aerospace

MDPI

Article

Explainable Reinforcement Learning Framework for Autonomous Windshear Escape with Policy Distillation

Yitan Wang [1], Yangyang Zhang [2,*] and Zhenxing Gao [2]

[1] Shandong Airlines, Jinan 250014, China
[2] College of General Aviation and Flight, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; z.x.gao@nuaa.edu.cn
* Correspondence: zhang_yy@nuaa.edu.cn; Tel.: +86-185-5178-1583

Abstract

Low-altitude micro downbursts pose a severe threat to aviation safety, yet conventional control approaches and standard deep reinforcement learning (DRL) often fail due to explicit modeling difficulties and sparse reward constraints. To address these challenges, this study proposes an explainable, data-driven framework integrating active-reward proximal policy optimization (AR-PPO). A bilevel optimization architecture driven by meta-gradients is developed to dynamically discover optimal reward functions without human intervention. Furthermore, a policy distillation pipeline utilizing wavelet-multivariate singular spectrum analysis (W-MSSA) and classification and regression trees (CART) is proposed to translate high-frequency continuous neural outputs into discrete, pilot-readable rules. Simulation results on a B737-800 model demonstrate that AR-PPO effectively overcomes the "stall trap" by autonomously learning to trade altitude for airspeed, outperforming static-reward baselines and empirical human pilots in extreme, zero-shot windshear encounters (22.0 m/s downdraft). Ultimately, the proposed framework successfully distills black-box AI strategies into verifiable, physics-informed standard operating procedures (SOPs), providing a highly transparent and robust solution for autonomous windshear escape and future competency-based flight training.

Keywords: deep reinforcement learning; explainable reinforcement learning; micro downburst; policy distillation; autonomous flight control

1. Introduction

With the rapid expansion of global civil aviation, ensuring flight safety in complex environments has become a key challenge. The transition from traditional automation to intelligent flight control is widely seen as an essential path to address this issue [1]. While modern autopilot systems are highly reliable under nominal conditions, their performance often drops significantly when facing unforeseen disturbances or extreme weather. Therefore, developing adaptive control algorithms that can make autonomous decisions in dynamic environments has become a major focus of aerospace research.

While recent advances in Doppler weather radar, low-level windshear alert systems, and onboard predictive windshear detection technologies have significantly reduced the likelihood of unexpected encounters, sudden low-altitude micro downbursts can still occur with minimal warning. Although microburst-related incidents have become relatively rare due to these improved detection capabilities and enhanced competency-based pilot training, they remain a severe and highly asymmetric threat when encountered during the

Check for updates

Academic Editor: Jules Simo

Received: 3 May 2026
Revised: 29 July 2026
Accepted: 5 August 2026
Published: 13 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

critical final approach phase, owing to their potential to cause a rapid and massive loss of aerodynamic energy. To address this, adaptive flight envelope protection methods have been developed to employ command filtering and haptic feedback to improve situational awareness while keeping the aircraft within its physical limits [2]. However, even advanced robust optimization frameworks are limited by wind field estimation errors during autonomous flight [3]. Conventional control approaches typically depend on an explicit model of the wind field, which is difficult to identify in real-time. In highly unpredictable conditions, the delay between disturbance detection and controller response often exceeds the narrow safety margins required for low-altitude recovery.

The emergence of deep reinforcement learning (DRL) offers a model-free alternative that can potentially mitigate these modeling difficulties. Research has demonstrated the effectiveness of DRL in handling multimodal data for crosswind takeoffs and coordinating complex maneuvers for dynamic soaring without explicit trajectory planning [4–6]. Despite these successes, the application of DRL to windshear escape reveals significant stability issues. In various UAV applications, DRL-based strategies exhibit substantial tracking errors under severe turbulence [7]. To solve this, hybrid architectures, such as combining TD3 agents with conventional PID loops or using self-attention mechanisms, have been proposed to manage the high-bandwidth dynamics of altitude and attitude channels [8,9]. Unfortunately, the applications of the RL algorithm mentioned above, which rely on manually designed reward functions, often lead to low sample efficiency and potential mission failure in complex real-world tasks.

