actuators

IMPACT FACTOR 2.523

CITESCORE 2.7

Article

# Robust Attitude Control of an Agile Aircraft Using Improved Q-Learning

Mohsen Zahmatkesh, Seyyed Ali Emami, Afshin Banazadeh and Paolo Castaldi

Special Issue

Fault Diagnosis, Fault-Tolerant Control and Their Applications to Aerospace and Mechatronic Systems

Edited by

Dr. Paolo Castaldi and Dr. Silvio Simani

MDPI

https://doi.org/10.3390/act11120374

actuators

MDPI

*Article*

# Robust Attitude Control of an Agile Aircraft Using Improved Q-Learning

**Mohsen Zahmatkesh [1], Seyyed Ali Emami [1], Afshin Banazadeh [1] and Paolo Castaldi [2,*]**

[1] Department of Aerospace Engineering, Sharif University of Technology, Tehran 1458889694, Iran
[2] Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi", University of Bologna, Via Dell'Universit'a 50, 40126 Cesena, Italy
* Correspondence: paolo.castaldi@unibo.it; Tel.: +39-0547-339242

**Abstract:** Attitude control of a novel regional truss-braced wing (TBW) aircraft with low stability characteristics is addressed in this paper using Reinforcement Learning (RL). In recent years, RL has been increasingly employed in challenging applications, particularly, autonomous flight control. However, a significant predicament confronting discrete RL algorithms is the dimension limitation of the state-action table and difficulties in defining the elements of the RL environment. To address these issues, in this paper, a detailed mathematical model of the mentioned aircraft is first developed to shape an RL environment. Subsequently, Q-learning, the most prevalent discrete RL algorithm, will be implemented in both the Markov Decision Process (MDP) and Partially Observable Markov Decision Process (POMDP) frameworks to control the longitudinal mode of the proposed aircraft. In order to eliminate residual fluctuations that are a consequence of discrete action selection, and simultaneously track variable pitch angles, a Fuzzy Action Assignment (FAA) method is proposed to generate continuous control commands using the trained optimal Q-table. Accordingly, it will be proved that by defining a comprehensive reward function based on dynamic behavior considerations, along with observing all crucial states (equivalent to satisfying the Markov Property), the air vehicle would be capable of tracking the desired attitude in the presence of different uncertain dynamics including measurement noises, atmospheric disturbances, actuator faults, and model uncertainties where the performance of the introduced control system surpasses a well-tuned Proportional–Integral–Derivative (PID) controller.

**Keywords:** reinforcement learning; q-learning; fuzzy q-learning; attitude control; truss-braced wing; flight control



**Citation:** Zahmatkesh, M.; Emami, S.A.; Banazadeh, A.; Castaldi, P. Robust Attitude Control of an Agile Aircraft Using Improved Q-Learning. *Actuators* **2022**, *11*, 374. https://doi.org/10.3390/act11120374

Academic Editor: Ronald M. Barrett

Received: 1 November 2022
Accepted: 6 December 2022
Published: 12 December 2022



## 1. Introduction

The aviation industry is expeditiously growing due to world demands such as reducing fuel burn, emissions, and cost, as well as providing faster and safer flights. This motivates the advent of new airplanes with novel configurations. In addition, the scope clause agreement limits the number of seats in each aircraft and flight outsourcing to protect the union pilot jobs. This factor leads to an increase in the production of Modern Regional Jet (MRJ) airplanes. In this regard, the importance of a safe flight becomes more vital considering more crowded airspace and new aircraft configurations having the ability to fly faster. Truss-braced wing aircraft is one of the re-raised high-performance configurations, which has attracted significant attention from both academia [1] and industry [2] due to its fuel burn efficiency. As a result, there would be a growing need for reliable modeling and simulations, analyzing the flight handling quality, and stability analysis for such configurations [3,4], while very few studies have addressed the flight control design for this aircraft category.

In the last decades, various classic methods for aircraft attitude control have been developed to enhance control performance. However, the most significant deficiency of

these approaches is the insufficient capacity to deal with unexpected flight conditions, while typically requiring a detailed dynamic model of the system.

Recently, the application of Reinforcement Learning (RL) has been extended to real problems, particularly, flight control design [5]. Generally, there are two main frameworks to incorporate RL in the control design process, i.e., the high-level and low-level control systems. In [6], a Soft Actor-Critic (SAC) algorithm was implemented in a path planning problem for a long-endurance solar-powered UAV with energy-consuming considerations. Another work [7] concentrated on the low-level control of a Skywalker X8 using SAC and comparing it with a PID controller. In [8], an ANN-based Q-learning horizontal trajectory tracking controller was developed based on an MDP model of an airship with fine stability characteristics. Apart from the previous method, Proximal Policy Optimization (PPO) was utilized in [9] for orientation control of a highly dynamically coupled fixed-wing aircraft in the stall condition. The PPO was successfully converged after 100,000 episodes. The application of PPO is not limited to fixed-wing aircraft. In [10], An improved proximal policy optimization algorithm is introduced to train a quadrotor for low-level control loops in various tasks such as take-off, precise flight, and hover.

Additionally, several papers have focused on particular maneuvers (such as the landing phase control) both in low-level and high-level schemes. For instance, in [11], Deep Q-learning (DQL) is used to guide an aircraft to land in the desired field. In [12], a Deep Deterministic Policy Gradient (DDPG) was implemented for a UAV to control either path-tracking for landing glide slope or attitude control for the landing flare section. Similarly, a DDPG method in [13] is used to control the outer loop of a landing procedure in the presence of wind disturbance. Additionally, the DDPG method is utilized in [14] to track desired values of skid steering vehicles under fault situations. This method could perform dual-channel control over the yaw rate and longitudinal speed. The works which have been referred to so far accompanied ANNs to be able to converge. However, to our best knowledge, there is research in attitude control using discrete RL without aiming at ANNs. In [15], a Q-learning algorithm was implemented to control longitudinal and lateral orientations in a general aviation aircraft (Cessna 172). This airplane profits from suitable stability characteristics, and the desired angles are zero. Furthermore, there are some Fuzzy adaptations of the work in [16], such as [17] where the Q-functions and action selection strategy are inferred from Fuzzy rules. Additionally, ref. [18] proposed a dynamic Fuzzy Q-learning (FQL) for online and continuous tasks in mobile robots.

Motivated by the above discussions, the main contributions of the current study can be summarized as follows:

(1) Alongside the response to global aviation society demands, a TBW aircraft (Chaka 50) (Figure 1) with poor stability characteristics has been chosen for the attitude control problem;
(2) It will be demonstrated that the proposed reward function is able to provide a robust control system even in low-stability and high-degree-of-freedom plants;
(3) The performance of the Q-learning controller will be evaluated in both MDP and POMDP problem definitions using different control criteria. Moreover, by proposing an effective Fuzzy Action Assignment (FAA) algorithm, continuous elevator commands could be generated using the discrete optimal Q-table (policy). Such a control approach illustrates well that the tabular Q-learning method can be a strong learning approach resulting in an effective controller for complex systems under uncertain conditions;
(4) In order to prove the reliability and robustness of the proposed control method, it is examined under different flight conditions consisting of sensor measurement noises, atmospheric disturbances, actuator faults, and model uncertainties, while the training process is only performed for ideal flight conditions.

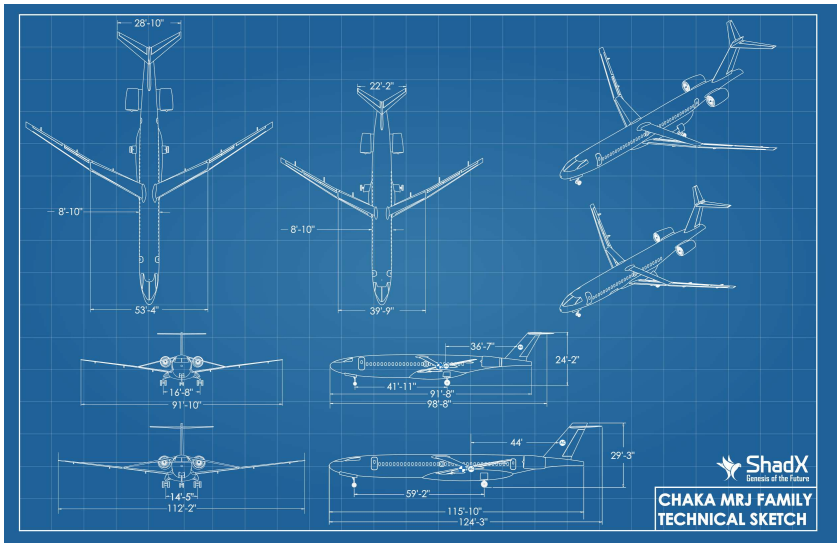


**Figure 1.** Chaka Modern Regional Jet (MRJ) Family [4].

## 2. Modeling and Simulation

In this section, the nonlinear equations of motion are derived for the airplane based on [19,20] in order to provide a nonlinear RL environment. There are some open-source environments based on Gym and Flight Gear such as GymFG [21]. Yet the specifics of the novel configuration enforce this research to simulate from scratch. By considering $F_B = [O_B, x_B, y_B, z_B]$, $F_S = [O_S, x_S, y_S, z_S]$, and $F_E = [O_E, x_E, y_E, z_E]$ as body, stability, and inertial frames as noted in Figure 2, the translational and rotational equations in the body frame are as follows:

