is given by

$$\mathbf{L_1} = \mathbf{I_{c1}}\boldsymbol{\Omega_1} + m_1\boldsymbol{\rho_{c1}} \times (\dot{\mathbf{r}}_\mathbf{10} + \dot{\rho}_\mathbf{c1}) \tag{12}$$

Within (12), $\mathbf{r_{10}}$ denotes the inertial position of the left wing; $\rho_\mathbf{c1}$ is the relative position vector pointing from the left wing centroid O_1 to the fuselage centroid O_0, and the sum of their time derivatives yields the absolute translational velocity of the left wing as measured in the inertial frame. The quantity $\mathbf{I_{c1}}$ is the left wing inertia tensor about its own center of mass O_1, and m_1 denotes the left wing mass. Differentiating (12) with respect to time produces the rate form presented below.

$$\dot{\mathbf{L_1}} = \mathbf{I_{c1}}\frac{\partial \mathbf{\Omega_1}}{\partial t} + \mathbf{\Omega_1} \times \mathbf{I_{c1}}\mathbf{\Omega_1} + m_1\dot{\rho}_{\mathbf{c1}} \times (\dot{\mathbf{r}}_{\mathbf{10}} + \dot{\rho}_{\mathbf{c1}}) + m_1\rho_{\mathbf{c1}} \times (\ddot{\mathbf{r}}_{\mathbf{10}} + \ddot{\rho}_{\mathbf{c1}}) \tag{13}$$

Applying the angular momentum principle, the net external torque on the left wing satisfies

$$\dot{\mathbf{L_1}} + \mathbf{v_1} \times \mathbf{p_1} = \mathbf{M_{a1}} + \mathbf{M_{01}} + \mathbf{M_{G1}} \tag{14}$$

In (14), $\mathbf{M_{a1}}$, $\mathbf{M_{01}}$, and $\mathbf{M_{G1}}$ denote, respectively, the aerodynamic couple acting directly on the left wing, the reaction torque imposed by the fuselage on the left wing at the hinge, and the gravitational contribution to the torque about the fuselage reference point. All moments are taken with the fuselage center of mass serving as the moment reference center. $\mathbf{p_1}$ is the linear momentum of the left wing.

An entirely analogous development is carried out for the right wing. Here $\rho_{\mathbf{c2}}$ represents the position vector from the right wing centroid O_2 to the fuselage centroid O_0, whose time derivative contributes to the inertial velocity of the right wing. The inertia tensor of the right wing about its own centroid is $\mathbf{I_{c2}}$, and O_2 is the right wing centroid, and m_2 denotes the right wing mass.

$$\mathbf{L_2} = \mathbf{I_{c2}}\mathbf{\Omega_2} + m_2\rho_{\mathbf{c2}} \times (\dot{\mathbf{r}}_{20} + \dot{\rho}_{\mathbf{c2}})$$

Taking the time derivative of the right wing angular momentum yields

$$\dot{\mathbf{L_2}} = \mathbf{I_{c2}}\frac{\partial \mathbf{\Omega_2}}{\partial t} + \mathbf{\Omega_2} \times \mathbf{I_{c2}}\mathbf{\Omega_2} + m_2\dot{\rho}_{\mathbf{c2}} \times (\dot{\mathbf{r}}_{\mathbf{20}} + \dot{\rho}_{\mathbf{c2}}) + m_2\rho_{\mathbf{c2}} \times (\ddot{\mathbf{r}}_{\mathbf{20}} + \ddot{\rho}_{\mathbf{c2}}) \tag{15}$$

Invoking the moment-of-momentum theorem for the right wing gives

$$\dot{\mathbf{L_2}} + \mathbf{v_2} \times \mathbf{p_2} = \mathbf{M_{a2}} + \mathbf{M_{02}} + \mathbf{M_{G2}} \tag{16}$$

Referring to (16), $\mathbf{M_{a2}}$, $\mathbf{M_{02}}$, and $\mathbf{M_{G2}}$ denote the aerodynamic torque on the right wing, the constraint torque from the fuselage on the right wing, and the gravitational torque on the right wing, respectively—all taken about the fuselage centroid. $\mathbf{p_2}$ is the linear momentum of the right wing.

Now sum together the left- and right-hand sides of Equations (11), (14), and (16). The internal reaction torque pairs $\mathbf{M_{10}}$ and $\mathbf{M_{01}}$, as well as $\mathbf{M_{20}}$ and $\mathbf{M_{02}}$, form action-reaction couples of equal magnitude and opposite direction by Newton's third law; they therefore cancel upon summation. Carrying out this addition leads directly to (17).

$$\begin{aligned}\mathbf{I_{c0}}\frac{\partial \mathbf{\Omega_0}}{\partial t} + \mathbf{\Omega_0} \times \mathbf{I_{c0}}\mathbf{\Omega_0} + \mathbf{I_{c1}}\frac{\partial \mathbf{\Omega_1}}{\partial t} + \mathbf{\Omega_1} \times \mathbf{I_{c1}}\mathbf{\Omega_1} \\ + m_1\dot{\rho}_{\mathbf{c1}} \times (\dot{\mathbf{r}}_{\mathbf{10}} + \dot{\rho}_{\mathbf{c1}}) + m_1\rho_{\mathbf{c1}} \times (\ddot{\mathbf{r}}_{\mathbf{10}} + \ddot{\rho}_{\mathbf{c1}}) + \mathbf{I_{c2}}\frac{\partial \mathbf{\Omega_2}}{\partial t} + \mathbf{\Omega_2} \times \mathbf{I_{c2}}\mathbf{\Omega_2} \\ + m_2\dot{\rho}_{\mathbf{c2}} \times (\dot{\mathbf{r}}_{\mathbf{20}} + \dot{\rho}_{\mathbf{c2}}) + m_2\rho_{\mathbf{c2}} \times (\ddot{\mathbf{r}}_{\mathbf{20}} + \ddot{\rho}_{\mathbf{c2}}) + \mathbf{v_1} \times \mathbf{p_1} + \mathbf{v_2} \times \mathbf{p_2} \\ = \mathbf{M_{a0}} + \mathbf{M_{a1}} + \mathbf{M_{G1}} + \mathbf{M_{a2}} + \mathbf{M_{G2}}\end{aligned} \tag{17}$$

3.2 Aerodynamic moment characterization under asymmetric morphing

Examining the right-hand side of (17), the resultant external torque acting on the complete vehicle can be partitioned into three physically distinct contributions: the aerodynamic moments $\mathbf{M_{a0}}$, $\mathbf{M_{a1}}$, and $\mathbf{M_{a2}}$; supplementary torques arising from the inertial coupling induced by the morphing process itself; and disturbance moments unique to asymmetric wing configurations—these encompass aerodynamic asymmetry effects, gravity-induced imbalances, and additional cross-coupling disturbances. The morphing-induced inertial terms are governed by multibody kinematics and are already accounted for on the left-hand side of (17).

Because aerodynamic moment generation depends on a complex interplay of flow conditions, vehicle geometry, and surface deflections, a closed-form expression derived from first principles is seldom attainable. The conventional approach, adopted here, resolves the aerodynamic moment vector into body-axis components M_x, M_y, and M_z, each expressed as the product of the dynamic pressure, a reference geometric length or area, and a corresponding dimensionless moment coefficient:

$$\begin{cases} M_x = QS_{ref}L_{ref}C_{mx} \\ M_y = QS_{ref}L_{ref}C_{mz} \\ M_z = QS_{ref}L_{ref}C_{my} \end{cases} \tag{18}$$

The three body-axis aerodynamic moment coefficients are: the rolling moment coefficient C_{mx}, the yawing moment coefficient C_{my}, and the pitching moment coefficient C_{mz}. In the definitions above, $Q = \frac{1}{2}\rho V^2$ represents the freestream dynamic pressure, L_{ref} is the mean aerodynamic chord, and S_{ref} denotes the wing reference area. These coefficients are constructed from stability-derivative expansions as summarized below. Throughout this paper, C_{mx}, C_{my}, and C_{mz} denote the rolling, yawing, and pitching moment coefficients, respectively.

Rolling moment coefficient

$$C_{mx} = C_{mx,0} + C_{mx}^{\beta}\beta + \frac{C_{mx}^{\omega_x}\omega_x L_{ref}}{2V} + C_{mx}^{\delta_a}\delta_a + C_{mx}^{\delta_e}\delta_e + C_{mx}^{\delta_r}\delta_r + w_{mx}$$

Yawing moment coefficient

$$C_{my} = C_{my,0} + C_{my}^{\beta}\beta + \frac{C_{my}^{\omega_z}\omega_z L_{ref}}{2V} + \frac{C_{my}^{\omega_x}\omega_x L_{ref}}{2V} + C_{my}^{\delta_a}\delta_a + C_{my}^{\delta_e}\delta_e + C_{my}^{\delta_r}\delta_r + w_{my}$$

Pitching moment coefficient

$$C_{mz} = C_{mz,0} + C_{mz}^{\alpha}\alpha + \frac{C_{mz}^{\omega_y}\omega_y \bar{c}}{V} + C_{mz}^{\delta_e}\delta_e + C_{mz}^{\delta_a}\delta_a + w_{mz}$$

In the preceding set of aerodynamic coefficient formulas, $(\omega_x,\ \omega_y,\ \omega_z)$ denotes the body-axis angular rate components (roll, pitch, and yaw rates); (w_{mx}, w_{my}, w_{mz}) is

the incremental aerodynamic moment coefficient perturbation produced by configurational asymmetry; and $\bar{c}$ is the mean aerodynamic chord. The subscript 0 designates the baseline (reference-configuration) value of each aerodynamic derivative, while the superscript notation ($\alpha,\ \beta,\ \delta_e,\ \delta_a,\ \delta_r, \omega_x,\ \omega_y,\ \omega_z$) indicates a partial derivative of the respective coefficient with respect to the local airflow angle, a control-surface deflection, an angular rate component, or an asymmetry parameter.

3.3 State-space formulation for control synthesis

The kinematic relationship governing the orientation of the morphing aircraft involves the attitude state vector $\chi_{\mathbf{1}} = [\vartheta,\ \ \psi,\ \ \varphi\ ]^T$. Its evolution is described by the standard Euler-angle kinematic differential equation

$$\dot{\chi}_{\mathbf{1}} = \mathbf{R}_{\mathbf{gtb}}^{\mathbf{T}} \mathbf{\Omega_0} \tag{19}$$

where $\mathbf{R_{gtb}}$ denotes the direction-cosine transformation (rotation matrix) that maps vectors from the Earth-fixed inertial frame into the fuselage body frame; it is parameterized by the Euler triplet $\chi_{\mathbf{1}}$. The vector $\chi_{\mathbf{1}} = (\vartheta, \psi, \varphi)^T$ collects the three Euler angles, and $\mathbf{\Omega_0} = (\omega_{\mathbf{x}}, \omega_{\mathbf{y}}, \omega_{\mathbf{z}})^T$ is the fuselage angular velocity expressed in body coordinates.

$$\mathbf{R_{gtb}}(\chi_{\mathbf{1}}) = \begin{bmatrix} \cos\psi\cos\varphi & \cos\psi\sin\varphi & -\sin\psi \\ -\cos\vartheta\sin\varphi + \sin\vartheta\sin\psi\cos\varphi & \cos\vartheta\cos\varphi + \sin\vartheta\sin\psi\sin\varphi & \sin\vartheta\cos\psi \\ \sin\vartheta\sin\varphi + \cos\vartheta\sin\psi\cos\varphi & -\sin\vartheta\cos\varphi + \cos\vartheta\sin\psi\sin\varphi & \cos\vartheta\cos\psi \end{bmatrix}$$

Select $\mathbf{\Omega_0}$ as the composite system state, i.e., $\chi_{\mathbf{2}} = \mathbf{\Omega_0} = [\omega_x,\ \omega_y,\ \omega_z]^T$ and $\mathbf{u} = [\delta_a, \delta_r, \delta_e]^T$. Assembling the dynamical relationships derived in the preceding subsections—specifically, consolidating Equations (17) and (18) with the kinematic equation above—the complete angular motion of the morphing vehicle admits the following compact state-space description:

$$\begin{cases} \dot{\chi}_1 = \mathbf{\Phi}_1\left(\chi_1, \chi_2, \mathbf{w}_1\right) = \mathbf{f}_1\left(\chi_1\right) \cdot \chi_2 + \mathbf{w}_1 \\ \dot{\chi}_2 = \mathbf{\Phi}_2\left(\chi_1, \chi_2, \mathbf{u}, \mathbf{w}_2\right) \end{cases} \tag{20}$$

In the system representation (20), the kinematic drift term is $\mathbf{f_1}\left(\chi_{\mathbf{1}}\right) = \mathbf{R_{gtb}}^T\left(\chi_{\mathbf{1}}\right)$ (see (19)), while the angular-rate drift $\mathbf{\Phi_2}\left(\chi_{\mathbf{1}}, \chi_{\mathbf{2}}, u, \mathbf{w_2}\right)$ is obtained by combining the multibody rotational dynamics of (17) with the aerodynamic moment model of (18). The vector signals $\mathbf{w_1}$ and $\mathbf{w_2}$ aggregate, respectively, the exogenous disturbances and the structured as well as unstructured uncertainties affecting the kinematic loop and the angular-rate dynamics.

$$\begin{cases} \eta_1 = \chi_1 \\ \eta_2 = \dot{\chi}_1 = \mathbf{\Phi}_1\left(\chi_1, \chi_2, \mathbf{w}_1\right) = \mathbf{f}_1\left(\chi_1\right) \cdot \chi_2 + \mathbf{w}_1 \end{cases}$$

Consequently,

$$\begin{cases} \dot{\eta}_1 = \eta_2 \\ \dot{\eta}_2 = \frac{d\mathbf{f}_1(\chi_1)}{dt}\chi_2 + \mathbf{f}_1(\chi_1)\,\mathbf{\Phi}_2(\chi_1, \chi_2, \mathbf{u}, \mathbf{w}_2) + \frac{d\mathbf{w}_1}{dt} = \mathbf{D}(\chi_1, \chi_2, \mathbf{u}) + \mathbf{g_r u} + \mathbf{w}_0 \end{cases} \tag{21}$$

Here $\mathbf{w_0}$ aggregates the effect of external perturbations and modeling inaccuracies into a single lumped disturbance term entering the angular-rate dynamics channel. Under the assumption that the input mapping is norm-bounded, the channel is split into the known component $\mathbf{g_r u}$ and the residual lumped term $\mathbf{w_0}$, which may be partially known a priori and is otherwise estimated online by the observer designed in Section 4.2.