To overcome the sparse reward issues, inverse reinforcement learning (IRL) and imitation learning (IL) have been developed. The advantage of IRL lies in its ability to infer deep reward function motivations from expert behavior, while IL achieves rapid cloning of complex policies by directly mapping states to actions. However, IRL suffers from reward ambiguity and high computational cost, while IL is prone to policy collapse due to error accumulation when encountering rare states that deviate from the expert demonstration distribution [10–12]. Recently, researchers have increasingly adopted intrinsic motivation and active exploration frameworks. This represents a paradigm shift from handcrafted rules toward autonomously discovered learning laws [10]. Foundational mechanisms, such as curiosity-driven exploration, leverage self-supervised prediction errors to encourage agents to investigate environmental features that influence their dynamics [11]. More robust methods, including random network distillation, achieve state-of-the-art performance in environments where extrinsic rewards are absent or extremely rare [12]. In continuous control domains, these directed exploratory behaviors aim to maximize value function coverage [13]. Recent breakthroughs even enable the meta-learning of RL rules or the end-to-end discovery of reward functions for embodied agents [14,15], often utilizing bilevel optimization and meta-gradients to autonomously discover generalized optimal rewards without human intervention [16–18]. Furthermore, specialized algorithms like active exploration deep reinforcement learning (AEDRL) utilize Gaussian processes to model dynamic uncertainty, optimizing action selection to minimize epistemic uncertainty [19,20]. By maximizing skill diversity or leveraging mutual information to enhance controllability, agents can develop robust skills that generalize to unseen, high-risk scenarios [21,22]. While frameworks like ActSafe ensure that such exploration occurs within a pessimistic safe set, guaranteeing safety during training while pursuing near-optimal recovery policies [23].

Despite these advancements in optimal policies, the unexplainability of DRL remains a significant bottleneck for airworthiness certification. To establish operational trust, explainable reinforcement learning (XRL) must transition from basic visualization to providing mechanistic evidence of decision-making. Recent findings suggest that model-free agents can internally develop planning structures that predict long-term consequences, similar

to emergent planning in humans [24]. To decode these internal representations, model-level approaches like shapley value-based interpretable policy via explanation verification (SILVER) utilize feature attribution and RL-guided labeling to generate behaviorally consistent datasets for surrogate models [25]. However, these methods predominantly focus on extracting the cognitive intent or feature importance within the network. Furthermore, distilling DRL policies into differentiable decision trees or hierarchical skilltrees enables the reduction of complex action spaces into interpretable discrete skills for long-horizon control [26,27]. These methodologies are increasingly applied to high-dimensional problems, such as compressor design optimization [28]. For multi-agent systems, temporally-aware frameworks allow for the visualization of causal effects and decision histories [29,30]. The most rigorous approach, Programmatically Interpretable RL (PIRL), represents policies using domain-specific languages amenable to symbolic verification [31]. Such programmatic policies often yield smoother trajectories and superior transferability, while the discovery of governing equations ensures that the learned logic remains physically grounded [32].

In summary, while significant progress has been made in autonomous flight control and active exploration, integrating safety-constrained active exploration with data-reconstruction-based feature extraction remains an open challenge, particularly in extreme windshear environments like micro downbursts. Moreover, current XRL methods primarily focus on interpreting the internal rationale of neural networks (i.e., explaining why decisions are made). This poses a major limitation in aerospace engineering, where strict airworthiness certification demands a precise understanding of how control commands are physically executed and coordinated. To address these issues, this study proposes a data-driven, explainable reinforcement learning framework. The main contributions of this paper are summarized as follows:

Proposal of an autonomous reward discovery framework via bilevel optimization. A novel active-reward proximal policy optimization (AR-PPO) framework driven by meta-gradients is proposed to tackle the inherent sparse reward and temporal credit assignment problems in low-altitude windshear encounters. Instead of relying on handcrafted, static reward shaping, the proposed framework employs a bilevel optimization architecture to autonomously map sparse episodic outcomes, such as survival or crash, into dense, state-action-level evaluations, effectively resolving the severe gradient interference between multiple safety constraints.