$$m \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix}^B + m \begin{bmatrix} 0 & -r & q \\ r & 0 & -p \\ -q & p & 0 \end{bmatrix}^B \begin{bmatrix} u \\ v \\ w \end{bmatrix}^B = \begin{bmatrix} F_{A_x} \\ F_{A_y} \\ F_{A_z} \end{bmatrix}^B + \begin{bmatrix} F_{T_x} \\ F_{T_y} \\ F_{T_z} \end{bmatrix}^B + m \begin{bmatrix} g_x \\ g_y \\ g_z \end{bmatrix}^B, \tag{1}$$

$$\begin{bmatrix} I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_z \end{bmatrix}^B \begin{bmatrix} \dot{p} \\ \dot{q} \\ \dot{r} \end{bmatrix}^B + \begin{bmatrix} 0 & -r & q \\ r & 0 & -p \\ -q & p & 0 \end{bmatrix}^B \begin{bmatrix} I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_z \end{bmatrix}^B \begin{bmatrix} p \\ q \\ r \end{bmatrix}^B = \begin{bmatrix} L_A \\ M_A \\ N_A \end{bmatrix}^B + \begin{bmatrix} L_T \\ M_T \\ N_T \end{bmatrix}^B, \tag{2}$$

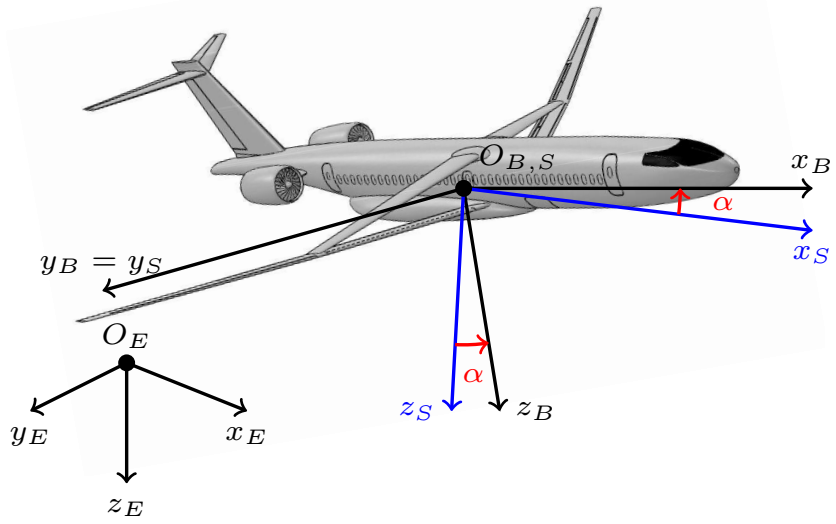


**Figure 2.** Chaka-50 body, stability, and inertial frames of reference.

In the above, $[V]^B = [u \ v \ w]^T$ is the velocity vector, and $[\omega]^B = [p \ q \ r]^T$ denotes the roll, pitch, and yaw angular rates vector in body frame along $x_B$, $y_B$, and $z_B$. Here, $]^B$, $]^S$,

and $]^E$ are used for body, stability, and inertial frames. Assuming that the thrust forces do not generate net moments and the thrust is in the $x$-direction, we have $L_T = M_T = N_T = F_{T_y} = F_{T_z} = 0$. In this work, the proposed control is implied in the longitudinal channel. Therefore, the aerodynamic forces and moments are formulated as follows (assuming no coupling between the longitudinal and the lateral modes):

$$\begin{bmatrix} L \\ D \\ M_A \end{bmatrix} = \bar{q}S\bar{c} \begin{bmatrix} c_{L_0} & c_{L_\alpha} & c_{L_{\dot{\alpha}}} & c_{L_u} & c_{L_q} & c_{L_{\delta_E}} \\ c_{D_0} & c_{D_\alpha} & c_{D_{\dot{\alpha}}} & c_{D_u} & c_{D_q} & c_{D_{\delta_E}} \\ c_{m_0} & c_{m_\alpha} & c_{m_{\dot{\alpha}}} & c_{m_u} & c_{m_q} & c_{m_{\delta_E}} \end{bmatrix} \begin{bmatrix} 1 \\ \alpha \\ \frac{\dot{\alpha}\bar{c}}{2V_1} \\ \frac{u}{V_1} \\ \frac{q\bar{c}}{2V_1} \\ \delta_E \end{bmatrix}. \tag{3}$$

In addition, obtaining aerodynamic forces in the body frame $[F_A]^B = [F_{A_x} F_{A_y} F_{A_z}]^T$ requires a transfer from the stability to body frame as follows:

$$\begin{bmatrix} F_{A_x} \\ F_{A_y} \\ F_{A_z} \end{bmatrix}^B = \begin{bmatrix} cos\alpha & 0 & -sin\alpha \\ 0 & 1 & 0 \\ sin\alpha & 0 & cos\alpha \end{bmatrix}^{BS} \begin{bmatrix} -D \\ 0 \\ -L \end{bmatrix}^S. \tag{4}$$

The vector of the gravity acceleration in the body axis (defined by (1)) is as follows:

$$\begin{bmatrix} g_x \\ g_y \\ g_z \end{bmatrix}^B = \begin{Bmatrix} -g\sin\theta \\ g\cos\theta\sin\phi \\ g\cos\theta\cos\phi \end{Bmatrix}. \tag{5}$$

Additionally, the rotational kinematic equations are necessary for transfer from the body to the inertial frame to acquire aircraft orientation in the inertial frame and to define the reward function in later sections. In this case, there are three approaches according to [20]. The Euler angle differential equation is used in this research due to the simpler initialization.

$$\begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 1 & \sin\varphi\tan\theta & \cos\varphi\tan\theta \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi/\cos\theta & \cos\varphi/\cos\theta \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix}^B. \tag{6}$$

To calculate velocity and position vector in the inertial frame, a transfer from the body to the inertial frame is necessary. Based on Figure 2, this transformation contains three rotations as follows:

$$[T]^{EB} = [T(\psi)]^{XB}[T(\theta)]^{YX}[T(\phi)]^{EY}. \tag{7}$$

Thus, using (1) and (6), the velocity and position in the inertial frame can be obtained:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix}^E = \begin{bmatrix} cos\psi & -sin\psi & 0 \\ sin\psi & cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}^{XB} \begin{bmatrix} cos\theta & 1 & sin\theta \\ 0 & 1 & 0 \\ -sin\theta & 0 & cos\theta \end{bmatrix}^{YX} \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\phi & -sin\phi \\ 0 & sin\phi & cos\phi \end{bmatrix}^{EY} \begin{bmatrix} u \\ v \\ w \end{bmatrix}^B. \tag{8}$$

Stability and control derivatives for the Chaka-50 are reported in [4] based on Computational Fluid Dynamics (CFD). The summary of these derivatives is presented in Table 1. Before six-degree-of-freedom (6-DoF) simulation using Equations (1) and (2), the trim conditions in a wings-level flight are calculated for simulation verification based on trim equations in [22]. In the drag equation, the absolute value of $\delta_E$, $i_{H_1}$, and $\alpha_1$ is considered. Additionally, flight path angle $\gamma_1$, motor installation angle $\phi_T$, and horizontal tail incidence angle $i_H$ are zero. The elevator deflection $\delta_E$ and required thrust $F_{T_x}$ for a trim flight are obtained and shown in Table 2. The values in the aforesaid table are important for 6-DoF simulation validation. By substituting these coefficients in the trim equations, the presented

airplane should be stabilized in the simulation and maintain its Angle of Attack (AoA) given in Table 2. In this case, the simulation is verified for upcoming steps. Additionally, the geometric, mass, and moment of inertia data are given in Table 3. All simulations are performed in MATLAB R2022a.