4 TD3-MRAC: a learning-based model reference adaptive control framework

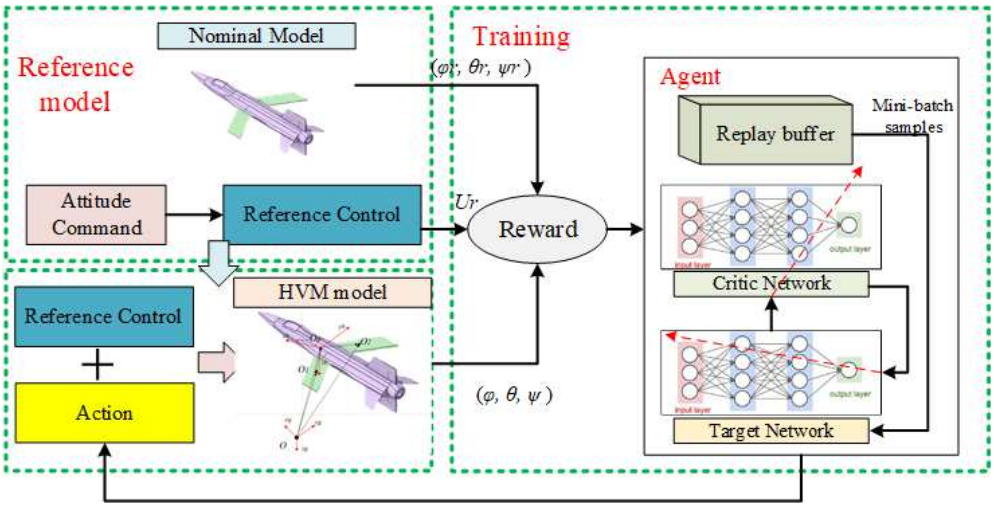


Fig. 3: Architecture of the proposed TD3-MRAC attitude controller

The overall architecture of the proposed TD3-MRAC attitude controller is illustrated in Fig. 3. The design philosophy is grounded in the MRAC paradigm: a fixed-wing aircraft model is adopted as the reference model that encodes the desired closed-loop behavior, and an adaptive mechanism is employed to drive the morphing vehicle to replicate this reference response in the presence of configuration-dependent uncertainties and external disturbances. In conventional MRAC, the adaptation law is derived analytically—typically from Lyapunov stability criteria or gradient-based update rules. The central idea of TD3-MRAC is to replace the hand-crafted adaptation law with a learned policy, realized by the TD3 algorithm, which is trained offline to generate corrective control increments that compensate for the structured

mismatch between the morphing vehicle and the reference model. A baseline backstepping controller, synthesized from the known reference dynamics, provides a stabilizing feedforward-feedback backbone, while the TD3 adaptation module supplies an online correction that refines the control signal in response to time-varying morphing effects and unmodeled dynamics. This separation of concerns—nominal stabilization via classical control and adaptive compensation via deep RL—preserves the interpretability and safety guarantees of the baseline design while leveraging the expressive power of neural networks to handle complex, high-dimensional uncertainties. The closed-loop stability of this architecture is established in Section 4.3.

4.1 Reference model and baseline controller

The reference model—taken as the dynamics of a conventional fixed-wing aircraft—defines the target closed-loop response that the morphing vehicle is required to emulate. Let the reference dynamics be expressed in state-space form as

$$\begin{cases} \dot{\mathbf{x}}_{1r} = \mathbf{x}_{2r} \\ \dot{\mathbf{x}}_{2r} = \varphi_r\left(\mathbf{x}_{1r}, \mathbf{x}_{2r}\right) + \mathbf{g_r u_r} \end{cases} \tag{22}$$

Since both $\varphi_r(\mathbf{x_{1r}}, \mathbf{x_{2r}})$ and g_r in (22) are fully known by construction, a baseline nonlinear controller u_r can be designed to stabilize the reference model and enforce accurate command following. In this work, a backstepping control law is adopted as the nominal design. Let x_d be the commanded attitude trajectory and define the tracking error as $\varepsilon_{\mathbf{1r}} = \mathbf{x_d} - \mathbf{x_{1r}}$. A virtual control is then introduced as follows, where $\mathbf{x_d} = [\phi_d, \theta_d, \psi_d]^T$ denotes the HMV attitude command and $\mathbf{x_{1r}} = [\phi_r, \theta_r, \psi_r]$ is the attitude angle of the reference model.

$$\mathbf{x_{2rd}} = \mathbf{C_1}\varepsilon_{\mathbf{1r}} + \dot{\mathbf{x}}_{\mathbf{d}}$$

with $\mathbf{C_1} > 0$ being a positive design parameter. Let the virtual control tracking error be $\varepsilon_{\mathbf{2r}} = \mathbf{x_{2rd}} - \mathbf{x_{2r}}$. The baseline (reference) control law is then constructed as

$$\mathbf{u_r} = \mathbf{g_r}^{-1}\left(\mathbf{C_2}\varepsilon_{\mathbf{2r}} + \varepsilon_{\mathbf{1r}} + \dot{\mathbf{x}}_{\mathbf{2rd}} - \varphi_r\left(\mathbf{x_{1r}}, \mathbf{x_{2r}}\right)\right) \tag{23}$$

where $\mathbf{C_2} > 0$ is a second tunable gain and $\mathbf{u_r} = [\delta_{ar}, \delta_{er}, \delta_{rr}]$ denotes the resulting baseline control signal. Because the reference model is completely characterized, alternative nonlinear design techniques—feedback linearization, sliding-mode control, or active disturbance rejection control (ADRC)—are equally applicable in this role. In scenarios where the morphing aircraft dynamics are highly complex and the internal reconfiguration effects resist explicit parametric description, an ADRC-type baseline may be preferable [37, 38]. The baseline controller serves as the fixed-structure component of the TD3-MRAC architecture, guaranteeing nominal stability of the reference loop, as proven in Theorem 1 (Section 4.3). Within the overall architecture the reference loop is implemented internally as a command shaper: it is simulated inside the controller using the known fixed-wing dynamics and generates the reference trajectory $\mathbf{x}_{1r}(t)$, $\mathbf{x}_{2r}(t)$ and their derivatives, which the plant-side tracking law of Section 4.3 is designed to follow.

4.2 TD3-based adaptive compensation mechanism

4.2.1 Extended state observation and problem formulation

A fundamental practical consideration in MRAC implementation is that not all internal states of the controlled plant may be directly measurable. An Extended State Observer (ESO) is therefore integrated into the TD3-MRAC framework to reconstruct the full state vector and to estimate the lumped disturbance acting on the system. The ESO augments the plant model by treating the total disturbance as an additional state variable; all states—including the disturbance estimate—are then recovered from the available input-output data. A key practical merit of this scheme is that it requires only low-order output measurements to simultaneously observe the plant states and the aggregate uncertainty.

Consider a second-order nonlinear plant of the form

$$\begin{cases} \dot{\mathbf{x}}_1 = \mathbf{x}_2 \\ \dot{\mathbf{x}}_2 = f\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{u}\right) + \mathbf{d} \end{cases} \tag{24}$$

For this plant, the observer is designed in the linear extended-state-observer (LESO) form following Han [39] and the bandwidth parameterization of Gao [40], which treats the total disturbance $\mathbf{h} := \mathbf{f}\left(\mathbf{x_1}, \mathbf{x_2}, u\right) + \mathbf{d}$ as an additional state. Let $\mathbf{y} = \mathbf{x_1}$ denote the measured low-order output (the attitude angles). The third-order LESO reads

$$\dot{\widehat{\mathbf{z}}}_1 = \widehat{\mathbf{z}}_2 + 3\omega_\mathbf{o}\left(\mathbf{y} - \widehat{\mathbf{z}}_1\right) \tag{25}$$

$$\dot{\widehat{\mathbf{z}}}_2 = \widehat{\mathbf{z}}_3 + 3{\omega_\mathbf{o}}^2\left(\mathbf{y} - \widehat{\mathbf{z}}_1\right) + \mathbf{g_r}\mathbf{u} \tag{26}$$

$$\dot{\widehat{\mathbf{z}}}_3 = {\omega_\mathbf{o}}^3\left(\mathbf{y} - \widehat{\mathbf{z}}_1\right) \tag{27}$$

where $\omega_\mathbf{o} > 0$ is the observer bandwidth and $\widehat{\mathbf{z}}_\mathbf{1}$, $\widehat{\mathbf{z}}_\mathbf{2}$, $\widehat{\mathbf{z}}_\mathbf{3}$ estimate $\mathbf{x_1}$, $\mathbf{x_2}$, and $\mathbf{h}$, respectively; the gains follow the bandwidth parameterization $(\beta_\mathbf{1}, \beta_\mathbf{2}, \beta_\mathbf{3}) = \left(3\omega_\mathbf{o}, 3{\omega_\mathbf{o}}^2, {\omega_\mathbf{o}}^3\right)$ and correspond to the ESO entries of Table 2. The disturbance estimate is used for feedforward compensation,

$$\mathbf{u_{ff}} = -{\mathbf{g_r}}^{-1}\widehat{\mathbf{z}}_\mathbf{3} \tag{28}$$

which removes the estimated lumped disturbance from the input channel; the residual estimation error is proven small in Theorem 2 (Section 4.3). Defining the estimation error

$$\mathbf{e} = \left[\widehat{\mathbf{z}}_\mathbf{1} - \mathbf{x_1}, \widehat{\mathbf{z}}_\mathbf{2} - \mathbf{x_2}, \widehat{\mathbf{z}}_\mathbf{3} - \mathbf{h}\right]^T \tag{29}$$

its dynamics read

$$\dot{\mathbf{e}} = \mathbf{A_e}\mathbf{e} + \mathbf{B_e}\dot{\mathbf{h}} \tag{30}$$

where $\mathbf{A_e}$ is Hurwitz with all eigenvalues at $-\omega_\mathbf{o}$, and the input-to-state stability of the error system with respect to $\dot{\mathbf{h}}$ is established in Theorem 2.

The baseline control law derived from the reference model provides a feedforward component that enables satisfactory command tracking when the morphing-induced deviations from the reference dynamics are small. The role of the TD3 adaptation module is to generate an incremental correction that compensates for the residual mismatch, thereby driving the closed-loop behavior of the morphing vehicle toward that of the reference model. Two complementary objectives thus define the MRAC problem: (a) the plant output should accurately track the reference model response, and (b) the tracking error with respect to the external attitude command should be minimized.