Zero-shot adaptation and robust recovery in unseen extreme wind fields. Driven by the meta-gradient mechanism, the agent could successfully break the heuristic stall trap, actively learning to prioritize kinetic energy (airspeed preservation) over potential energy (altitude maintenance) during vortex core penetration. This provides a globally optimal recovery trajectory in extreme downdrafts. By leveraging the active exploration mechanism to thoroughly explore state-space uncertainties, the trained RL control law achieves rapid zero-shot adaptation to unencountered Micro downbursts with higher intensities and varying core radii without the need for retraining.

Explainable policy distillation for human-readable SOPs. This study mitigates the black-box limitations of DRL by shifting the interpretability focus from internal network intent to physical execution. Using wavelet analysis combined with multivariate singular spectrum analysis (W-MSSA), the high-frequency control outputs are precisely denoised and decoupled. By integrating phase plane mapping with classification and regression tree (CART) policy distillation, the learned extreme recovery strategies are translated into human-readable "IF–THEN" standard operating procedures (SOPs). This could provide a verifiable, physics-based reference for future competency-based training and assessment (CBTA) and evidence-based training (EBT).

This paper is organized as follows. In Section 2, the methodology integrating an active exploration reinforcement learning framework is described, including flight dynamic model with micro downburst model, reward function design, multivariable decoupling and signal reconstruction via W-MSSA and rule extraction and SOPs generation, which are employed in Section 3. The proposed framework is demonstrated with simulation experiments evaluating zero-shot adaptation in unseen micro downburst and extracting human-readable SOPs. The discussion and conclusions are presented in Section 4 and Section 5, respectively.

2. Materials and Methods

2.1. Dynamic Environment of RL

2.1.1. Aircraft Equations of Motion

As shown in Figure 1, the overall framework is divided into two primary stages: the training procedure and the explainable policy distillation procedure. During the training phase, the agent continuously interacts with a simulated environment that integrates a flight dynamic model and a windshear model, optimizing its control policy based on state and reward feedback. Once the agent is fully trained, its generated action sequences are fed into the distillation pipeline. Here, W-MSSA is first applied for multivariable decoupling and signal reconstruction to filter out algorithmic noise. Subsequently, sequential pattern mining is employed alongside comparisons with empirical pilot action sequences. This final step serves to identify human operational issues, provide quantitative flight guidance, and fundamentally refine emergency response procedures.

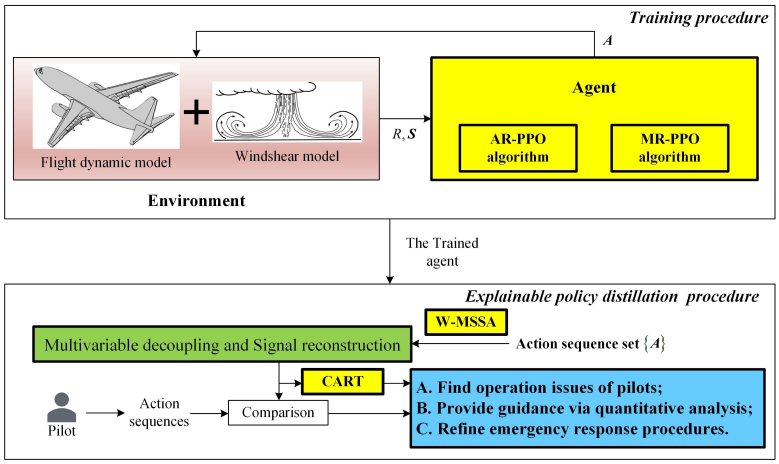


Figure 1. Overall framework of this study.