**Table 1.** Stability and control derivatives (1/rad).

| Longitudinal Derivatives | Take-Off | Cruise | −10% Model Uncertainty | +10% Model Uncertainty |
|---|---|---|---|---|
| $c_{D_0}$ | 0.0378 | 0.0338 | 0.0304 | 0.0371 |
| $c_{L_0}$ | 0.3203 | 0.3180 | 0.2862 | 0.3498 |
| $c_{m_0}$ | −0.07 | −0.061 | −0.054 | −0.067 |
| $c_{D_\alpha}$ | 0.95 | 0.8930 | 0.8037 | 0.9823 |
| $c_{L_\alpha}$ | 11.06 | 14.88 | 13.39 | 16.37 |
| $c_{m_\alpha}$ | −12.18 | −11.84 | −10.65 | −13.02 |
| $c_{D_u}$ | 0.040 | 0.041 | 0.0369 | 0.0415 |
| $c_{L_u}$ | 0 | 0.081 | 0.0729 | 0.0891 |
| $c_{m_u}$ | 0 | −0.039 | −0.0351 | −0.0429 |
| $c_{D_q}$ | 0 | 0 | 0 | 0 |
| $c_{L_q}$ | 11.31 | 12.53 | 11.27 | 13.78 |
| $c_{m_q}$ | −40.25 | −40.69 | −36 | −44 |
| $c_{D_{\delta_E}}$ | 0.1550 | 0.1570 | 0.1413 | 0.1727 |
| $c_{L_{\delta_E}}$ | 0.96 | 0.78 | 0.702 | 0.858 |
| $c_{m_{\delta_E}}$ | −6.15 | −5.98 | −5.38 | −6.57 |

**Table 2.** Trim parameters of Chaka MRJ.

| Required Thrust ($F_{T_x}$) (lbs) | Angle of Attack ($\alpha°$) | Required Elevator ($\delta_E^\circ$) |
|---|---|---|
| 21,433.02 | 0.39 | −2.28 |

**Table 3.** Chaka-50 required specifics for simulation.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Wing Area ($m^2$) | 43.42 | $I_{xx}$ (kg·$m^2$) | 378,056.535 |
| Mean Aerodynamic Chord (m) | 1.216 | $I_{yy}$ (kg·$m^2$) | 4,914,073.496 |
| Span (m) | 28 | $I_{zz}$ (kg·$m^2$) | 5,670,084.803 |
| Mass (kg) | 18,418.27 | $I_{xz}$ (kg·$m^2$) | 0 |

### *2.1. Atmospheric Disturbance and Sensor Measurement Noise*

Atmospheric disturbance is generally air turbulence in minuscule regions in the atmosphere. According to the literature, atmospheric disturbance is defined as a stochastic process which is characterized by velocity spectral. There are two prevalent models which are usually implemented. In this research, the Dryden turbulence model is utilized [23] for its simpler mathematical formulation.

$$\begin{aligned} G_u(s) &= \sigma_u \sqrt{\frac{2L_u}{\pi V_1}} \left[ \frac{1}{1 + (\frac{L_u}{V_1} s)} \right], \\ G_w(s) &= \sigma_w \sqrt{\frac{2L_w}{\pi V_1}} \left[ \frac{1 + 2\sqrt{3}\frac{L_w}{V_1} s}{(1 + \frac{2L_w}{V_1} s)^2} \right]. \end{aligned} \tag{9}$$

Here, $L_w$, and $L_u$ are the scaling length, and $\sigma_u$, $\sigma_w$ represent the intensity of turbulence. Additionally,

$$\begin{aligned} \sigma_u &= \frac{\sigma_w}{(0.177 + 0.000823z)^{0.4}}, \\ \sigma_w &= 0.1 u_{20}, \end{aligned} \tag{10}$$

where $u_{20}$ is the wind speed at the height of $20ft$, and $V_1 = 160\frac{m}{s}$. Subsequently, the equations of motions are updated according to [24] considering wind effects. The wind components and their derivatives are computed in the inertial frame but usually, the wind calculation in the body frame is more easygoing as an alternative, where the first and third components of wind velocity vector $[W]^B = [W_x \;\; W_y \;\; W_z]^T$ in the body frame are implemented in 6-DoF simulation.

$$\begin{aligned}
\dot{W}_x &= \left[\frac{\partial W_x}{\partial x}\right]^B (u + W_x) + \left[\frac{\partial W_x}{\partial y}\right]^B (v + W_y) + \left[\frac{\partial W_x}{\partial z}\right]^B (w + W_z) + \left[\frac{\partial W_x}{\partial t}\right]^B, \\
\dot{W}_y &= \left[\frac{\partial W_y}{\partial x}\right]^B (u + W_x) + \left[\frac{\partial W_y}{\partial y}\right]^B (v + W_y) + \left[\frac{\partial W_y}{\partial z}\right]^B (w + W_z) + \left[\frac{\partial W_y}{\partial t}\right]^B, \\
\dot{W}_z &= \left[\frac{\partial W_z}{\partial x}\right]^B (u + W_x) + \left[\frac{\partial W_z}{\partial y}\right]^B (v + W_y) + \left[\frac{\partial W_z}{\partial z}\right]^B (w + W_z) + \left[\frac{\partial W_z}{\partial t}\right]^B.
\end{aligned} \tag{11}$$

The spatial derivatives of the wind speed, which are often stated in the inertial frame, must be transferred to the body frame:

$$[\nabla W]^B = [T]^{BE}[\nabla W]^E[\bar{T}]^{BE}. \tag{12}$$

The effect of wind on angular rates $\omega_w^E$ can be defined as a rigid solid air caused by fluid stresses, and it is expressed in the inertial frame as

$$^E = \frac{1}{2}\left[\left(\frac{\partial W_z}{\partial y} - \frac{\partial W_y}{\partial z}\right)\right]^E i + \frac{1}{2}\left[\left(\frac{\partial W_x}{\partial z} - \frac{\partial W_z}{\partial x}\right)\right]^E j + \frac{1}{2}\left[\left(\frac{\partial W_y}{\partial x} - \frac{\partial W_x}{\partial y}\right)\right]^E k. \tag{13}$$

The above equation must be transferred to the body axis so as to be used in the 6-DoF simulation:

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix}^B = \begin{bmatrix} p \\ q \\ r \end{bmatrix}^B - [T]^{BE} \begin{bmatrix} \left(\frac{\partial W_z}{\partial y} - \frac{\partial W_y}{\partial z}\right) \\ \left(\frac{\partial W_x}{\partial z} - \frac{\partial W_z}{\partial x}\right) \\ \left(\frac{\partial W_y}{\partial x} - \frac{\partial W_x}{\partial y}\right) \end{bmatrix}^E. \tag{14}$$

In addition, sensor noises are considered as much as $\pm 10\%$ of the measured pitch angle.

### *2.2. Elevator Fault*

An actuator fault is a kind of failure influencing the inputs of the plant. Due to malformed operation, material aging, or inappropriate maintenance processes, actuator faults may occur in the aircraft. In this work, the actuator fault is formulated as two terms; the first is a multiplicative term that is defined as the elevator's incompetency in reaching the desired quantity, and the second is meant as output quantity bias, namely, an additive term.

$$\begin{aligned}
\delta_{E_t} = 0.6\delta_{E_t} - 0.7^\circ, &\qquad \text{If } t > 12s, \\
0.7\delta_{E_t} + 0.6^\circ, &\qquad \text{If } t > 8s, \\
0.8\delta_{E_t} - 0.5^\circ, &\qquad \text{If } t > 4s,
\end{aligned} \tag{15}$$

The Equation (15) expresses the elevator working with its 60%'s power, and simultaneously a $-0.7^\circ$ output alteration after 12 s of flight. It is assumed the fault will be aggravated over time. Analogous faults with different parameter values occurred after 4 and 8 s of flight where 20% deficiency and $0.5^\circ$ additive bias after 4 s and 30% deficiency and $0.6^\circ$ additive bias after 8 s were considered in this work.

## 3. Attitude Control Using Q-Learning

Truss-braced wing aircraft usually suffer from inadequate stability owing to their narrow mean aerodynamic chord (MAC). More precisely, this fact can be verified by comparing the Phugoid and Short-period modes of Boeing N+3 TTBW [3] and Chaka 50 [4]

with those of Cessna 172 [25]. A summary of numerical values for the longitudinal modes of above-mentioned aerial vehicles has been gathered in Table 4. This table can verify the aforementioned claim. For clarification, the Chaka and Boeing with analogous configurations suffer from poor stability characteristics thanks to their agile characteristics. On the opposite, the Cessna 172 benefits from stable dynamics behaviors. To better demonstrate the dynamic behavior of the air vehicle, simulation results of trim conditions are depicted in Figure 7 over 1500 s. Needless to say, the angle of attack is converged to its theoretical trim value (Table 2). Additionally, the pitch angle is converged in accordance with the angle of attack. However, low-stability existence (resulting in long-time fluctuations) is affected by the damping ratio of the Phugoid mode.