Given the reference model (22) and the true plant dynamics (24), the structural mismatch between them is quantified as

$$\begin{cases} \dot{\epsilon}_1 = \dot{\mathbf{x}}_1 - \dot{\mathbf{x}}_{1r} = \mathbf{x}_2 - \mathbf{x}_{2r} \\ \dot{\epsilon}_2 = \dot{\mathbf{x}}_2 - \dot{\mathbf{x}}_{2r} = f\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{u}\right) + \mathbf{d} - \varphi_r\left(\mathbf{x}_{1r}, \mathbf{x}_{2r}\right) - \mathbf{g_r}\mathbf{u_r} \end{cases} \tag{31}$$

Under mild regularity conditions, the implicit function theorem permits $f(\mathbf{x_1}, \mathbf{x_2}, \mathbf{u})$ to be reparameterized as $F(\mathbf{x_1}, \mathbf{x_2}, \mathbf{u}) + \mathbf{g_r}\left(\mathbf{\Delta u} + \mathbf{u_r}\right)$, where $F(\mathbf{x_1}, \mathbf{x_2}, \mathbf{u})$ subsumes the unresolved (unknown) dynamics and $\mathbf{\Delta u}$ is the adaptive compensation term to be supplied by the TD3 policy. Equation (31) can therefore be rewritten in the form

$$\begin{cases} \dot{\epsilon}_1 = \mathbf{x}_2 - \mathbf{x}_{2r} \\ \dot{\epsilon}_2 = \left(\mathbf{F} - \varphi_\mathbf{r}\right) + \mathbf{g_r}\mathbf{\Delta u} + \mathbf{d} = \mathbf{g_r}\mathbf{\Delta u} + \mathbf{D} \end{cases} \tag{32}$$

In (32), $\mathbf{D}$ aggregates all unmodeled dynamics, parametric uncertainties, and external perturbations into a single total disturbance signal. The term $\mathbf{\Delta u} = [\Delta\delta_{ar}, \Delta\delta_{\varepsilon r}, \Delta\delta_{rr}]$ is the adaptive control increment—this is precisely the quantity that the TD3-MRAC framework learns to generate, replacing the analytically derived adaptation law of classical MRAC with a data-driven policy optimized through reinforcement learning.

4.2.2 TD3-MRAC: learning the adaptive compensation policy

In the TD3-MRAC framework, the adaptive compensation policy is trained via the TD3 algorithm operating on a MDP formulation of the flight control problem. As established in Section 2, an MDP is characterized by the tuple (s_t, a_t, r_t, s_{t+1}). For the present MRAC setting, these components are specialized as follows to capture the interplay between the reference model, the baseline controller, the ESO, and the learned adaptation mechanism.

(i) State space

The state representation in TD3-MRAC must furnish the agent with sufficient contextual information to produce effective adaptive compensation on top of the existing baseline controller. Unlike a stand-alone RL controller that acts in isolation, the TD3-MRAC agent operates within a structured control architecture: it receives the observed states from the ESO, the current output of the baseline controller, and a record of its own recent compensatory actions. The composite state vector is assembled as follows:

$$\mathbf{S_t} = \Big[\mathbf{k_1}\left((\mathbf{E_1}(t))^2\right), \mathbf{k_2}\left((\mathbf{E_2}(t))^2\right), \mathbf{k_3}\left(\mathbf{\Delta u}^2\right), \mathbf{k_4}\left({\varepsilon_1}^2\right), \mathbf{k_5}\left({\varepsilon_2}^2\right), \mathbf{k_6}\left(\mathbf{\Lambda}(t)\right), \mathbf{k_7}\left(\widehat{\mathbf{D}}(t)\right), \mathbf{k_8}\left(\mathbf{u_r}\right), \mathbf{k_9}\left(\mathbf{S}_{t-2}\right)\Big] \quad (33)$$

which expands to

$$S_{sub}(t) = [\|\mathbf{E_1}(t)\|^2, \|\mathbf{E_2}(t)\|^2, \|\mathbf{\Delta u}\|^2, \|\epsilon_1\|^2, \Lambda(t), \widehat{\mathbf{D}}(t), \mathbf{u_r}]$$

Remark 1 (State space design rationale) In (33), $\mathbf{k_i}$ $(i = 1, 2, \ldots, 9)$ are diagonal normalization matrices that scale the training inputs to a numerically well-conditioned range, facilitating stable convergence of the TD3 algorithm. The state vector incorporates the following components, each motivated by the MRAC structure: the three-channel attitude tracking error $\mathbf{E}(t)$ between the morphing vehicle and the attitude command; the model-matching error $\varepsilon(t)$ between the plant and the reference model, which directly encodes the MRAC objective of emulating the reference behavior; the baseline control output $\mathbf{u_r}$, which informs the agent of the current nominal control action; the agent's own compensatory increment $\mathbf{\Delta u}$ from the preceding time step; the total disturbance estimate $\widehat{\mathbf{D}}(t)$ supplied by the ESO; the morphing command $\mathbf{\Lambda}$ that specifies the current wing configuration; and a short history buffer $\mathbf{S}_{t-2} = [\mathbf{u}_{t-2}, \Delta u_{t-2}]$ of the combined control effect, enabling the agent to assess how its past compensatory actions interacted with the baseline controller and to learn a temporally coherent adaptation strategy.

(ii) Action space

The action corresponds to the RL-generated compensatory increment that serves as the adaptive control term in the MRAC architecture. To respect the physical saturation limits of the control surfaces, the action space is bounded as

$$\begin{aligned} \mathbf{a_t} = \mathbf{\Delta u} &= [\Delta\delta_a, \Delta\delta_\varepsilon, \Delta\delta_r] \\ \Delta u &\in [\mathbf{\Delta \underline{u}},\ \mathbf{\Delta \overline{u}}] \end{aligned} \quad (34)$$

where $\mathbf{\Delta \underline{u}}$ and $\mathbf{\Delta \overline{u}}$ denote the lower and upper saturation limits of the compensation action, respectively.

(iii) Reward function

In the TD3-MRAC framework, the reward function encodes the MRAC performance objectives into a scalar learning signal. Its design critically influences the convergence behavior, steady-state accuracy, and closed-loop stability of the trained policy. The reward is structured to simultaneously enforce several desiderata: minimizing the attitude tracking error, maintaining fidelity to the reference model response, limiting control effort and actuation rate, and respecting safety margins. According to their functional roles, the reward terms are organized into 8 components.

$$
\begin{aligned}
\mathbf{r_1} &= -\mathbf{w_1}\left(\mathbf{E_1}(t)\right)^2 \\
\mathbf{r_2} &= -\mathbf{w_2}\left(\mathbf{E_2}(t)\right)^2 \\
\mathbf{r_3} &= -\mathbf{w_3}{\varepsilon_1}^2 \\
\mathbf{r_4} &= -\mathbf{w_4}{\varepsilon_2}^2 \\
\mathbf{r_5} &= -\mathbf{w_5}\left\|\mathbf{u_r} + \mathbf{\Delta u}\right\| \\
\mathbf{r_6} &= \mathbf{w_6}\mathbb{1}\left(\left\|\varepsilon_1(t)\right\| \leq \bar{\varepsilon}\right) \\
\mathbf{r_7} &= -\mathbf{w_7}\left\|\mathbf{a_t} - 2\mathbf{a}_{t-1} + \mathbf{a}_{t-2}\right\|^2 \\
\mathbf{r_8} &= -\mathbf{w_8}\left\|\varepsilon_\mathbf{t} - 2\varepsilon_{t-1} + \varepsilon_{t-2}\right\|^2
\end{aligned}
\tag{35}
$$

The term $\mathbf{r_1}$ penalizes the attitude tracking residual between the morphing vehicle state and the commanded attitude, while $\mathbf{r_3}$ penalizes the model-matching discrepancy between the plant response and the reference model output—this second penalty directly embodies the MRAC objective of enforcing reference model following. The terms $\mathbf{r_2}$ and $\mathbf{r_4}$ suppress the time derivatives of the above errors to limit rapid transients and improve the damping characteristics of the closed-loop response. The term $\mathbf{r_5}$ penalizes the total control magnitude, thereby limiting control effort and promoting smooth actuation. The term $\mathbf{r_6}$ provides a sparse bonus reward when the model-matching error falls below a prescribed tolerance band. To further suppress oscillatory behavior, two second-order penalty terms are introduced: $\mathbf{r_7}$ penalizes the second-order temporal difference of the compensation action, which directly quantifies high-frequency chattering of the control surfaces—minimizing it promotes temporally coherent actuation and reduces mechanical wear on the servo mechanisms; and $\mathbf{r_8}$ penalizes the second-order difference of the attitude tracking error, which captures oscillatory ringing in the closed-loop attitude response, improving transient smoothness and preventing overshoot cycles.

4.3 Stability analysis of the TD3-MRAC closed loop

This section establishes the theoretical guarantees of the proposed framework. Following the separation principle of structured-learning control, the analysis proceeds in three layers: (i) the reference closed loop formed by the fixed-wing reference model and the baseline backstepping controller, which is proven globally exponentially stable in Theorem 1; (ii) the linear extended state observer, whose estimation error is proven input-to-state stable with respect to the derivative of the total disturbance, with a steady-state bound tunable through the observer bandwidth (Theorem 2); and (iii) the composite closed loop including the bounded learned compensation policy, which is proven uniformly ultimately bounded with an explicit ultimate bound expressed in terms of the design parameters (Theorem 3, Corollary 1). Importantly, these guarantees require only boundedness of the learned action (Lemma 1) and do not rely on the convergence of the TD3 algorithm; the role of learning is to improve the closed-loop performance within the guaranteed tube, as quantified in Section 5.

4.3.1 Standing assumptions

The analysis is carried out under the following standing assumptions, each of which is standard in the flight-control literature [32, 41] and is verified numerically for the morphing vehicle in Section 5.

A1 (Bounded command). The attitude command and its first two derivatives are bounded and piecewise continuous: $\|\mathbf{x_d}\| \leq \overline{\mathbf{x}}_{d1}, \|\dot{\mathbf{x}_d}\| \leq \overline{\mathbf{x}}_{d2}, \|\ddot{\mathbf{x}_d}\| \leq \overline{\mathbf{x}}_{d3}$

A2 (Nonsingular input channel). The reference-model input channel is known and nonsingular: $\det(\mathbf{g_r}) \neq 0$.

A3 (Bounded total disturbance). The total disturbance $\mathbf{h} := \mathbf{f}(\mathbf{x_1}, \mathbf{x_2}, u) + \mathbf{d}$ and its derivative are bounded on the operating envelope: $\|\mathbf{h}\| \leq \overline{\mathbf{h}}, \left\|\dot{\mathbf{h}}\right\| \leq \overline{\mathbf{H}}$.

A4 (Morphing envelope). The morphing command evolves within a compact set with bounded rate: $\mathbf{\Lambda} \in \overline{\mathbf{\Lambda}}, \left\|\dot{\mathbf{\Lambda}}\right\| \leq \dot{\mathbf{\Lambda}}_{\max}$, and the induced deviations of the inertia, aerodynamic-coefficient, and input-channel maps from the fixed-wing reference model are uniformly bounded on the envelope.

A5 (Bounded input-channel mismatch). The input-channel mismatch satisfies $\|\mathbf{\Delta g}(\mathbf{x_1}, \Lambda)\| \leq \overline{\mathbf{g}}$, and the control input respects the actuator limits $\|\mathbf{u}\| \leq \overline{\mathbf{u}}$.

A6 (Bounded kinematic disturbance). The kinematic disturbance is bounded: $\|\mathbf{w_1}\| \leq \overline{\mathbf{w}}_{\mathbf{1}}$.

A7 (Bounded compensation action). The learned compensation satisfies $\|\mathbf{\Delta u}\| \leq \overline{\mathbf{a}}$; by Lemma 1 this is a design guarantee rather than an assumption.

A8 (Positive baseline gains). The baseline gains are $\mathbf{C_1} = \mathbf{c_1 I_3}, \mathbf{C_2} = \mathbf{c_2 I_3}$ with $\mathbf{c_1}, \mathbf{c_2} > 0$.

4.3.2 Stability of the reference closed loop

The reference loop consists of the reference model (22) closed by the backstepping controller (23). Its stability is established first, since the reference states $\mathbf{x_{1r}}$, $\mathbf{x_{2r}}$ and their derivatives serve as the bounded tracking objective of the actual vehicle.

Theorem 1 (Global exponential stability of the reference closed loop) *Consider the reference model (22) driven by a command satisfying A1, and the backstepping controller*

$$\varepsilon_{\mathbf{1r}} = \mathbf{x_d} - \mathbf{x_{1r}} \tag{36}$$

$$\varepsilon_{\mathbf{2r}} = \mathbf{x_{2rd}} - \mathbf{x_{2r}} \tag{37}$$

$$\mathbf{u_r} = \mathbf{g_r}^{-1}\left[\varepsilon_{\mathbf{1r}} + \mathbf{C_2}\varepsilon_{\mathbf{2r}} + \dot{\mathbf{x}_{\mathbf{2rd}}} - \varphi_{\mathbf{r}}(\mathbf{x_{1r}}, \mathbf{x_{2r}})\right] \tag{38}$$

with $\mathbf{c_1}, \mathbf{c_2} > 0$ *and* $\det(\mathbf{g_r}) \neq 0$. *Then the reference tracking errors obey*

$$\dot{\varepsilon_{\mathbf{1r}}} = -\mathbf{C_1}\varepsilon_{\mathbf{1r}} + \varepsilon_{\mathbf{2r}} \tag{39}$$

$$\dot{\varepsilon_{\mathbf{2r}}} = -\varepsilon_{\mathbf{1r}} - \mathbf{C_2}\varepsilon_{\mathbf{2r}} \tag{40}$$

and the origin of the error system is globally exponentially stable:

$$\left\|\left[{\varepsilon_{\mathbf{1r}}}^T, {\varepsilon_{\mathbf{2r}}}^T\right]^T(t)\right\| \leq \left\|\left[{\varepsilon_{\mathbf{1r}}}^T, {\varepsilon_{\mathbf{2r}}}^T\right]^T(0)\right\| \exp\left(-\min\{\mathbf{c_1}, \mathbf{c_2}\}\, t\right) \tag{41}$$

Proof Consider the Lyapunov function candidate

$$\mathbf{V_r} = \frac{1}{2}{\varepsilon_{\mathbf{1r}}}^T \varepsilon_{\mathbf{1r}} + \frac{1}{2}{\varepsilon_{\mathbf{2r}}}^T \varepsilon_{\mathbf{2r}} \tag{42}$$

Along the trajectories of (39)–(40), the cross terms cancel and

$$\dot{\mathbf{V}}_\mathbf{r} = -\mathbf{c_1} \left\|\varepsilon_{\mathbf{1r}}\right\|^2 - \mathbf{c_2} \left\|\varepsilon_{\mathbf{2r}}\right\|^2 \leq -2 \min\left\{\mathbf{c_1}, \mathbf{c_2}\right\} \mathbf{V_r} \tag{43}$$

and the claim follows from the comparison lemma [41]. Consequently, the reference states and their derivatives remain bounded for all $\mathbf{t} \geq 0$ and define a valid tracking objective for the plant-side law (53).