The flight of an aircraft is not only influenced by gravity, but also by aerodynamics, engine thrust, and external disturbances. Due to these forces and moments, the aircraft undergoes translational motion along the three axes and rotational motion around the three axes in body frame, as shown in Figure 2.

The motion of an aircraft could be described as 12 state equations, including the center of mass motion equations, kinematic equations, moment equations, and navigation equations [33]. The equations of aircraft center of mass motion under body frame are:

$$\begin{cases} \dot{V}_x = \frac{X}{m} - \dot{W}_{xB} - g\sin\theta - [q(V_z + W_{zB}) - r(V_y + W_{yB})] \\ \dot{V}_y = \frac{Y}{m} - \dot{W}_{yB} + g\sin\phi\cos\theta - [r(V_x + W_{xB}) - p(V_z + W_{zB})] \\ \dot{V}_z = \frac{Z}{m} - \dot{W}_{zB} + g\cos\phi\cos\theta - [p(V_y + W_{yB}) - q(V_x + W_{xB})] \end{cases} \quad (1)$$

where V_x, V_y and V_z are the triaxial components of true airspeed $V = \sqrt{V_x^2 + V_y^2 + V_z^2}$ in body frame, and $\dot{V}_x$, $\dot{V}_y$ and $\dot{V}_z$ are the derivative of V_x, V_y and V_z over time. X, Y and Z are the force on the three-axis of body frame, and m is the mass of aircraft. ϕ and θ are the roll angle and pitch angle, and g is the gravitational acceleration. p, q and r are the angular velocities in body frame. W_{xB}, W_{yB} and W_{zB} are the triaxial components of the wind speed in body frame, and the coordinate conversion from inertial frame to body frame is:

$$\begin{bmatrix} W_{xB} \\ W_{yB} \\ W_{zB} \end{bmatrix} = \begin{bmatrix} \cos\psi\cos\theta & -\sin\psi\cos\phi + \cos\psi\sin\theta\sin\phi & \sin\psi\sin\phi + \cos\psi\sin\theta\cos\phi \\ \sin\psi\cos\theta & \cos\psi\cos\phi + \sin\psi\sin\theta\sin\phi & -\sin\phi\cos\psi + \sin\psi\sin\theta\cos\phi \\ -\sin\theta & \cos\theta\sin\phi & \cos\theta\cos\phi \end{bmatrix} \begin{bmatrix} W_x \\ W_y \\ W_z \end{bmatrix} \tag{2}$$

where ψ is the yaw angle, and W_x, W_y and W_z are the wind speed in inertial frame. The kinematic equations are:

$$\begin{cases} \dot{\phi} = p + \tan\theta(q\sin\phi + r\cos\phi) \\ \dot{\theta} = q\cos\phi - r\sin\phi \\ \dot{\psi} = (q\sin\phi + r\cos\phi)/\cos\theta \end{cases} \tag{3}$$

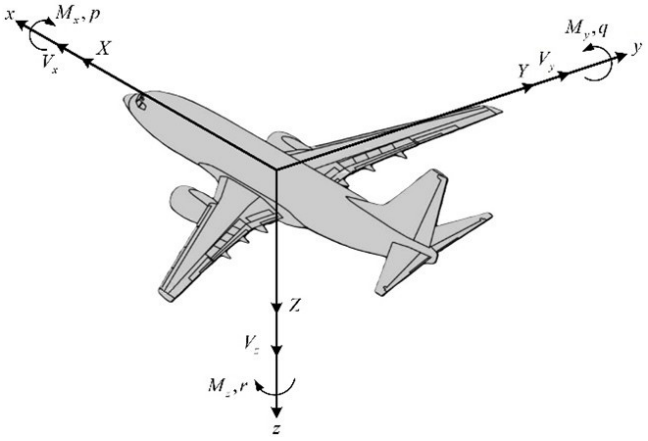


Figure 2. Six degrees of freedom motion of an aircraft.