**Table 4.** Longitudinal dynamics characteristics of presented regional aircraft in comparison with other related works.

| Aircraft/ Roots | Short Period Roots | Phugoid Roots |
|---|---|---|
| Chaka 50 | $-0.8 \pm 0.61i$ | $-0.0064 \pm 0.05i$ |
| Cessna 172 | $-3.23 \pm 5.71i$ | $-0.025 \pm 0.19i$ |
| Boeing N+3 | $-0.35 \pm 0.35i$ | $-0.0082 \pm 0.07i$ |

### *3.1. MDP and POMDP Definition in Attitude Control*

To make the attitude control problem suitable for an RL environment, one should define the control problem as an MDP in which the next state of the system could be determined using only the current value of the action and system states [26]. To this end, the problem is formulated as follows: at each time-step $t$, the controller receives the state's information including $\theta_t \in \mathcal{S}_1$, and $\dot{\theta}_t \in \mathcal{S}_2$ from the environment. Based on that, using the current policy of the system, the controller selects an action (which is the elevator deflection, $\delta_{E_t} \in \mathcal{A}(s)$. Subsequently, by applying $\delta_{E_t}$ to the aircraft, the system proceeds to the next step, $\theta_{t+1}$, $\dot{\theta}_{t+1}$, and achieves a reward, $R_{t+1} \in \mathcal{R}$, which is used to evaluate and improve the performance of the current policy. This process continues until reaching final states $\theta_T$, $\dot{\theta}_T$.

$$\theta_0,\ \dot{\theta}_0,\ \delta_{E_0},\ R_1,\ \theta_1,\ \dot{\theta}_1,\ \ldots,\ \theta_T,\ \dot{\theta}_T \tag{16}$$

Accordingly, by observing both $\theta_t$ and $\dot{\theta}_t$ in the state vector of the system, the problem satisfies the Markov property. Moreover, as will be discussed in the following, using only the pitch angle ($\theta_t$) as the system states, which leads to a Partially Observable MDP (POMDP), can also result in a stable control system, though with reduced performance. Now, the purpose of the RL is to find an optimal policy that achieves maximum reward over time. In this regard, the state-action value function $Q_\pi(\theta, \dot{\theta}, \delta_E)$ defines the expected return, which is the sum of discounted rewards starting from one specific state following policy $\pi$ to the terminal state $\theta_T$, $\dot{\theta}_T$.

$$Q_\pi(\theta,\ \dot{\theta},\ \delta_E) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, \theta_t = \theta,\ \dot{\theta}_t = \dot{\theta},\ \delta_{E_t} = \delta_E\right], \tag{17}$$

where $0 < \gamma < 1$ denotes the discount factor.

### *3.2. Structure of Q-Learning Controller*

In this work, the optimal policy in each state is approximated directly using an early breakthrough in RL, namely, the Q-learning [16]. Q-learning is an off-policy, model-free control algorithm, which is based on the Temporal Difference (TD) method as a combination of the dynamic programming and Monte Carlo approaches. Generally, considering an RL environment for the attitude control problem, using the current control command ($\delta_{E_t}$), the system state ($\theta_t, \dot{\theta}_t$) is obtained at each time step, which is then used to modify the action selection policy. Such an iterative process is continued for several episodes so as to reach an optimal strategy. The pseudocode of the Q-learning method is given in Algorithm 1.

**Algorithm 1:** Q-learning Aircraft Attitude Controller

**Data:** Learning Rate $\alpha$, Discount Factor $\gamma$, Desired Angle $\theta_{des} = 1$deg
**Result:** $Q_{\pi^*}(\theta, \dot{\theta}, \delta_E)$
$Q(\theta_0, \dot{\theta}_0, \delta_{E_0}) \leftarrow 0$; **for all** $\theta \in S_1, \dot{\theta} \in S_2, \delta_E \in A(s)$;
**for** *Episode Number = 1 to 20000* **do**
  Initialize 6-DoF simulation with a random $\theta_0 \in [0\ 2]$ deg.
  **for** *time-step (0.01) = 0 to 5 s* **do**
    Select an action $\delta_E$ based on the $\epsilon$-greedy strategy.;
    Execute 6-DoF simulation using computed $\delta_{E_t}$, observe $R_{t+1}$, $\theta_{t+1}$, $\dot{\theta}_{t+1}$.;
    Update the state-action value function:

$$Q(\theta_t, \dot{\theta}_t, \delta_{E_t}) = Q(\theta_t, \dot{\theta}_t, \delta_{E_t}) + \bar{\alpha}\left[R_{t+1} + \gamma \underset{\delta_E}{max} Q(\theta_{t+1}, \dot{\theta}_{t+1}, \delta_E) - Q(\theta_t, \dot{\theta}_t, \delta_{E_t})\right]$$

    Substitute simulation parameters in time-step $t$ with $t + 1$.
  **end**
**end**
Return $Q_{\pi^*}(\theta, \dot{\theta}, \delta_E)$

Due to the nonlinearity and poor stability characteristics of truss-braced wing aircraft, the Q-learning implementation without utilizing NNs can be a challenging problem. However, in the following, it will be shown that by defining a comprehensive reward function along with employing an introduced Fuzzy Action Assignment (FAA) scheme, the air vehicle is capable of following the desired pitch angle even in the presence of sensor measurement noises, atmospheric disturbance, actuator faults, and model uncertainties.

### *3.3. Reward Function and Action Space Definition*

The definition of an effective reward function plays a key role in the convergence of the learning process. Therefore, this research has carefully concentrated on the reward function design and hyper-parameters tuning. In this way, the reward function would be computed in three consecutive steps, while it contains different system variables including $\theta$, $q$, $\delta_E$.

First of all, to restrict the high operating frequency of the elevator, it is essential to give a large punishment in the case of aggressive elevator selection. Thus, in case of an elevator altering more than 0.1 radians, this punishment is applied.

$$Reward_t = -10000, \quad \text{If } \left(|\delta_{E_t}| - |\delta_{E_{t-1}}|\right) > 5.73^\circ. \tag{18}$$

Subsequently, if the operation rate of the elevator is satisfactory, the reward function will be computed as follows if the aircraft is in the vicinity of the desired state.

$$\begin{aligned} Reward_t = (300, &\quad \text{If } |e_{\theta_t}| < 0.05^\circ) \\ + (300, &\quad \text{If } |e_{\theta_t}| < 0.02^\circ) \\ + (400, &\quad \text{If } |q_t| < 0.04^\circ) \\ + (600, &\quad \text{If } |q_t| < 0.02^\circ) \\ + (800, &\quad \text{If } |q_t| < 0.005^\circ), \end{aligned} \tag{19}$$

where $e_{\theta_t} = \theta_t - \theta_{des_t}$ is proportional error. This definition checks the status of pitch tracking first. Then, after the early episodes, the controller finds and prioritizes fewer pitch rates using more reward allocations. The mentioned terms were specified for learning process convergence. In other words, they are involved when the simulated states are in proximity to desired values. However, it is vital to guide the learning process from the first state using another term. Consequently, if none of the above two conditions are met, we should encourage the air vehicle to move towards the desired state. This can be achieved using the following reward function:

$$Reward_t = -(100\,|e_{\theta_t}|)^2 - (40\,|q_t|)^2. \tag{20}$$

Accordingly, the farther the system is from the desired state, the less reward it receives. Additionally, a derivative term (the second term) has been incorporated into the reward function to avoid high pitch rates. The mentioned reward function has been carefully developed during precise dynamics consideration of flight simulations. More specifically, the presence of the pitch rate ($q_{sim}$) in Equation (19), as well as its weight in Equation (20), plays a substantial role in the convergence rate.