□

4.3.3 Convergence of the extended state observer

The convergence properties of the LESO (25)–(27) are stated next; the proof uses the standard bandwidth-scaling argument of [40].

Theorem 2 (Input-to-state stability of the LESO) *Consider the plant (24) under A3 and the LESO (25)–(27). The estimation error (29) obeys (30), where $\mathbf{A_e}$ is Hurwitz with all eigenvalues at $-\omega_\mathbf{o}$, and the error system is input-to-state stable with respect to $\dot{\mathbf{h}}$:*

$$\left\|\mathbf{e}(t)\right\| \leq \kappa {\omega_\mathbf{o}}^2 \left\|\mathbf{e}(0)\right\| \exp\left(-\frac{\omega_\mathbf{o} t}{2\overline{\lambda}}\right) + \frac{\gamma_\mathbf{e}}{\omega_\mathbf{o}} \overline{\mathbf{H}} \tag{44}$$

where $\overline{\lambda} := \lambda_{\max}(\mathbf{P})$ and $\mathbf{P} = \mathbf{P}^T \succ 0$ solves the normalized Lyapunov equation $\overline{\mathbf{A}}^T \mathbf{P} + \mathbf{P}\overline{\mathbf{A}} = -\mathbf{I_9}$, with

$$\overline{\mathbf{A}} = \begin{bmatrix} -3\mathbf{I_3} & \mathbf{I_3} & 0 \\ -3\mathbf{I_3} & 0 & \mathbf{I_3} \\ -\mathbf{I_3} & 0 & 0 \end{bmatrix};$$

the constants are $\kappa = \sqrt{\frac{\overline{\lambda}}{\lambda_{\min}(\mathbf{P})}}$ and $\gamma_\mathbf{e} = \frac{2\overline{\lambda}\|\mathbf{P}\|}{\lambda_{\min}(\mathbf{P})}$. In particular, $\lim_{t\to\infty} \|\mathbf{e}(t)\| \leq \frac{\gamma_\mathbf{e}\overline{\mathbf{H}}}{\omega_\mathbf{o}}$, i.e., the steady-state estimation error is $O\left(\frac{\overline{\mathbf{H}}}{\omega_\mathbf{o}}\right)$ and decreases as the bandwidth increases, up to the noise–bandwidth tradeoff of Remark 2.

Proof Apply the scaling transformation

$$\eta = T(\omega_\mathbf{o})\, \mathbf{e} \tag{45}$$

$$\dot{\eta} = \omega_\mathbf{o} \overline{\mathbf{A}} \eta + \mathbf{B_0} \dot{\mathbf{h}} \tag{46}$$

with $T(\omega_\mathbf{o}) = \operatorname{diag}\left({\omega_\mathbf{o}}^2 \mathbf{I_3}, \omega_\mathbf{o} \mathbf{I_3}, \mathbf{I_3}\right)$ and $\mathbf{B_0} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{I_3} \end{bmatrix}$. Since $\overline{\mathbf{A}}$ is Hurwitz, $\mathbf{P}$ as above exists, and with $\mathbf{V_e} = \eta^T \mathbf{P} \eta$ one obtains, using $\lambda_{\min}(\mathbf{P}) \|\eta\|^2 \leq \mathbf{V_e} \leq \overline{\lambda} \|\eta\|^2$ and $\left\|\eta^T \mathbf{P} \mathbf{B_0} \dot{\mathbf{h}}\right\| \leq \|\mathbf{P}\| \|\eta\| \overline{\mathbf{H}}$,

$$\mathbf{V_e} = \eta^T \mathbf{P} \eta \tag{47}$$

$$\dot{\mathbf{V}}_{\mathbf{e}} \leq -\frac{\omega_{\mathbf{o}}}{\overline{\lambda}}\mathbf{V}_{\mathbf{e}} + 2\left\|\mathbf{P}\right\|\overline{\mathbf{H}}\sqrt{\frac{\mathbf{V}_{\mathbf{e}}}{\lambda_{\min}\left(\mathbf{P}\right)}} \tag{48}$$

With $\mathbf{W} := \sqrt{\mathbf{V}_{\mathbf{e}}}$, the comparison lemma [41] yields

$$\mathbf{W}(t) \leq \mathbf{W}(0)\exp\left(-\frac{\omega_{\mathbf{o}}t}{2\overline{\lambda}}\right) + \frac{2\overline{\lambda}\left\|\mathbf{P}\right\|\overline{\mathbf{H}}}{\omega_{\mathbf{o}}\sqrt{\lambda_{\min}\left(\mathbf{P}\right)}}\left(1-\exp\left(-\frac{\omega_{\mathbf{o}}t}{2\overline{\lambda}}\right)\right) \tag{49}$$

Mapping back with $\left\|\mathbf{e}\right\| \leq \left\|\eta\right\| \leq \frac{\mathbf{W}}{\sqrt{\lambda_{\min}(\mathbf{P})}}$ for $\omega_{\mathbf{o}} \geq 1$ and $\mathbf{W}(0) \leq \sqrt{\overline{\lambda}}\left\|\eta(0)\right\| \leq \sqrt{\overline{\lambda}}{\omega_{\mathbf{o}}}^{2}\left\|\mathbf{e}(0)\right\|$ yields (44).

□

Remark 2 (Noise–bandwidth tradeoff) With measurement noise of intensity $\sigma_{\mathbf{n}}$ in $\mathbf{y}$, the steady-state estimation variance grows polynomially with $\omega_{\mathbf{o}}$ (the noise gain of the $\widehat{\mathbf{z}}_{\mathbf{2}}$ channel scales as $O\left(\omega_{\mathbf{o}}\right)$ and that of $\widehat{\mathbf{z}}_{\mathbf{3}}$ as $O\left({\omega_{\mathbf{o}}}^{3}\right)$), so $\omega_{\mathbf{o}}$ must balance disturbance-rejection speed against noise amplification; this tradeoff is inherited from all high-gain observers [41].

Remark 3 (Peaking) The transient factor $\kappa{\omega_{\mathbf{o}}}^{2}\left\|\mathbf{e}(0)\right\|$ in (44) reflects the standard peaking behavior of high-gain observers; it can be mitigated by saturating the observer output or initializing $\widehat{\mathbf{z}}_{\mathbf{1}}(0) = \mathbf{y}(0)$.

4.3.4 Boundedness of the learned compensation

Lemma 1 (Bounded compensation action) *If the TD3 actor outputs the compensation through a saturating layer—e.g., a tanh scaling or a hard clamp onto the box* (34)*—then*

$$\left\|\boldsymbol{\Delta}\mathbf{u}(t)\right\| \leq \overline{\mathbf{a}} := \sqrt{3}\cdot\max\left\{\left\|\boldsymbol{\Delta}\mathbf{u}_{lb}\right\|, \left\|\boldsymbol{\Delta}\mathbf{u}_{ub}\right\|\right\} \tag{50}$$

for all $\mathbf{t} \geq 0$*, deterministically and independently of the learning outcome. This lemma is the keystone of the separation principle: the stability results below require only that the learned policy be bounded, not that the RL training converges. Convergence of TD3 under nonlinear function approximation remains an open problem [36], and the present guarantees do not rely on it.*

4.3.5 Uniform ultimate boundedness of the composite closed loop

To make the composite error dynamics dissipative, the baseline law is evaluated on the actual plant state to track the reference trajectory, and the observer estimate is fed forward:

$$\mathbf{u} = \mathbf{u_b} + \mathbf{u_{ff}} + \Delta\mathbf{u} \tag{51}$$

$$\mathbf{e_1} = \mathbf{x_1} - \mathbf{x_{1r}}, \mathbf{x_{2d}} = \dot{\mathbf{x_{1r}}} - \mathbf{C_1}\mathbf{e_1}, \mathbf{e_2} = \mathbf{x_2} - \mathbf{x_{2d}} \tag{52}$$

$$\mathbf{u_b} = {\mathbf{g_r}}^{-1}\left[\dot{\mathbf{x_{2d}}} - \varphi_{\mathbf{r}}\left(\mathbf{x_1}, \mathbf{x_2}\right) - \mathbf{e_1} - \mathbf{C_2}\mathbf{e_2}\right] \tag{53}$$

Theorem 3 (Uniform ultimate boundedness of the composite closed loop) *Consider the plant (20)/(24) under A1–A8 and the composite law (51)–(53) with the feedforward (28). Define the residual mismatch*

$$\widetilde{\mathbf{D}} = (\mathbf{h} - \widehat{\mathbf{z}}_{\mathbf{3}}) - \varphi_{\mathbf{r}}(\mathbf{x_1}, \mathbf{x_2}) + \Delta g(\mathbf{x_1}, \Lambda)\,\mathbf{u} \tag{54}$$

whose norm satisfies $\left\|\widetilde{\mathbf{D}}\right\| \le \delta_{\mathbf{e}}(t) + \overline{\varphi}_{\mathbf{r}} + \overline{\mathbf{g}}\overline{\mathbf{u}}$ *with* $\delta_{\mathbf{e}}(t) := \frac{\gamma_{\mathbf{e}}\overline{\mathbf{H}}}{\omega_{\mathbf{o}}} + \kappa{\omega_{\mathbf{o}}}^2 \|\mathbf{e}(0)\| \exp\left(-\frac{\omega_{\mathbf{o}} t}{2\lambda}\right)$ *the time-decaying ESO error bound of Theorem 2, and* $\overline{\varphi}_{\mathbf{r}} := \sup \|\varphi_{\mathbf{r}}(\mathbf{x_1}, \mathbf{x_2})\|$ *over the operating envelope.*

Proof Substituting (51)–(53) into the plant (20) with the kinematic map $\mathbf{f_1} = \mathbf{I_3}$ valid near level flight (the general case follows by absorbing the bounded term $(\mathbf{f_1} - \mathbf{I_3})\,\mathbf{x_2}$ into $\mathbf{w_1}$), the error dynamics read

$$\dot{\mathbf{e_1}} = -\mathbf{C_1}\mathbf{e_1} + \mathbf{e_2} + \mathbf{w_1} \tag{55}$$

$$\dot{\mathbf{e_2}} = -\mathbf{e_1} - \mathbf{C_2}\mathbf{e_2} + \widetilde{\mathbf{D}} + \mathbf{g_r}\boldsymbol{\Delta}\mathbf{u} \tag{56}$$

where the derivation uses $\mathbf{g_r}\mathbf{u_b} = \dot{\mathbf{x}}_{\mathbf{2d}} - \varphi_{\mathbf{r}}(\mathbf{x_1}, \mathbf{x_2}) - \mathbf{e_1} - \mathbf{C_2}\mathbf{e_2}$ and $\mathbf{h} = \mathbf{f} + \mathbf{d}$. Consider the Lyapunov function candidate

$$\mathbf{V} = \frac{1}{2}\|\mathbf{e_1}\|^2 + \frac{1}{2}\|\mathbf{e_2}\|^2 \tag{57}$$

$$\dot{\mathbf{V}} \le -\mathbf{c_0}\|\mathbf{e}\|^2 + \overline{\mathbf{b}}\|\mathbf{e}\| \tag{58}$$

with $\overline{\mathbf{b}}$ as in (60), where the cross terms of the backstepping structure cancel and the bounds of A5–A7, Lemma 1, and Theorem 2 were used. With $\mathbf{W} := \|\mathbf{e}\| = \sqrt{2\mathbf{V}}$, $\dot{\mathbf{W}} \le -\mathbf{c_0}\mathbf{W} + \overline{\mathbf{b}}$, and the comparison lemma [41] gives the uniform ultimate boundedness bound

$$\|\mathbf{e}(t)\| \le \|\mathbf{e}(0)\| \exp(-\mathbf{c_0}t) + \frac{\overline{\mathbf{b}}}{\mathbf{c_0}}(1 - \exp(-\mathbf{c_0}t)) \tag{59}$$

with $\mathbf{c_0} := \min\{\mathbf{c_1}, \mathbf{c_2}\}$ and

$$\overline{\mathbf{b}} = \overline{\mathbf{w}}_{\mathbf{1}} + \overline{\varphi}_{\mathbf{r}} + \overline{\mathbf{g}}\overline{\mathbf{u}} + \|\mathbf{g_r}\|\,\overline{\mathbf{a}} + \frac{\gamma_{\mathbf{e}}\overline{\mathbf{H}}}{\omega_{\mathbf{o}}} \tag{60}$$

in particular, the closed loop is uniformly ultimately bounded with ultimate bound

$$\lim_{t\to\infty} \|\mathbf{e}(t)\| \le \frac{\overline{\mathbf{b}}}{\mathbf{c_0}} \tag{61}$$

□

The bound is explicit in all design parameters, decreases with $\mathbf{c_1}, \mathbf{c_2}$ and $\omega_{\mathbf{o}}$ (subject to Remark 2), and is independent of the RL training procedure.