The rotational motion equations in body frame are:

$$\begin{cases} \Gamma\dot{p} = I_{xz}(I_x - I_y + I_z)pq - [I_z(I_z - I_y) + I_{xz}^2]qr + I_z M_x + I_{xz}M_z \\ I_y\dot{q} = (I_z - I_x)pr - I_{xz}(p^2 - r^2) + M_y \\ \Gamma\dot{r} = [(I_x - I_y)I_x + I_{xz}^2]pq - I_{xz}(I_x - I_y + I_z)qr + I_{xz}M_x + I_x M_z \end{cases} \tag{4}$$

where $\Gamma = I_x I_z - I_{xz}^2$, and I_x is the moment of inertia of x-axis in body frame, and I_y is the moment of inertia of y-axis in body frame, and I_z is the moment of inertia of z-axis in body frame. I_{xz} is the inertia product of x-axis and z-axis. $\dot{p}$, $\dot{q}$ and $\dot{r}$ are the derivative of p, q and r over time. M_x is the roll moment, and M_y is the pitch moment, and M_z is the yaw moment. The navigation equations with disturbance wind influence are:

$$\begin{cases} \dot{x}_E = V_x\cos\theta\cos\psi + V_y(-\cos\phi\sin\psi + \sin\phi\sin\theta\cos\psi) \\ \quad + V_z(\sin\phi\sin\psi + \cos\phi\sin\theta\cos\psi) + W_x \\ \dot{y}_E = V_x\cos\theta\sin\psi + V_y(\cos\phi\cos\psi + \sin\phi\sin\theta\sin\psi) \\ \quad + V_z(-\sin\phi\cos\psi + \cos\phi\sin\theta\sin\psi) + W_y \\ \dot{z}_E = V_x\sin\theta - V_y\sin\phi\cos\theta - V_z\cos\phi\cos\theta + W_z \end{cases} \tag{5}$$

where $\dot{x}_E$, $\dot{y}_E$ and $\dot{z}_E$ are the ground speed in inertial frame. And the angle of attack α and sideslip angle β are calculated as:

$$\begin{cases} \alpha = \arctan(V_z / V_x) \\ \beta = \arcsin(V_y / V) \end{cases} \tag{6}$$

2.1.2. Micro Downburst Model

For the follow-up simulation, this study built a micro downburst, a type of windshear, based on vortex rings [34]. Figure 3 shows the principle of modeling the windshear field using vortex rings.

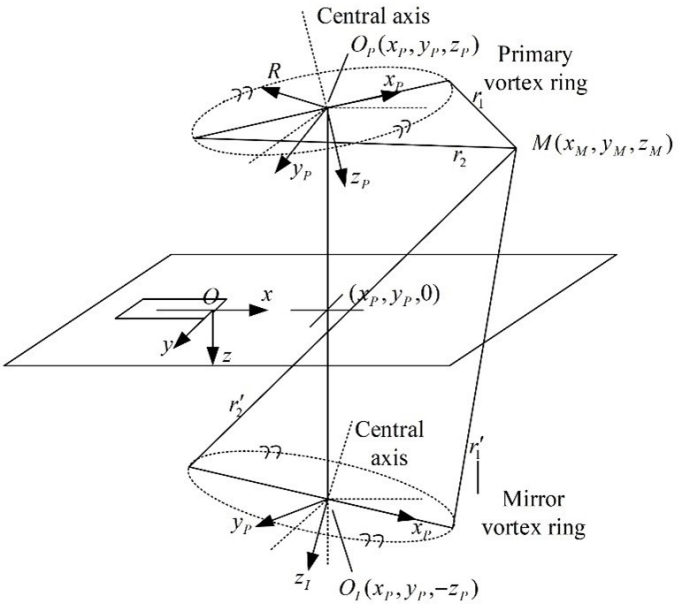


Figure 3. Windshear modelling by using vortex rings.