Considering the tabular Q-learning, the elevator commands are obtained discretely. Thus, the elevator commands are divided into $-0.25$ to $+0.25$ radians with 0.025 intervals, corresponding to 21 elevator deflections. Additionally, the $\epsilon$-greedy action selection strategy with an epsilon decay is used in this research to enhance the greedy action selection probability in the last episodes. This epsilon decay scheme (Table 5) eliminates the uncertainties in action selections in the last episodes.

$$\delta_{E_t} = \begin{cases} \arg\max\ Q(\theta_t, \dot{\theta}_t, \delta_E) & \text{with probability } 1-\epsilon \\ \text{random action} & \text{with probability } \epsilon \end{cases} \tag{21}$$

**Table 5.** Q-learning and PID simulation parameters.

| Parameter | Value |
|---|---|
| Epsilon($\epsilon$) | $[0.1 : 3 \times 10^{-6}\ ^{1} : 0.04]$ |
| Alpha ($\tilde{\alpha}$) | $[0.02 : 9 \times 10^{-7} : 0.002]$ |
| Gamma ($\gamma$) | 0.99 |
| Episode number | 20,000 |
| $\theta_{observed}$ (rad) | $[-10, -0.024 : 0.002 : -0.002, -0.001, 0]$ |
| $\dot{\theta}_{observed}$ (rad) | $[-10, -0.04, -0.02, -0.005]$ |
| $V_1(\frac{m}{s})$ | 160 |
| $h_{init}(m)$ | 300 |
| $K_p, K_i, K_d$ | $-15, -4, -2$ |

[1] The intervals are written in blue.

### *3.4. Structure of Fuzzy Action Assignment*

The previous discussions were focused on the main structure of the learning process which generates discrete MDP and POMDP models. In order to generate continuous elevator commands, a Fuzzy Action Assignment (FAA) method is proposed here to enhance the performance of the basic Q-learning. In this method, instead of taking a discrete greedy action in a given state $\theta$, $\dot{\theta}$, a relative weight (known also as the validity function or membership function) is assigned to each cell of the grid of system states (see Figure 3) according to the current value of the state-action value function. More precisely, the membership function corresponding to a cell of the grid with centers $\theta_i$ and $\dot{\theta}_j$ is defined as follows:

$$MF_{i,j} = \exp\left(-\frac{1}{2}\left(\frac{\theta_t - \theta_i}{\sigma_\theta}\right)^2\right) \exp\left(-\frac{1}{2}\left(\frac{q_t - \dot{\theta}_j}{\sigma_{\dot{\theta}}}\right)^2\right), \tag{22}$$

where $\sigma_\theta$ and $\sigma_{\dot{\theta}}$ represent the validity widths of membership functions.

Subsequently, $\delta_E$ is calculated at each time-step using a weighted average as follows:

$$\delta_{E_t} = \frac{\sum_i \sum_j MF_{i,j} \arg\max\ Q(\theta_i, \dot{\theta}_j, \delta_E)}{\sum_i \sum_j MF_{i,j}}. \tag{23}$$

The pseudocode of the proposed control strategy is summarized in Algorithm 2. As will be seen in the following section, the employment of the FAA approach results in a robust control system in the presence of different types of uncertain dynamics, while generating feasible control commands. In addition, it should be noted that more effective

approaches to determining the membership functions such as using adaptive membership functions [27,28] can be involved in the design to improve closed-loop performance, which are beyond the scope of this paper.

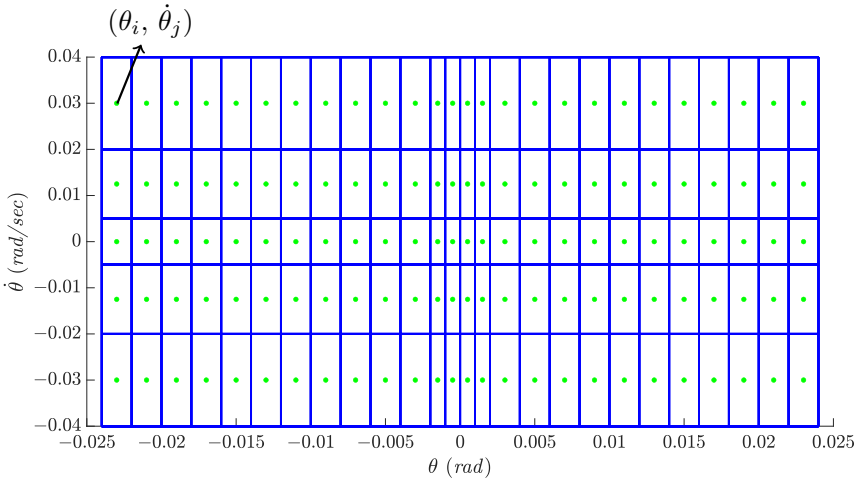


**Figure 3.** Grid of state variables used for tabular Q-learning (The center of each cell, which is used to compute the membership function of the cell is shown by a circle point.)

**Algorithm 2:** Q-learning controller improved by FAA scheme

**Data:** $Q_{\pi^*}(\theta, \dot{\theta}, \delta_E)$
**Result:** Trajectory of the system
Initialize 6-DoF simulation using predefined initial conditions;
**for** *time-step (0.01) = 0 to 5 s* **do**
  Determine the virtual value of the pitch angle (in the case of tracking a variable trajectory);
  Compute the membership function corresponding to each cell of the grid using Equation (22);
  Select an action $\delta_E$ according to FAA technique Equation (23);
  Execute 6-DoF simulation using computed $\delta_E$, compute the next system states;
  Substitute simulation parameters in time-step $t$ with $t + 1$.
**end**

## 4. Simulation Results and Discussion

In this section, flight simulations are performed in two problem frameworks as MDP and POMDP. The difference between them is observing $\dot{\theta}$ in the MDP model, while the POMDP neglects that. Obviously, the state-action table in the MDP model is three-dimensional ($\theta \times \dot{\theta} \times \delta_E$) whereas in the POMDP model, it is shaped as a two-dimensional table ($\theta \times \delta_E$). In addition to the number of observed states, the bounds and intervals of states are important in the convergence time. In this way, to train the RL controller efficiently, it is momentous to divide $\theta$ and $\dot{\theta}$ intervals knowledgeably so as not to lose any important information. As a synopsis, the simulation parameters and the intervals are listed in Table 5.

As seen, a linear decay is used for $\epsilon$ and $\bar{\alpha}$ during episodes. Additionally, the state-action table intervals including $\theta$ and $\dot{\theta}$ are mentioned in this table, where blue numbers denote the interval length. These values are considered symmetrically with positive signs for the positive zone.

Figure 4 shows the rewards of each episode for MDP and POMDP modeling. In early episodes, POMDP results are better and cause fewer fluctuations. However, after about 4000 episodes, MDP starts achieving positive rewards. The MDP converges fairly in episode number 10000 and surpasses the POMDP method. It is worth mentioning that

POMDP never achieves positive rewards. Consequently, encompassing $\dot{\theta}$ plays a significant role in learning efficiency.

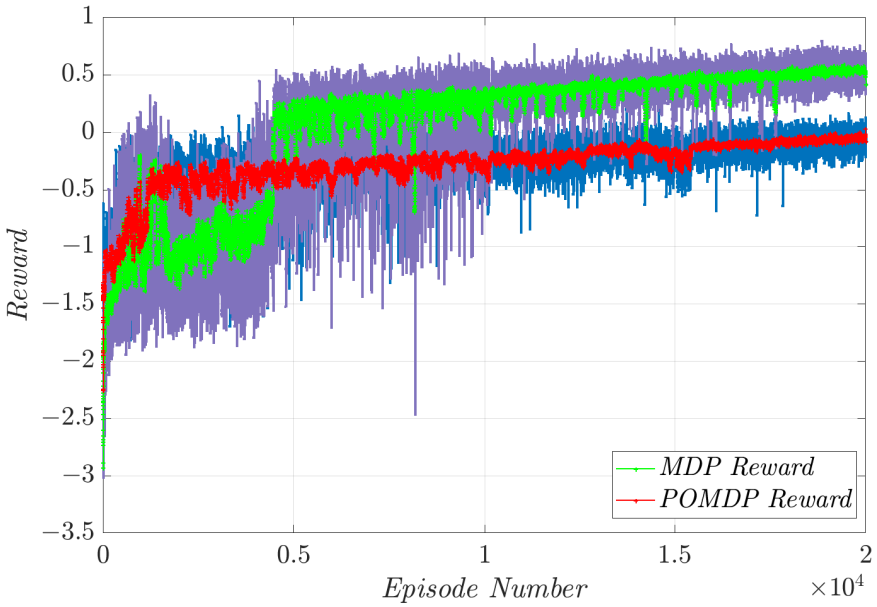


**Figure 4.** MDP and POMDP rewards over 20,000 episodes.

### *4.1. Constant Pitch Angle Tracking*

The result of the proposed Q-learning controller for tracking a constant pitch angle is illustrated in Figure 5 in comparison with a PID controller. Elevator deflections are shown in Figure 6, as well. Obviously, closed-loop performance in the case of POMDP is worse than others because the environment modeling does not satisfy the complete Markov Property. However, it is significantly better than the results shown in Figure 7 for open-loop simulation. To better compare the performance of different control systems, the tracking error is defined as follows:

$$TE = \frac{\int_0^t |\theta_t - \theta_{des}| dt}{t}. \tag{24}$$

Also the control effort is computed as:

$$CE = \frac{\int_0^t |\delta_E| dt}{t}. \tag{25}$$

As can be observed in Table 6, the tracking error using the FAA technique is less than PID, while having the same control effort. Additionally, the overshoot and the settling time in the case of FAA approach prove better tracking results. Furthermore, the FAA technique can satisfactorily solve the issue of oscillations caused by discrete action selections and it simultaneously reduces the control effort.