Corollary 1 (Command tracking) *Under the conditions of Theorems 1–3,*

$$\|\mathbf{x_1}(t) - \mathbf{x_d}(t)\| \le \|\mathbf{e_1}(t)\| + \|\mathbf{x_{1r}}(t) - \mathbf{x_d}(t)\| \tag{62}$$

and hence $\lim_{t\to\infty} \|\mathbf{x_1}(t) - \mathbf{x_d}(t)\| \le \frac{\overline{\mathbf{b}}}{\mathbf{c_0}}$*, i.e., the vehicle tracks the commanded attitude within a computable tube of radius* $\frac{\overline{\mathbf{b}}}{\mathbf{c_0}}$*.*

Remark 4 (Role of the learned compensation and the feedforward) The bound (61) is a worst-case guarantee: the policy $\boldsymbol{\Delta u}$ is trained to reduce the residual mismatch, so the effective tracking error is typically well below $\frac{\overline{\mathbf{b}}}{\mathbf{c_0}}$; the Monte Carlo study of Section 5.3 quantifies this tightening empirically. If the feedforward (28) is omitted, Theorem 3 remains valid with the ESO-error bound replaced by the total-disturbance bound $\overline{\mathbf{h}}$—a looser but still UUB guarantee—showing that the feedforward tightens rather than enables the bound.

Remark 5 (On RL convergence and generalization) The guarantees of Theorems 1–3 do not rest on the convergence of the TD3 algorithm; they require only the deterministic boundedness of the policy output (Lemma 1). The role of training is performance improvement within the guaranteed tube: Fig. 6 evidences empirical convergence, and the 500-scenario study evidences generalization. A probabilistic tightening is available: if the Monte Carlo ensemble verifies $\left\|\widetilde{\mathbf{D}} + \mathbf{g_r}\boldsymbol{\Delta u}\right\| \leq \overline{\rho}$ with empirical frequency $\mathbf{p_0}$ over the perturbation distribution, then the bound holds with probability at least $\mathbf{p_0}$ on that distribution.

4.3.6 Robustness margins

In the frequency domain, the learned compensation channel can be characterized as a sector-bounded nonlinearity, which yields a classical robustness margin statement for the structured loop.

Proposition 1 (Sector-bounded compensation) *Suppose the learned compensation satisfies the sector bound*

$$\|\boldsymbol{\Delta u}(s)\| \leq \rho \|\mathbf{s}\| + \rho_{\mathbf{0}} \tag{63}$$

with constants $(\rho, \rho_{\mathbf{0}})$ estimated from the Monte Carlo ensemble of Section 5.3. Then, for each control loop, if the Nyquist locus of the baseline loop linearized at a frozen morphing configuration does not enter the disk associated with the sector bound, the closed loop remains stable for every policy satisfying (63) by the circle criterion [42]. The classical gain and phase margins of the baseline loop at frozen morphing configurations provide a complementary frequency-domain margin characterization, and the nonlinear-domain counterpart of these margins is the explicit ultimate bound of Theorem 3.

5 Numerical simulation

In this section, the proposed TD3-MRAC (TD3-based Model Reference Adaptive Control) method and the benchmark RMRAC (Robust Model Reference Adaptive Control) approach are evaluated through numerical simulations. The initial conditions, morphing mode, and attitude command profiles are first specified. The nominal-case tracking performance of both controllers is then presented and compared. Finally, a comprehensive multi-source perturbation analysis is conducted to assess and contrast the robustness characteristics of the two control strategies.

5.1 Simulation setup and initialization

The nonlinear dynamics of the HMV derived in Section 3 are employed as the simulation plant. A fixed-wing aircraft model is introduced as the nominal (baseline) model, on which a standard attitude controller is designed. The morphing-induced aerodynamic/structural variations are therefore treated as time-varying uncertainties and coupling effects with respect to the nominal fixed-wing dynamics, which is consistent with common control-oriented modeling assumptions in HMV studies. Initial conditions and command profiles: The closed-loop simulations are initialized with prescribed attitude/angle-rate conditions (or randomized initial conditions within a bounded set), and the attitude commands are selected to include both transient and maneuvering segments (e.g., step-like commands and time-varying commands) to evaluate tracking performance in both transient and steady-state regimes.

Table 1: Initial state of the morphing aircraft

Parameters	Values	Bias
Initial pitch angle	8.0°	$\pm 2^\circ$
Initial yaw angle	0.0°	$\pm 2^\circ$
Initial roll angle	0.0°	$\pm 5^\circ$

Table 2: Parameters of the controller

Parameters		Values
Reward parameters	w_1, w_2	0.5
	w_3, w_4	0.1
	w_5	0.05
	w_6	0.1
	w_7	150
	w_8, w_9	0.05
ESO parameters	$(\beta_1, \beta_2, \beta_3)$	$(100, 300, 350)$

Table 2 lists the reward weights and the ESO parameters used in this study.

Fig. 4 illustrates the prescribed sweep angle morphing command profile for the variable-sweep aircraft over a 70-second simulation horizon. The command signal follows a periodic trapezoid waveform with a period of 24 seconds, in which the sweep angle transitions linearly from 10° to 50° within a 2-second ramp phase, remains at the maximum deflection of 50° for a 10-second dwell phase, and subsequently returns to 10° through a symmetric ramp-down phase, followed by another 10-second hold at the minimum angle. This time-varying morphing schedule introduces significant

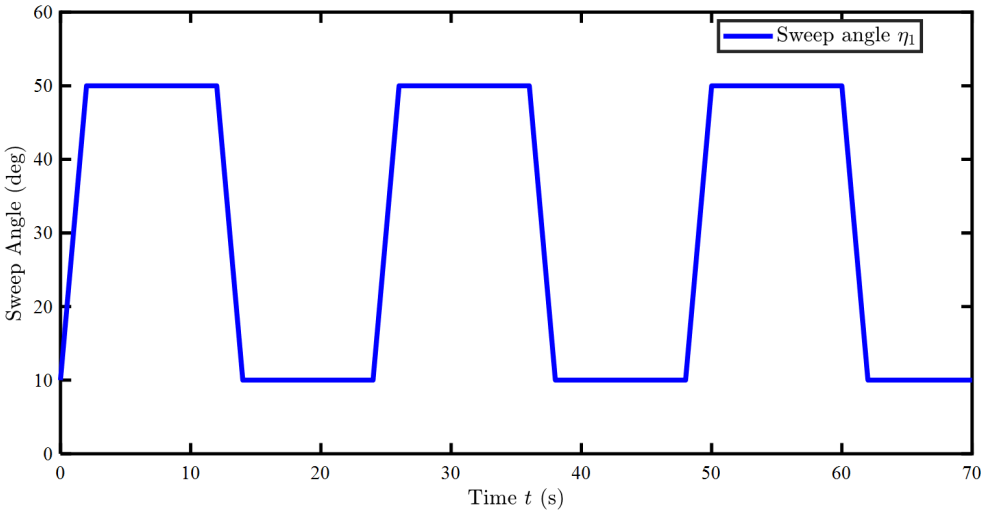


Fig. 4: Sweep angle morphing command

aerodynamic and inertial parameter variations, posing substantial challenges to attitude stability. The command serves as the reference morphing input for evaluating the proposed attitude stabilization controller under continuous structural reconfiguration.

Fig. 5 presents the three-channel attitude command profile prescribed for the variable-sweep aircraft over the 70-second simulation horizon. The roll command undergoes a smooth sigmoidal transition from a shallow bank angle to a sustained large roll, followed by a symmetric return, indicating a coordinated lateral maneuvering phase. Concurrently, the pitch command executes a trapezoidal step from a low nose-down attitude to a markedly elevated pitch angle, maintained over an extended hold phase before recovering, suggesting a pull-up and sustained climb maneuver. The yaw command remains constant throughout, implying no deliberate heading change is required. Collectively, these commands represent a compound banked climb-and-recovery maneuver, which, superimposed on the continuously varying sweep angle morphing schedule, constitutes a highly demanding multi-axis attitude stabilization task for the proposed controller.

5.2 TD3 training process and nominal-case verification

Baseline-plus-compensation architecture. The proposed control input is implemented in the additive form

$$\mathbf{u} = \mathbf{u_b} + \mathbf{u_{ff}} + \Delta\mathbf{u}$$

where $\mathbf{u_b}$ is the plant-side backstepping baseline law (53), $\mathbf{u_{ff}}$ is the ESO feedforward compensation (28), and $\boldsymbol{\Delta}\mathbf{u}$ is the learned compensation term. This structure follows the residual/auxiliary RL philosophy, in which the baseline controller preserves an interpretable stabilization mechanism, the ESO feedforward removes the estimated

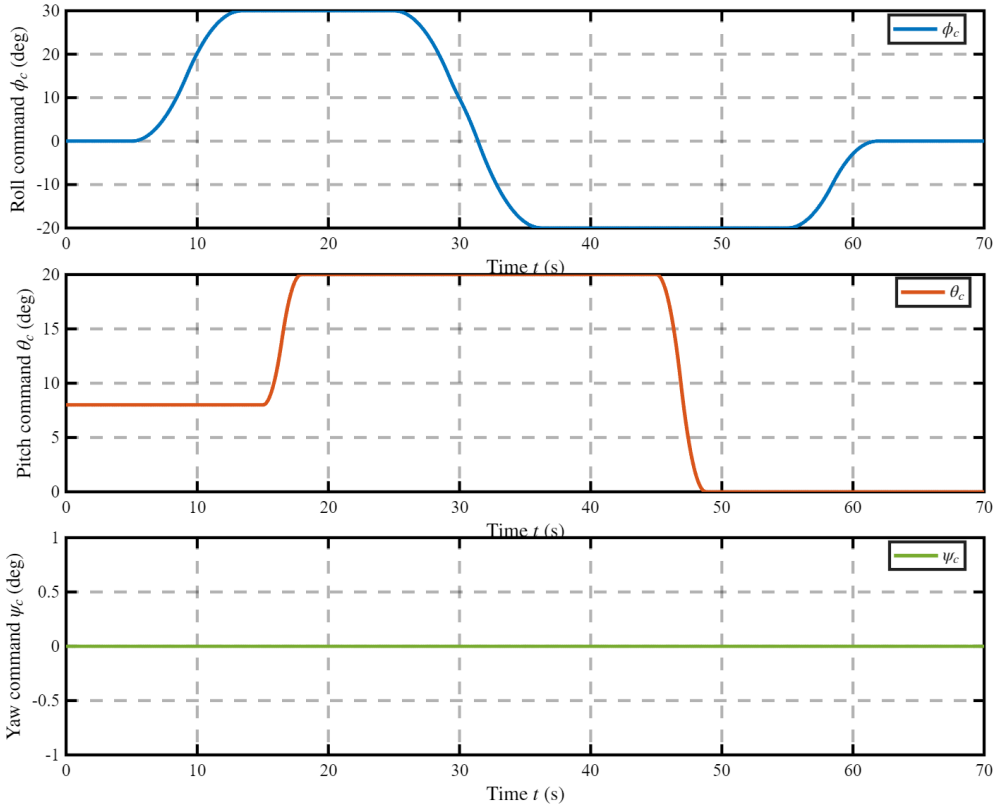


Fig. 5: Three-channel attitude command profile: roll (top), pitch (middle), yaw (bottom)

lumped disturbance, and RL focuses on the remaining unmodeled residual effects; the closed-loop guarantees of this architecture are established in Section 4.3.

The RMRAC benchmark follows the robust MRAC design of Lavretsky and Wise [43] with σ-modification and projection. Its control input is $\mathbf{u} = \mathbf{u_{bl}} + \mathbf{u_{ad}}$, where $\mathbf{u_{bl}} = -\mathbf{K_x}^T\mathbf{x} + \mathbf{k_r r}$ is a fixed baseline, $\mathbf{u_{ad}} = -\widehat{\mathbf{\Theta}}^T\mathbf{\Phi}(\mathbf{x})$ is the adaptive component with regressor $\mathbf{\Phi}(\mathbf{x})$, and the adaptation law is $\dot{\widehat{\mathbf{\Theta}}} = \mathbf{\Gamma} Proj\widehat{\mathbf{\Theta}}, \mathbf{\Phi e P}^T\mathbf{B} - \sigma\Gamma\widehat{\mathbf{\Theta}}$, where $\mathbf{e} = \mathbf{x} - \mathbf{x_m}$ is the model-matching error and $\mathbf{P} \succ 0$ solves the Lyapunov equation $\mathbf{A_m}^T\mathbf{P} + \mathbf{P A_m} = -\mathbf{Q}$. The parameters (baseline gain, adaptation gain, σ, projection bounds, and Q) were tuned on the same scenario set with an equal tuning budget. Both controllers share identical actuator limits, sampling period, initial conditions, morphing schedules, and the identical 500-scenario perturbation ensemble with fixed random seeds.

TD3 formulation and training details. The compensation policy is trained using TD3. TD3 improves training stability for continuous control by employing (i) two critic networks with clipped target estimation to reduce overestimation bias, (ii) delayed actor updates, and (iii) target policy smoothing.

The agent observation (state) includes attitude tracking error terms and nominal-model matching error terms (and their rates), while the action corresponds to the compensation command applied to the baseline controller. The reward is constructed as the weighted sum of the penalties/bonus terms defined in Section 4, encouraging simultaneous command tracking, nominal-dynamics consistency, and smooth/energy-efficient compensation. Key hyperparameters (e.g., replay buffer size, batch size, exploration noise, policy delay, and target smoothing noise) are kept fixed throughout training and are reported for reproducibility.

After training converges (as indicated by the learning curves of episodic return and/or tracking error statistics), the learned compensator is evaluated under nominal or weak-disturbance conditions. The results are presented in terms of (i) attitude command tracking responses, (ii) tracking error time histories, and (iii) baseline input, compensation input, and total control input profiles, with particular attention to the smoothness of Δu and the reduction of control-surface jitter.

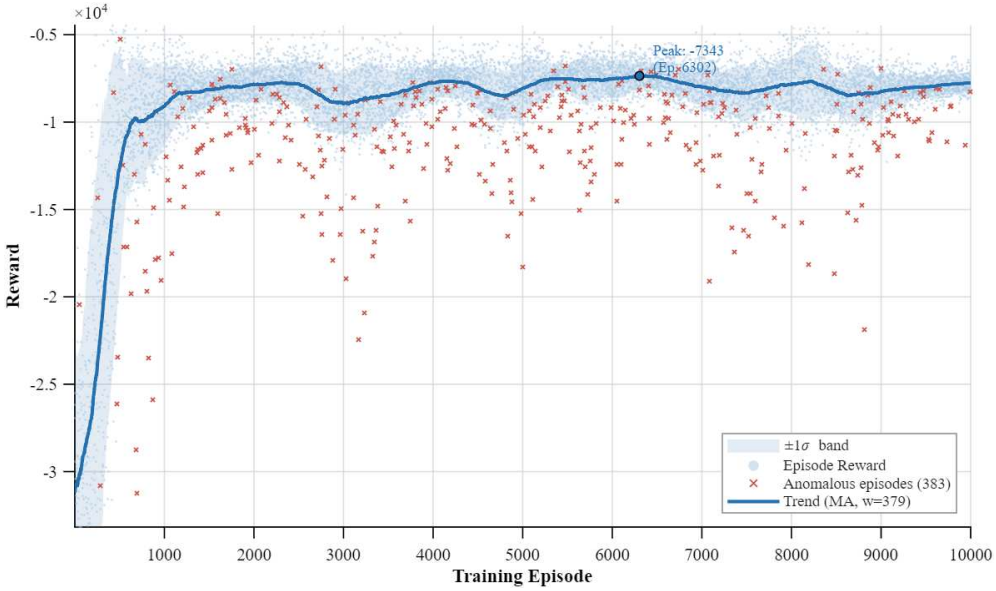


Fig. 6: Training reward curve

Fig. 6 illustrates the episodic reward curve of the TD3 agent over 10,000 training episodes. The agent initially exhibits poor performance, with rewards as low as approximately -3×10^4, reflecting the exploratory phase of early training. A rapid performance improvement is observed within the first 1,000 episodes, after which the reward converges and stabilizes in the range of $-7{,}500$ to $-9{,}000$. The smoothed trend (80-step moving average) confirms a monotonic ascending trajectory that plateaus around episode 1,000, with a peak trend value of $-7{,}343$ recorded at episode 6,302. Occasional anomalous episodes with significantly degraded rewards persist throughout training, indicative of the stochastic nature of the environment and policy exploration.

Overall, the results demonstrate that the TD3 algorithm achieves stable and consistent convergence within a relatively short training horizon.

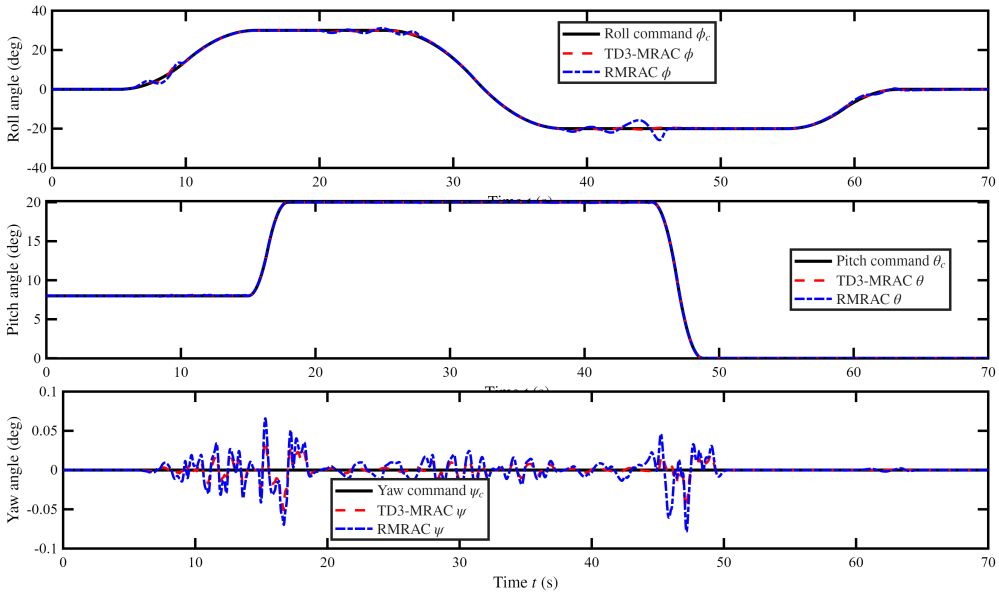


Fig. 7: Attitude tracking response under nominal conditions

As shown in Fig. 7, both controllers achieve satisfactory tracking of the prescribed multi-axis attitude commands. TD3-MRAC demonstrates competitive transient response with reduced overshoot in the pitch channel during the ramp-up phase, attributable to the learned compensation policy that actively counteracts morphing-induced aerodynamic variations. The RMRAC controller exhibits a characteristically smoother control profile with slightly more pronounced transient excursions, reflecting its reliance on adaptive parameter estimation during rapid configuration changes. Fig. 8 displays the control surface deflection histories for both controllers. TD3-MRAC produces control surface commands with higher-frequency content, reflecting the neural network policy output characteristics, whereas RMRAC generates comparatively smoother actuation profiles due to its continuous adaptive law structure.

5.3 Simulation verification under multi-source disturbances

Multi-source uncertainties and morphing scenarios: To assess robustness under realistic morphing conditions, multiple sources of uncertainties/disturbances are introduced, including:

a. Aerodynamic parameter perturbations: key aerodynamic coefficients/derivatives are perturbed (constant-bias or time-varying) to emulate modeling uncertainty;

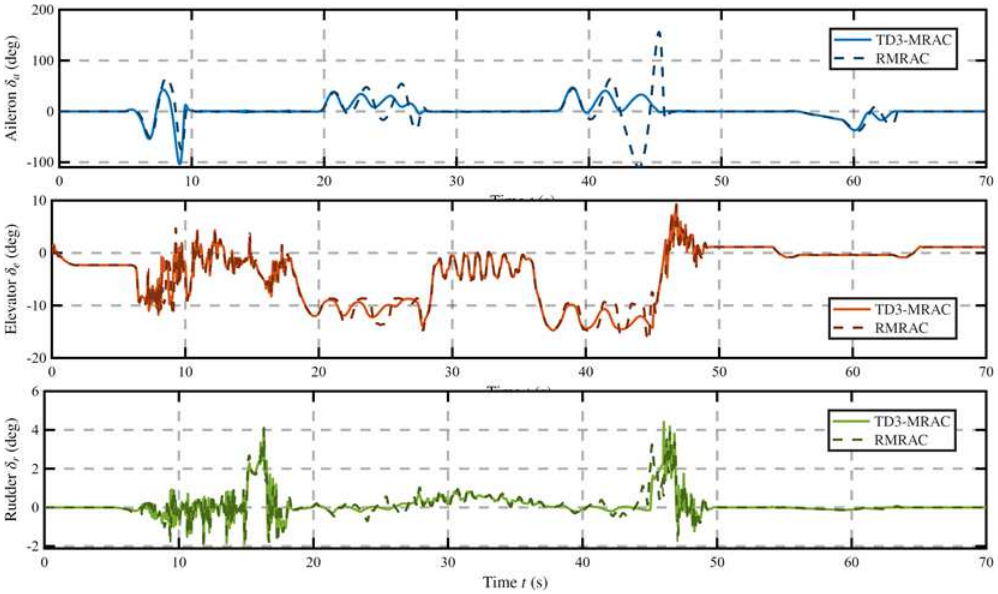


Fig. 8: Control surface deflection profiles under nominal conditions

b. Morphing-mode variations: several morphing schedules are considered (e.g., slow continuous morphing and fast/segmented morphing), producing strong time-varying nonlinearities and coupling effects during the transition phase;

c. External disturbances/noise: equivalent disturbance moments and sensor noise are injected to evaluate disturbance rejection and sensitivity. Such uncertainty-in-morphing-phase testing aligns with typical HMV validation practices.

Table 3: Disturbance term and disturbance magnitude

Parameters	Bias
Initial pitch angle	$\pm 2\,\mathrm{rand}(0,1)^\circ$
Initial yaw angle	$\pm 2\,\mathrm{rand}(0,1)^\circ$
Initial roll angle	$\pm 5\,\mathrm{rand}(0,1)^\circ$
Moment of pitch	$\pm 0.2\,\mathrm{rand}(0,1)$
Moment of roll	$\pm 0.2\,\mathrm{rand}(0,1)$
Moment of yaw	$\pm 0.1\,\mathrm{rand}(0,1)$

To evaluate the robustness of the proposed attitude controller for the morphing aircraft, a structured perturbation framework is established by introducing biases across six parameters, including initial attitude deviations in pitch, yaw, and roll angles, as well as moment disturbances in all three control channels, as summarized in Table 3. The perturbation magnitudes are defined as the product of a channel-specific

scaling factor and rand(0,1), where rand(0,1) denotes a scalar randomly sampled from a uniform distribution over the interval [0, 1], ensuring that the applied bias varies stochastically within the prescribed bounds.

To enable combinatorial scenario generation, each perturbation term is assigned a discrete indicator: 0 denotes no perturbation, +1 denotes a positive bias, and −1 denotes a negative bias. By permuting these indicators across all six parameters, a comprehensive set of disturbance combinations is constructed, with each combination defining a distinct operational scenario. The controller is subsequently evaluated across all scenarios, ensuring that the robustness assessment covers a broad spectrum of coupled disturbance conditions rather than isolated single-parameter deviations, thereby providing a rigorous validation of the controller's tolerance to parametric uncertainty.

5.3.1 Robustness analysis of the controller under multiple source disturbances

To systematically evaluate the robustness of the proposed TD3-based attitude controller under parametric uncertainty, a Monte Carlo-style batch simulation framework is established. A total of 500 perturbation scenarios are constructed by assigning discrete bias indicators of −1, 0, and +1 to six independent disturbance parameters, comprising initial attitude deviations in roll (φ), pitch (θ), and yaw (ψ), as well as additive biases on the aerodynamic moment coefficients C_m, C_l, and C_n. Each indicator combination defines a unique operational scenario, and the controller is evaluated under all 500 conditions using the same trained TD3 policy without any retraining or fine-tuning.