**Table 6.** Tracking error and control effort of four methods for $\theta_{des} = 1^\circ$.

| Controller | Tracking Error (deg) | Control Effort (deg) | Overshoot (%) | Settling Time (s) |
|---|---|---|---|---|
| MDP | 0.071 | 2.11 | 7.38 | - |
| POMDP | 0.13 | 3.48 | 30.15 | - |
| PID | 0.066 | 0.69 | 27.42 | 5.44 |
| FAA + MDP | 0.057 | 0.69 | 8.20 | 1.76 |

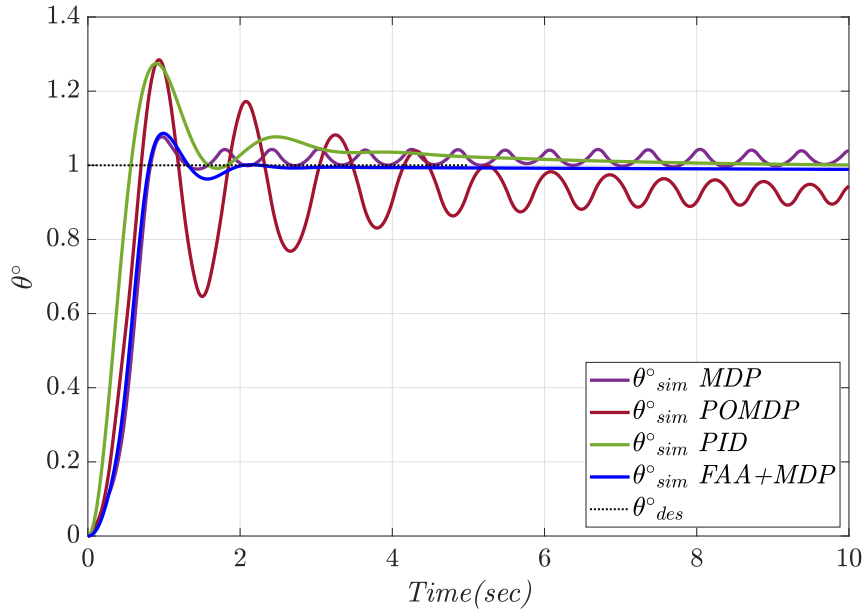


**Figure 5.** Performance of the improved Q-learning controller (using FAA) compared to MDP, POMDP, and PID controller.

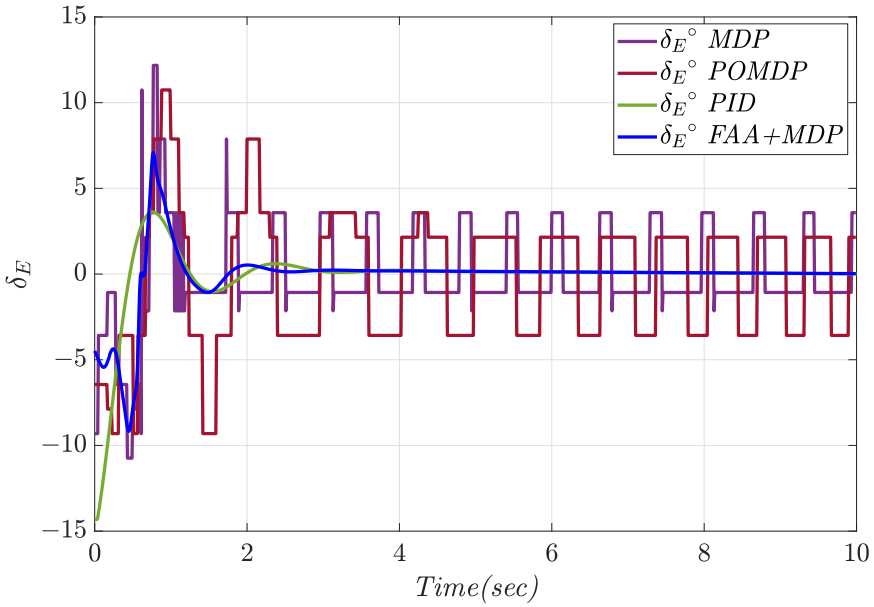


**Figure 6.** Elevator deflections for $\theta_{des} = 1^\circ$ tracking.

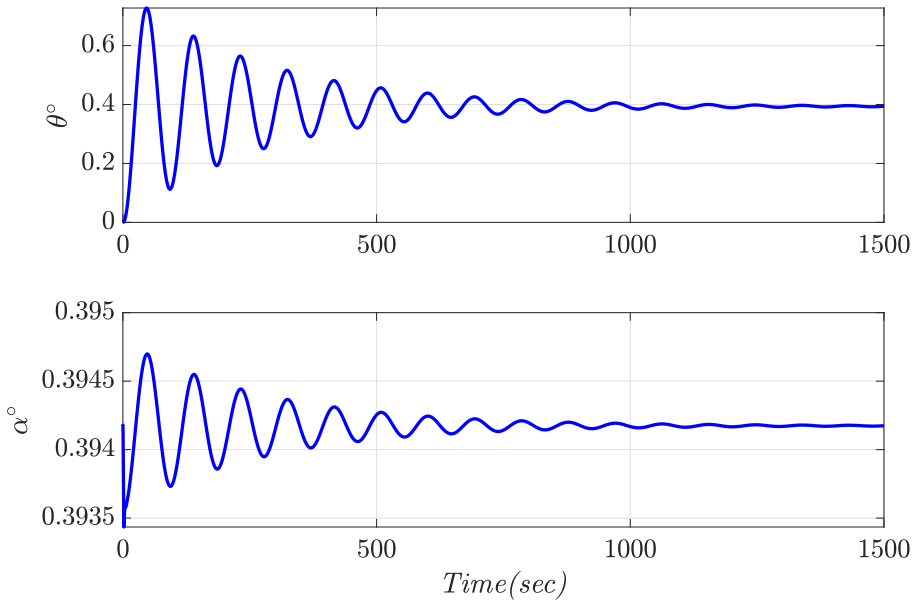


**Figure 7.** The angle of attack and pitch angle variation in the vicinity of trim conditions.

### *4.2. Variable Pitch Angle Tracking*

The tracking of a variable pitch is followed in this study for two reasons; firstly, the learned policy by a fixed desired pitch angle can be used for other angles without the need for retraining. Second, the desired pitch angle for takeoff and landing maneuvers is in the range of $\pm 3^\circ$ as usual. Therefore, the tracking of a variable pitch angle between $\pm 4^\circ$ is reasonable. The tracking result for a variable pitch angle is illustrated in Figure 8. At first glance, it is obvious that both MDP and FAA-improved MDP Q-learning were able to track variable $\theta_{des}$ in ideal conditions. However, the high working frequency of the elevator in the case of MDP (shown in Figure 9) proves the significant superiority of the FAA scheme.