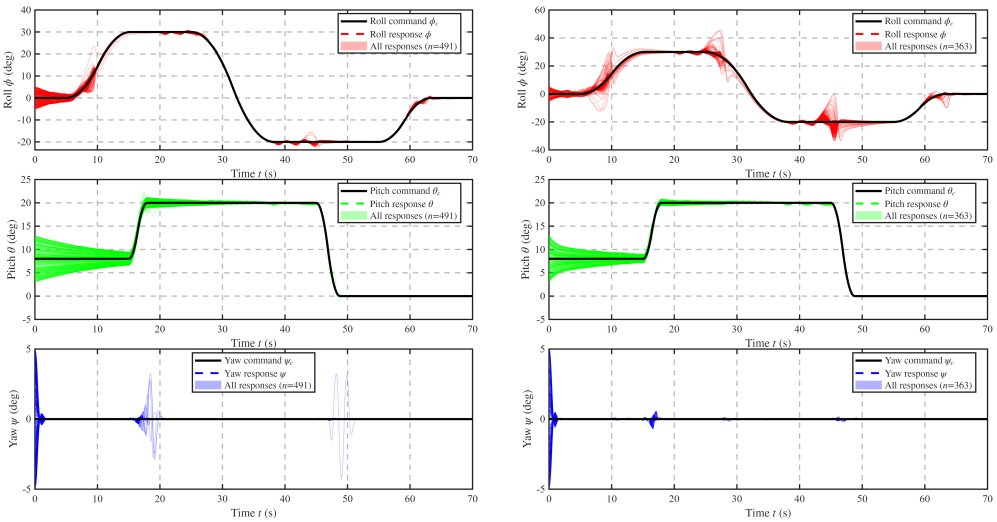
For each scenario, the closed-loop attitude tracking performance is quantified using the Integral Absolute Error (IAE)

$$IAE_k = \int_0^{T_f} |e_k(t)|\, dt, \qquad k \in \{\phi, \theta, \psi\}$$

where $e_k(t)$ denotes the tracking error of the $k-th$ attitude channel and $T_f = 70$ s is the simulation duration. The IAE is computed independently for each channel via trapezoidal numerical integration. The retained scenarios are subsequently visualized as overlaid time-history curves, and the IAE distributions across all valid scenarios are presented as channel-wise histograms to characterize the statistical robustness of the controller.

The robustness of the trained TD3 attitude controller is evaluated across 500 perturbation scenarios, with results summarized in the IAE histograms and parameter sensitivity radar chart. Fig. 9 presents the attitude tracking responses of TD3-MRAC and RMRAC across all valid perturbation cases. In each subplot, the black solid line denotes the attitude command, the colored dashed line represents the median response across all valid cases, and the translucent envelope captures the ensemble of all individual case responses. The spread of the response envelope reflects controller sensitivity to parameter perturbations: a narrower envelope indicates superior robustness to the perturbation set.

As shown in Fig. 9, the two controllers exhibit distinct channel-dependent behaviors. In the roll channel, TD3-MRAC achieves a markedly tighter response envelope



(a) (b)

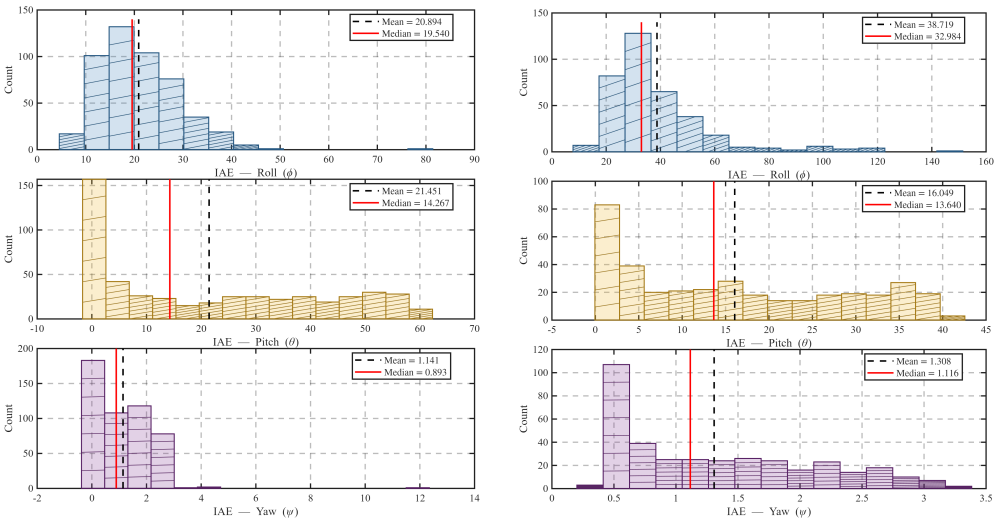
Fig. 9: Attitude tracking responses of TD3-MRAC and RMRAC across 500 perturbation cases. (a) TD3-MRAC. (b) RMRAC.

and tracks the median command more faithfully than RMRAC, consistent with the 46% mean IAE improvement reported in Table 4. In the pitch channel, RMRAC maintains a somewhat tighter envelope, reflecting its observer-based adaptive law's effectiveness against structured pitch-axis perturbations. In the yaw channel, both controllers produce comparably narrow envelopes with TD3-MRAC showing a slight edge in median tracking fidelity. Under both controllers, the roll channel exhibits the largest dispersion among the three axes, suggesting that roll dynamics are inherently more sensitive to the selected perturbation set than pitch and yaw channels.

5.3.2 Integral absolute error distribution

Fig. 10 shows the IAE histograms for both controllers, aggregated across all valid perturbation cases. The vertical dashed and solid lines mark the mean and median IAE, respectively. A distribution concentrated at lower IAE values with a shorter right tail indicates not only better nominal tracking accuracy but also greater resilience to worst-case perturbations.

The IAE distributions in Fig. 10 reveal a nuanced picture. In the roll channel, TD3-MRAC achieves substantially lower mean and median IAE (20.9 deg·s and 19.5 deg·s, respectively) compared with RMRAC (38.7 deg·s and 33.0 deg·s), representing a 46% improvement. In the yaw channel, TD3-MRAC also outperforms RMRAC, with a 13% lower mean IAE and a 20% lower median IAE. In the pitch channel, the situation reverses: RMRAC achieves a 34% lower mean IAE, although the median gap narrows to only 5%, indicating that TD3-MRAC's pitch performance is competitive in typical cases and degraded primarily in a subset of outlier scenarios. The RMRAC



(a) (b)

Fig. 10: IAE distribution across valid perturbation cases. (a) TD3-MRAC. (b) RMRAC.

distributions are generally more compact across all channels, reflecting the consistency benefits of its explicit disturbance-rejection structure. For TD3-MRAC, the wider IAE spread in pitch indicates an opportunity for targeted improvement through domain-randomized training or pitch-specific data augmentation.

5.3.3 Parameter sensitivity analysis

The radar charts in Fig. 11 reveal channel-specific sensitivity patterns. Under TD3-MRAC, the pitch channel is most sensitive to aerodynamic moment coefficient biases, which accounts for the performance gap relative to RMRAC observed in this channel. The roll and yaw channels show comparatively balanced sensitivity profiles, consistent with the strong IAE improvements in these axes. RMRAC exhibits a more uniformly distributed sensitivity footprint across all six perturbation sources, a direct benefit of its robust adaptive architecture that actively compensates for parametric uncertainties in real time. For both controllers, initial attitude biases contribute relatively less to IAE degradation than moment coefficient biases, indicating that the closed-loop dynamics are more tolerant of initial condition errors than of persistent aerodynamic discrepancies.

5.4 Quantitative comparison and discussion

Table 4 summarizes the key robustness metrics for both controllers, consolidating the findings from the perturbation analysis. The comparison encompasses the mean and median IAE per channel and the percentage improvement of TD3-MRAC relative to RMRAC.

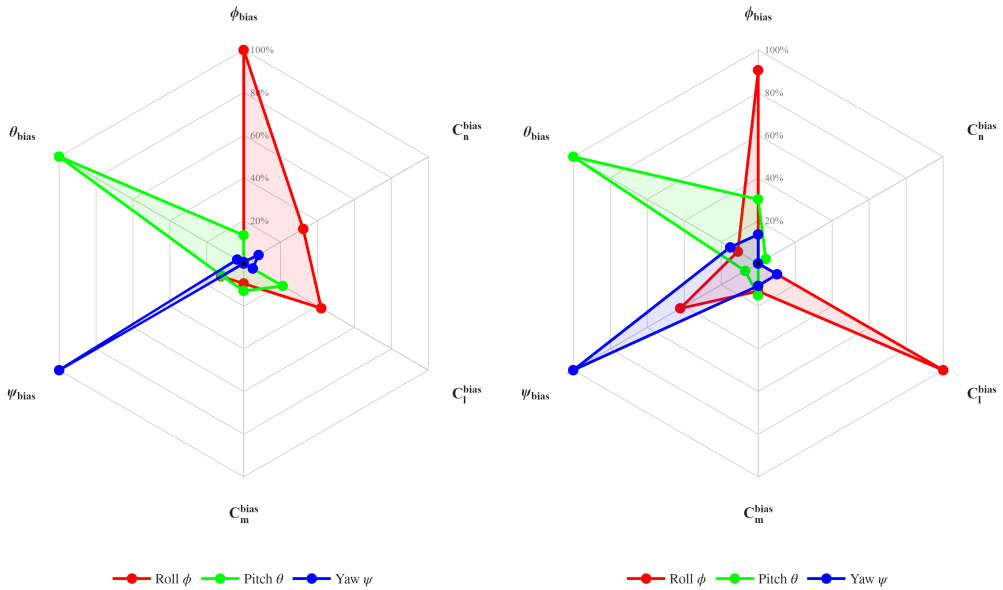


Fig. 11: Normalized parameter sensitivity radar charts. (a) TD3-MRAC. (b) RMRAC.

Table 4: Summary of robustness metrics for TD3-MRAC and RMRAC

Metric	Channel	TD3-MRAC	RMRAC	Improv. (%)
Mean IAE (deg·s)	Roll	20.894	38.719	+46.0
Mean IAE (deg·s)	Pitch	21.451	16.049	−33.7
Mean IAE (deg·s)	Yaw	1.141	1.308	+12.8
Median IAE (deg·s)	Roll	19.540	32.984	+40.8
Median IAE (deg·s)	Pitch	14.267	13.640	−4.6
Median IAE (deg·s)	Yaw	0.893	1.116	+20.0

The comparative analysis presented in this section reveals a more favorable picture for TD3-MRAC than a cursory inspection of response envelopes alone might suggest. Across the six quantitative metrics in Table 4, TD3-MRAC wins on four—roll and yaw, both mean and median—with improvements reaching 46% in roll mean IAE and 20% in yaw median IAE. RMRAC holds an advantage only in the pitch channel (34% mean, 5% median). This channel-asymmetric outcome is consistent with the two controllers' design philosophies: RMRAC's observer-based adaptive law is structurally well-suited to rejecting the specific pitch-axis perturbations included in the test suite, while TD3-MRAC's model-free compensation generalizes more broadly across the full perturbation space without relying on structured uncertainty assumptions.

A distinguishing strength of the TD3-MRAC approach, borne out by these results, is that it eliminates the need for explicit aerodynamic uncertainty characterization. Conventional model-based robust or adaptive methods such as RMRAC require the designer to specify the structure, bounds, and channels of parametric uncertainty—a nontrivial task for morphing vehicles whose aerodynamic coefficients vary continuously with the morphing state. The TD3-MRAC agent, by contrast, acquires its compensation strategy directly from simulated experience, implicitly learning the morphing-dependent uncertainty through the value function and policy networks. The substantial roll and yaw improvements demonstrate that this data-driven paradigm not only reduces the engineering burden of uncertainty characterization but also discovers compensation strategies that can exceed the performance of hand-designed adaptive laws in axes where the uncertainty structure is complex or poorly characterized.

Furthermore, the offline-training, online-deployment structure of the TD3-MRAC architecture is well-suited to practical flight control applications. At runtime, policy evaluation requires only forward passes through the neural networks, yielding deterministic, real-time control actions without the online parameter estimation or iterative optimization inherent to many adaptive control schemes. The sustained performance across the perturbation ensemble confirms that the trained policy has not overfitted to the nominal dynamics and generalizes effectively to off-nominal conditions—a critical requirement for certification-conscious RL applications in aerospace systems. The remaining pitch-channel gap, which narrows substantially in median (5%) versus mean (34%), suggests that targeted data augmentation or domain randomization during training could further close this gap. Taken together, the strong roll and yaw results, combined with the elimination of explicit uncertainty modeling, position TD3-MRAC as a practical and compelling approach for robust attitude control of morphing aircraft.

6 Conclusion

6.1 Summary

This paper has presented TD3-MRAC, a learning-enhanced model reference adaptive control framework for hypersonic morphing vehicle attitude control. The proposed architecture integrates a fixed-wing reference model, a baseline backstepping controller, a linear extended state observer with explicit feedforward compensation, and a bounded TD3-learned compensation policy. A complete layered stability analysis is provided: Theorem 1 proves global exponential stability of the reference closed loop; Theorem 2 proves the input-to-state stability of the LESO with a steady-state estimation bound of $O(\overline{\mathbf{H}}/\omega_o)$; Theorem 3 and Corollary 1 prove uniform ultimate boundedness of the composite closed loop with the explicit ultimate bound $\bar{b}/c_0$ expressed in terms of the design parameters; and Proposition 1 characterizes the robustness margin retained by the baseline loop in the presence of the sector-bounded learned compensation. These guarantees rely only on the deterministic boundedness of the learned action and are decoupled from RL convergence. Numerical simulations under a 500-scenario multi-source perturbation ensemble demonstrate that

TD3-MRAC substantially outperforms a conventional robust adaptive controller in the roll and yaw channels—with IAE improvements of up to 46%—while remaining competitive in pitch. The central contribution is twofold: the elimination of explicit aerodynamic uncertainty characterization from the control design workflow, and a stability-certified structured-learning architecture for morphing vehicles whose aerodynamic coefficients vary continuously in operation.

6.2 Future work

Several directions warrant further investigation. First, the fixed single-model architecture can be extended to a gain-scheduled set of reference models with hysteresis switching, whose stability follows from average-dwell-time arguments [34]; this would widen the certified morphing envelope. Second, the ultimate bound can be tightened by training the policy against adversarial perturbations and by certifying the sector constants of Proposition 1 via Lipschitz estimation of the actor network. Third, online adaptation mechanisms—such as meta-learning or safe continual learning—could refine the compensation behavior in response to previously unseen morphing configurations, e.g., through domain-randomized training [44], while the deterministic safety monitor preserves the guarantees outside the certified envelope. Finally, the framework should be validated through hardware-in-the-loop experiments and, ultimately, flight tests on a representative morphing platform.

Declarations

Funding. This work was supported by the National Natural Science Foundation of China (Grant Nos. 92471205, 92471301, 92371201, and 521926233).

Competing interests. The authors declare that they have no competing interests.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The data supporting the findings of this study will be made publicly available on GitHub upon publication of this article.

Materials availability. Not applicable.

Code availability. The code used in this study is not publicly available.

Author contributions. H. Chen conceived the idea, designed the framework, acquired the funding, and wrote the original draft. F. Meng developed the TD3-MRAC algorithm and performed the simulations. Y. Huang contributed to the theoretical analysis and the numerical validation. G. Chen supervised the work and reviewed and edited the manuscript. All authors discussed the results and approved the final version of the manuscript.

Supplementary information. No supplementary material is associated with this article.

References

[1] Parancheerivilakkathil, M.S., *et al.*: A review of control strategies used for morphing aircraft applications. Chin. J. Aeronaut. **37**(4), 436–463 (2024) https://doi.org/10.1016/j.cja.2023.12.035

[2] Barbarino, S., Bilgen, O., Ajaj, R.M., Friswell, M.I., Inman, D.J.: A review of morphing aircraft. J. Intell. Mater. Syst. Struct. **22**(9), 823–877 (2011) https://doi.org/10.1177/1045389X11414084

[3] Kim, Y., Kim, S., Suk, J., Shin, H.-S.: Robust controller design for UAM using incremental nonlinear dynamic inversion with double-loop structure. IEEE Trans. Aerosp. Electron. Syst. **61**(1), 76–92 (2025) https://doi.org/10.1109/TAES.2024.3443024

[4] Wu, K., Zhang, P., Wu, H.: A new control design for a morphing UAV based on disturbance observer and command filtered backstepping techniques. Sci. China Technol. Sci. **62**(10), 1845–1853 (2019) https://doi.org/10.1007/s11431-018-9377-8

[5] Madonski, R., Herbst, G., Stankovic, M.: ADRC in output and error form: Connection, equivalence, performance. Control Theory Technol. **21**(1), 56–71 (2023) https://doi.org/10.1007/s11768-023-00129-y

[6] Chen, W., Lu, J., Wang, X., Zhang, W.: Design of a controller for morphing aircraft based on backstepping/RHO. J. Beijing Univ. Aeronaut. Astronaut. **40**(8), 1060–1065 (2014) https://doi.org/10.13700/j.bh.1001-5965.2013.0510

[7] Bechlioulis, C.P., Rovithakis, G.A.: Robust adaptive control of feedback linearizable MIMO nonlinear systems with prescribed performance. IEEE Trans. Autom. Control **53**(9), 2090–2099 (2008) https://doi.org/10.1109/TAC.2008.929402

[8] Sönmez, S., et al.: Reinforcement Learning Based Prediction of PID Controller Gains for Quadrotor UAVs. arXiv:2502.04552 (2025). https://doi.org/10.48550/arXiv.2502.04552

[9] Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. In: International Conference on Learning Representations (ICLR), San Juan, Puerto Rico (2016)

[10] Zhang, B., Zhang, W., Mou, J., Yang, R., Zhang, Y.: Fuzzy PID controller for UAV based on reinforcement learning. In: Fu, W., Gu, M., Niu, Y. (eds.) Proceedings of 2022 International Conference on Autonomous Unmanned Systems (ICAUS 2022), pp. 1724–1732. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-0479-2_160

[11] Hezaveh, M.H., Ramezanlou, M.T., Schwartz, H., Lambadaris, I.: Enhancing UAV altitude control: PID tuning with reinforcement learning and genetic algorithm.

In: 2025 IEEE International Systems Conference (SysCon), pp. 1–6 (2025). https://doi.org/10.1109/SysCon64521.2025.11014809

[12] Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A.: Deep reinforcement learning: A brief survey. IEEE Signal Process. Mag. **34**(6), 26–38 (2017) https://doi.org/10.1109/MSP.2017.2743240

[13] Lin, J., Wang, L., Gao, F., Shen, S., Zhang, F.: Flying Through a Narrow Gap Using Neural Network: An End-to-End Planning and Control Approach. arXiv:1903.09088 (2019). https://doi.org/10.48550/arXiv.1903.09088

[14] Zhao, J., et al.: End-to-end deep reinforcement learning for image-based UAV autonomous control. Appl. Sci. **11**(18) (2021) https://doi.org/10.3390/app11188419

[15] Hao, G., Li, Y., Li, Y., Jiang, L., Zeng, Z.: Lyapunov-based safe reinforcement learning for microgrid energy management. IEEE Trans. Neural Netw. Learn. Syst. **36**(6), 9985–9999 (2025) https://doi.org/10.1109/TNNLS.2024.3496932

[16] Johannink, T., *et al.*: Residual reinforcement learning for robot control. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 6023–6029 (2019). https://doi.org/10.1109/ICRA.2019.8794127

[17] Dhiman, G., Tiumentsev, Y.V., Tskhay, R.A.: Combined use of dynamic inversion and reinforcement learning for optimal adaptive control of supersonic transport airplane motion. J. Comput. Syst. Sci. Int. **64**(1), 136–152 (2025) https://doi.org/10.1134/S106423072570011X

[18] Ran, M., Li, J., Xie, L.: Reinforcement-learning-based disturbance rejection control for uncertain nonlinear systems. IEEE Trans. Cybern. **52**(9), 9621–9633 (2022) https://doi.org/10.1109/TCYB.2021.3060736

[19] Alakuijala, M., Dulac-Arnold, G., Mairal, J., Ponce, J., Schmid, C.: Residual Reinforcement Learning from Demonstrations. arXiv:2106.08050 (2021). https://doi.org/10.48550/arXiv.2106.08050

[20] Luo, Y., Ma, T.: Learning Barrier Certificates: Towards Safe Reinforcement Learning with Zero Training-Time Violations. arXiv:2108.01846 (2022). https://doi.org/10.48550/arXiv.2108.01846

[21] Hou, Z., Liu, W., Knoll, A.: Safe reinforcement learning for autonomous driving by using disturbance-observer-based control barrier functions. IEEE Trans. Intell. Veh. **10**(6), 3782–3791 (2025) https://doi.org/10.1109/TIV.2024.3463468

[22] Kalaria, D., Lin, Q., Dolan, J.M.: Disturbance Observer-Based Control Barrier Functions with Residual Model Learning for Safe Reinforcement Learning. arXiv:2410.06570 (2024). https://doi.org/10.48550/arXiv.2410.06570

[23] Chen, J., Wang, J., Wang, W.: Model reference adaptive control for a class of aircraft with actuator saturation. In: 2018 37th Chinese Control Conference (CCC), pp. 2705–2710 (2018). https://doi.org/10.23919/ChiCC.2018.8484160

[24] Chen, M., Shen, Y., Skelton, R.E.: Model-based and Markov data-based linearized tensegrity dynamics and analysis of morphing airfoils. In: AIAA SCITECH 2024 Forum. American Institute of Aeronautics and Astronautics, Reston, VA, USA (2024). https://doi.org/10.2514/6.2024-2347

[25] Shekhar, A., Sharma, A.: Review of model reference adaptive control. In: 2018 International Conference on Information, Communication, Engineering and Technology (ICICET), pp. 1–5 (2018). https://doi.org/10.1109/ICICET.2018.8533713

[26] Gong, S., Xu, Z., Cheng, L., Huang, X.: Self-organizing model reference adaptive control for aircraft with enhanced persistent excitation. Aerosp. Sci. Technol. **145**, 108875 (2024) https://doi.org/10.1016/j.ast.2024.108875

[27] Tao, G., Tang, X., Chen, S., Fei, J., Joshi, S.M.: Adaptive failure compensation of two-state aircraft morphing actuators. IEEE Trans. Control Syst. Technol. **14**(1), 157–164 (2006) https://doi.org/10.1109/TCST.2005.859639

[28] Kiumarsi, B., Vamvoudakis, K.G., Modares, H., Wunsch, D.C., Lewis, F.L.: Optimal and autonomous control using reinforcement learning: A survey. IEEE Trans. Neural Netw. Learn. Syst. **29**(6), 2042–2062 (2018) https://doi.org/10.1109/TNNLS.2017.2773458

[29] Joshi, G., Virdi, J., Chowdhary, G.: Asynchronous Deep Model Reference Adaptive Control. arXiv:2011.02920 (2020). https://doi.org/10.48550/arXiv.2011.02920

[30] Cheng, Y., Zhao, P., Wang, F., Block, D.J., Hovakimyan, N.: Improving the robustness of reinforcement learning policies with $\mathcal{L}_1$ adaptive control. IEEE Robot. Autom. Lett. **7**(3), 6574–6581 (2022) https://doi.org/10.1109/LRA.2022.3169309

[31] Rohrs, C.E., Valavani, L., Athans, M., Stein, G.: Robustness of continuous-time adaptive control algorithms in the presence of unmodeled dynamics. IEEE Trans. Autom. Control **30**(9), 881–889 (1985) https://doi.org/10.1109/TAC.1985.1104070

[32] Ioannou, P.A., Sun, J.: Robust Adaptive Control. Prentice Hall, Upper Saddle River, NJ, USA (1996)

[33] U.S. Department of Defense: Flying Qualities of Piloted Aircraft. MIL-STD-1797B (2006)

[34] Liberzon, D.: Switching in Systems and Control. Birkhäuser, Boston, MA, USA (2003)

[35] Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.: Deterministic policy gradient algorithms. In: Proceedings of the 31st International Conference on Machine Learning (ICML), Beijing, China, pp. 387–395 (2014)

[36] Fujimoto, S., van Hoof, H., Meger, D.: Addressing function approximation error in actor-critic methods. In: Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, pp. 1587–1596 (2018)

[37] Cheng, C., HuiXin, Y.: Heading rate controller for unmanned helicopters based on modified ADRC. Sci. China Technol. Sci. **66**(5), 1255–1262 (2023)

[38] Kumar, N., Kishore, G., Saderla, S., Kim, Y.: An adaptive law based online system identification of cropped delta unmanned aerial vehicles from flight tests. Aerosp. Sci. Technol. **141**, 108526 (2023) https://doi.org/10.1016/j.ast.2023.108526

[39] Han, J.: From PID to active disturbance rejection control. IEEE Trans. Ind. Electron. **56**(3), 900–906 (2009) https://doi.org/10.1109/TIE.2008.2011621

[40] Gao, Z.: Scaling and bandwidth-parameterization based controller tuning. In: Proceedings of the American Control Conference (ACC), Denver, CO, USA, pp. 4989–4996 (2003). https://doi.org/10.1109/ACC.2003.1242516

[41] Khalil, H.K.: Nonlinear Systems, 3rd edn. Prentice Hall, Upper Saddle River, NJ, USA (2002)

[42] Zames, G.: On the input-output stability of time-varying nonlinear feedback systems—part i: Conditions derived using concepts of loop gain, conicity, and positivity. IEEE Trans. Autom. Control **11**(2), 228–238 (1966) https://doi.org/10.1109/TAC.1966.1098316

[43] Lavretsky, E., Wise, K.A.: Robust and Adaptive Control: With Aerospace Applications, 2nd edn. Springer, London, U.K. (2024). https://doi.org/10.1007/978-1-4471-4396-3

[44] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, pp. 23–30 (2017). https://doi.org/10.1109/IROS.2017.8202133