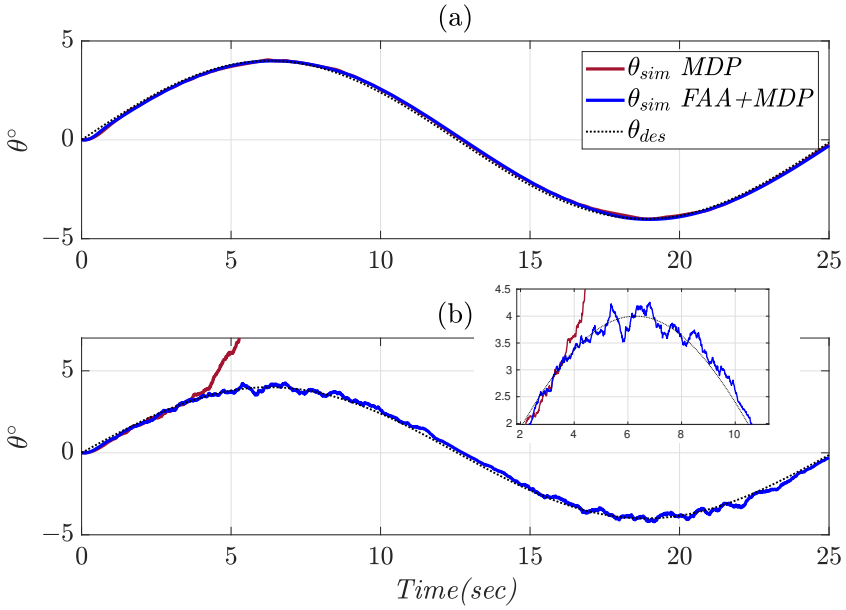


**Figure 8.** Variable $\theta$ tracking using MDP and FAA-improved MDP (**a**) in ideal conditions, (**b**) in the presence of sensor measurement noises and atmospheric disturbances.

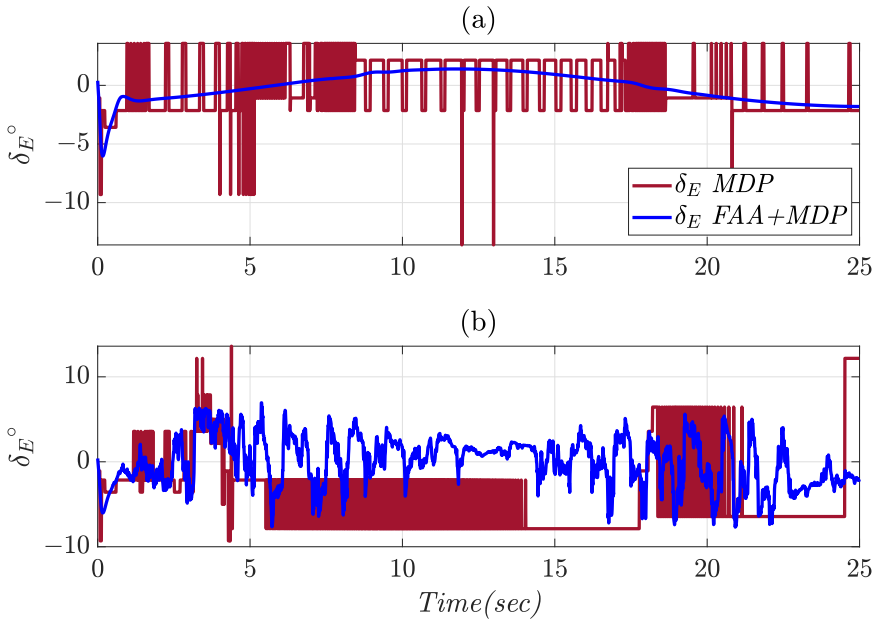


**Figure 9.** Elevator deflections in variable $\theta$ tracking (**a**) in ideal conditions, (**b**) in the presence of sensor measurement noises and atmospheric disturbances.

Now, to evaluate the robustness of the proposed control scheme, closed-loop response is investigated in the presence of different types of uncertain dynamics including measurement noises, atmospheric disturbance, actuator faults, and model parameter uncertainties. Table 7 illustrates a comparison between various flight conditions. Considering ideal flight conditions, the analogous tracking error between MDP and FAA is noticeable. However, the less control effort using FAA proves its efficiency. On the other hand, in the presence of sensor measurement noises and atmospheric disturbances, the proposed method is significantly superior compared to the basic Q-learning. More specifically, using the FAA approach, the control effort even considering measurement noises and external disturbances is less than that of the MDP in ideal conditions. It should be noted that, in the simultaneous presence of sensor measurement noises and atmospheric disturbances, the basic Q-learning controller is unable to stabilize the system (see Figure 8b), while using the proposed method, the air vehicle could efficiently follow the desired trajectory.

**Table 7.** Tracking error and control effort of two methods in variable $\theta_{des}$ tracking in different flight conditions.

| Controller | Flight Condition | Tracking Error (deg) | Control Effort (deg) |
|---|---|---|---|
| MDP | Ideal | **0.108** [1] | 2.224 |
| FAA + MDP | Ideal | 0.112 | **1.008** |
| MDP | Noise + Disturbance | 23.816 | 5.424 |
| FAA + MDP | Noise + Disturbance | **0.132** | **2.032** |
| MDP | Actuator Fault | 0.304 | 1.796 |
| FAA + MDP | Actuator Fault | **0.136** | **1.004** |
| FAA + MDP | −10% Uncertainty | 0.116 | 1.16 |
| FAA + MDP | +10% Uncertainty | 0.116 | **0.972** |

[1] The best results of each group are written in bold.

Another simulation is performed to evaluate the proposed method's robustness in the presence of actuator faults and model uncertainties. In this regard, a sequence of both multiplicative and additive actuator faults is applied to the elevator control surface based on Equation (15). Additionally, the model parameter uncertainties are mentioned in Table 1.

The system response is obtained as shown in Figure 10. As can be observed, in the MDP case, the system experienced more tracking errors between 12 to 20 s but the FAA controller produced satisfactory outcomes considering both control effort and tracking errors. Apart from that, the generated $\delta_E$ in the MDP case is not applicable in real applications owing to high working frequency. The FAA method is able to generate satisfactory findings even in the presence of model uncertainties in the range of $\pm 10\%$. As a result, the proposed method is effectively capable of dealing with different uncertain terms, which are inevitable in real flight experiments.

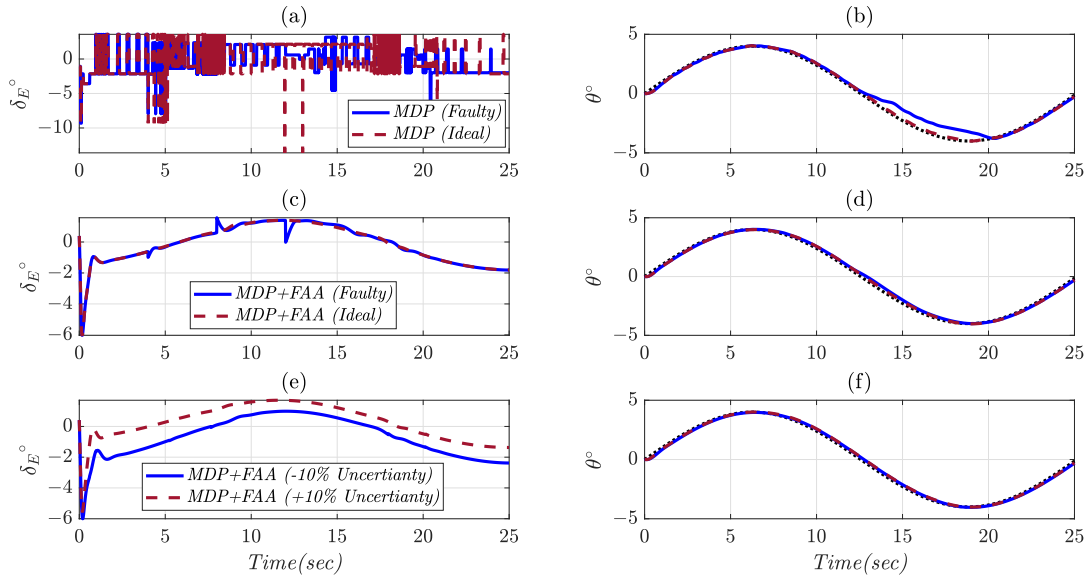


**Figure 10.** (**a**) Elevator deflection of basic MDP Q-learning in faulty and ideal flight conditions; (**b**) Pitch angle tracking related to (**a**). (**c**) Elevator deflection of FAA-improved Q-learning in ideal and faulty conditions; (**d**) Pitch angle tracking related to (**c**). (**e**) Elevator deflection of FAA-improved Q-learning in the presence of $\pm 10\%$ model uncertainties; (**f**) Pitch angle tracking related to (**e**).

## 5. Conclusions

This work proposed a robust attitude control system for an agile aircraft using an improved Q-learning under both MDP and POMDP problem modeling. The aircraft is a novel regional truss-braced wing airplane, which was treated as an RL environment. It was shown that by defining an appropriate reward function and satisfying the Markov property, Q-learning, even considering the tabular case, is convergent for such a challenging system. This method was verified in constant and variable $\theta_{des}$ tracking simulations, where the variable pitch angle tracking became possible using a novel efficient transformation, which maps the real pitch angle into a virtual variable according to the instantaneous difference between the real pitch angle and the desired angle. Finally, the FAA-augmented method was introduced to construct continuous control commands, eliminate response fluctuations, and provide a robust control system in the presence of atmospheric disturbances, sensor measurement noises, actuator faults, and model uncertainties.

**Author Contributions:** Conceptualization, M.Z., S.A.E. and A.B.; methodology, M.Z. and S.A.E.; software, M.Z. and S.A.E.; validation, M.Z., S.A.E. and A.B.; formal analysis, M.Z. and S.A.E.; investigation, M.Z. and S.A.E.; resources, M.Z. and S.A.E.; data curation, M.Z. and S.A.E.; writing—original draft preparation, M.Z.; writing—review and editing, M.Z. and S.A.E.; visualization, M.Z.; supervision, S.A.E., A.B. and P.C.; project administration, M.Z.; funding acquisition, P.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Nomenclature

| | |
|---|---|
| $[u\ v\ w]$ | Velocity components in body frame |
| $[p\ q\ r]$ | Angular velocity components in body frame |
| $\phi, \theta, \psi$ | Roll, pitch, and yaw angles |
| $I_x, I_y, I_z$ | Moments of inertia in body frame |
| $m$ | Mass of airplane |
| $F_A$ | Aerodynamic force vector in body frame |
| $F_T$ | Engine thrust vector in body frame |
| $L_A$ | Aerodynamic moment vector |
| $L_T$ | Thrust moment vector in body frame |
| $g$ | Acceleration of gravity |
| $D$ | Drag |
| $L$ | Lift |
| $\alpha$ | Angle of attack |
| $\delta_E$ | Elevator deflection |
| $\bar{q}$ | Airplane dynamic pressure |
| $\bar{c}$ | Mean aerodynamic chord |
| $S$ | Wing Area |
| $Q(s, a)$ | State-action value function |
| $R$ | Reward |
| $\bar{\alpha}$ | Learning rate |
| $K_p, K_i, K_d$ | Proportional, integral, derivative coefficients of PID controller |

## References


1. Li, L.; Bai, J.; Qu, F. Multipoint Aerodynamic Shape Optimization of a Truss-Braced-Wing Aircraft. *J. Aircr.* **2022**, *59*, 1–16. [CrossRef]
2. Sarode, V.S. Investigating Aerodynamic Coefficients and Stability Derivatives for Truss-Braced Wing Aircraft Using OpenVSP. Ph.D. Thesis, Virginia Tech, Blacksburg, VA, USA, 2022.
3. Nguyen, N.T.; Xiong, J. Dynamic Aeroelastic Flight Dynamic Modeling of Mach 0.745 Transonic Truss-Braced Wing. In Proceedings of the AIAA SCITECH 2022 Forum, San Diego, CA, USA, 3–7 January 2022; p. 1325.
4. Zavaree, S.; Zahmatkesh, M.; Eghbali, K.; Zahiremami, K.; Vaezi, E.; Madani, S.; Kariman, A.; Heidari, Z.; Mahmoudi, A.; Rassouli, F.; et al. *Modern Regional Jet Family (Chaka: A High-Performance, Cost-Efficient, Semi-Conventional Regional Jet Family)*; AIAA: Reston, VA, USA, 2021. Available online: https://www.aiaa.org/docs/default-source/uploadedfiles/education-and-careers/university-students/design-competitions/winning-reports---2021-aircraft-design/2nd-place---graduate-team---sharif-university-of-technology.pdf?sfvrsn=41350e892 (accessed on 5 December 2022).
5. Emami, S.A.; Castaldi, P.; Banazadeh, A. Neural network-based flight control systems: Present and future. *Annu. Rev. Control* **2022**, *53*, 97–137. [CrossRef]
6. Xi, Z.; Wu, D.; Ni, W.; Ma, X. Energy-Optimized Trajectory Planning for Solar-Powered Aircraft in a Wind Field Using Reinforcement Learning. *IEEE Access* **2022**, *10*, 87715–87732. [CrossRef]
7. Bøhn, E.; Coates, E.M.; Reinhardt, D.; Johansen, T.A. Data-Efficient Deep Reinforcement Learning for Attitude Control of Fixed-Wing UAVs: Field Experiments. *arXiv* **2021**, arXiv:2111.04153.
8. Yang, X.; Yang, X.; Deng, X. Horizontal trajectory control of stratospheric airships in wind field using Q-learning algorithm. *Aerosp. Sci. Technol.* **2020**, *106*, 106100. [CrossRef]
9. Hu, W.; Gao, Z.; Quan, J.; Ma, X.; Xiong, J.; Zhang, W. Fixed-Wing Stalled Maneuver Control Technology Based on Deep Reinforcement Learning. In Proceedings of the 2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI), Fuzhou, China, 8–10 July 2022; pp. 19–25.
10. Xue, W.; Wu, H.; Ye, H.; Shao, S. An Improved Proximal Policy Optimization Method for Low-Level Control of a Quadrotor. *Actuators* **2022**, *11*, 105. [CrossRef]
11. Wang, Z.; Li, H.; Wu, H.; Shen, F.; Lu, R. Design of Agent Training Environment for Aircraft Landing Guidance Based on Deep Reinforcement Learning. In Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 8–9 December 2018; Volume 02, pp. 76–79. [CrossRef]
12. Yuan, X.; Sun, Y.; Wang, Y.; Sun, C. Deterministic Policy Gradient with Advantage Function for Fixed Wing UAV Automatic Landing. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; pp. 8305–8310. [CrossRef]

13. Tang, C.; Lai, Y.C. Deep Reinforcement Learning Automatic Landing Control of Fixed-Wing Aircraft Using Deep Deterministic Policy Gradient. In Proceedings of the 2020 International Conference on Unmanned Aircraft Systems (ICUAS), Athens, Greece, 1–4 September 2020; pp. 1–9. [CrossRef]
14. Dai, H.; Chen, P.; Yang, H. Fault-Tolerant Control of Skid Steering Vehicles Based on Meta-Reinforcement Learning with Situation Embedding. *Actuators* **2022**, *11*, 72. [CrossRef]
15. Richter, D.J.; Natonski, L.; Shen, X.; Calix, R.A. Attitude Control for Fixed-Wing Aircraft Using Q-Learning. In *Intelligent Human Computer Interaction*; Kim, J.H., Singh, M., Khan, J., Tiwary, U.S., Sur, M., Singh, D., Eds.; Springer International Publishing: Cham, Swizerland, 2022; pp. 647–658.
16. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
17. Glorennec, P.; Jouffe, L. Fuzzy Q-learning. In Proceedings of the 6th International Fuzzy Systems Conference, Barcelona, Spain, 5 July 1997; Volume 2, pp. 659–662. [CrossRef]
18. Er, M.J.; Deng, C. Online tuning of fuzzy inference systems using dynamic fuzzy Q-learning. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2004**, *34*, 1478–1489. [CrossRef] [PubMed]
19. Napolitano, M.R. *Aircraft Dynamics*; Wiley: Hoboken, NJ, USA, 2012.
20. Zipfel, P. *Modeling and Simulation of Aerospace Vehicle Dynamics*, 3rd ed.; AIAA: Reston, VA, USA, 2014.
21. Wood, A.; Sydney, A.; Chin, P.; Thapa, B.; Ross, R. GymFG: A Framework with a Gym Interface for FlightGear. *arXiv* **2020**, arXiv:2004.12481.
22. Roskam, J. *Airplane Flight Dynamics and Automatic Flight Controls*; DARcorporation: Lawrence, KS, USA, 1998.
23. Mil-f, V. *8785c: Flying Qualities of Piloted Airplanes*; US Air Force: Washington, DC, USA, 1980; Volume 5.
24. Frost, W.; Bowles, R.L. Wind shear terms in the equations of aircraft motion. *J. Aircr.* **1984**, *21*, 866–872. [CrossRef]
25. Çetin, E. System identification and control of a fixed wing aircraft by using flight data obtained from x-plane flight simulator. Master's Thesis, Middle East Technical University, Ankara, Turkey, 2018.
26. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
27. Emami, S.A.; Banazadeh, A. Intelligent trajectory tracking of an aircraft in the presence of internal and external disturbances. *Int. J. Robust Nonlinear Control* **2019**, *29*, 5820–5844. [CrossRef]
28. Emami, S.A.; Ahmadi, K.K.A. A self-organizing multi-model ensemble for identification of nonlinear time-varying dynamics of aerial vehicles. *Proc. Inst. Mech. Eng. Part I J. Syst. Control. Eng.* **2021**, *235*, 1164–1178. [CrossRef]