In the ground coordinate system O_{xyz}, the origin is set to a point on the airport runway, and the approach direction is along the axis x. The primary vortex ring has a mirror vortex ring symmetrical relative to the ground, whose coordinate origins are $O_P(x_P, y_P, z_P)$ and $O_I(x_P, y_P, -z_P)$, respectively. The center of the vortex ring is O_p. The radius of the vortex ring is R, and the radius of the vortex core is r. r_1 is the distance from point M to the nearest point of the primary vortex ring, and r_2 is the distance from point M to the farthest point of the primary vortex ring. r'_1 is the distance from point M to the nearest point of the mirror vortex ring, and r'_2 is the distance from point M to the farthest point of the mirror vortex ring.

The streamline equation of M can be expressed as:

$$\psi = -\frac{\Gamma}{2\pi}\left[\frac{0.788k^2(r_1 + r_2)}{0.25 + 0.75\sqrt{1 - k^2}} - \frac{0.788k'^2(r'_1 + r'_2)}{0.25 + 0.75\sqrt{1 - k'^2}}\right] \tag{7}$$

where Γ is the strength of the vortex ring, and $k = \left|\frac{r_2 - r_1}{r_2 + r_1}\right|$, $k' = \left|\frac{r'_2 - r'_1}{r'_2 + r'_1}\right|$. The induced velocities of a pair of vortex rings at any point M in space are:

$$\begin{cases} W_x = \left(\frac{x_M - x_P}{{r_M}^2}\right)\frac{\partial \psi}{\partial z_M} \\ W_y = \left(\frac{y_M - y_P}{{r_M}^2}\right)\frac{\partial \psi}{\partial z_M} \\ W_z = \left(-\frac{1}{r_M}\right)\frac{\partial \psi}{\partial r_M} \end{cases} \tag{8}$$

where r_M is the distance between point M and the center axis of the two vortex rings.

If M is at the central axis of the vortex ring, $r_M = 0$. Therefore, it cannot be solved using the third equation in Formula (8). Derive the induced velocity at the central axis of the vortex ring by introducing the potential function of the vortex ring:

$$W_z(z) = \frac{\Gamma}{2R} \frac{1}{\left(1 + \left(\frac{z}{R}\right)^2\right)^{2/3}} \tag{9}$$

where z is the distance between M and O_P. When $z = 0$, there is $W_z(0) = \frac{\Gamma}{2R}$, so that $\Gamma = 2RW_z(0)$. That is to say, the strength of the vortex ring is determined by the vertical induced velocity at the center of the vortex ring and the radius of the vortex ring.

2.2. RL Algorithm with Active Exploration Reward Function

2.2.1. Design of the Active Reward Function

To ensure full-envelope control capability, a comprehensive state-action space is defined. The state space is defined as $\mathbf{S}_t = [V, h, \theta, \psi, \phi]^T \in \mathbb{R}^5$, representing airspeed, altitude, pitch angle, yaw angle, and roll angle, respectively. This allows the agent to perceive both longitudinal and lateral-directional flight dynamics. The action space is defined as $\mathbf{A}_t = [\delta_{th}, \delta_e, \delta_r, \delta_a]^T \in \mathbb{R}^4$, comprising the throttle lever, elevator deflection, rudder deflection, and aileron deflection.

To address the inherent sparse reward issues in low-altitude windshear encounters, the active reward function with proximal policy optimization (AR-PPO) framework is proposed. The training procedure is formulated as a nested optimization problem, which dynamically transforms sparse environmental feedback into dense, instructive reward signals by decoupling the learning process into two interconnected stages, that is, a lower-level policy optimization and an upper-level reward parameter update.

The lower-level task aims to derive the optimal flight policy π_θ by maximizing the expected return of a parameterized, dense active reward function $R_\eta(\mathbf{S}, \mathbf{A})$. The cumulative objective function for a trajectory τ is defined as:

$$J_{lower}(\theta, \eta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \gamma^t R_\eta(\mathbf{S}_t, \mathbf{A}_t)\right] \tag{10}$$

where η denotes the learnable parameters of the reward function, designed to provide high-frequency, step-by-step guidance. During this phase, the policy network parameters θ are updated via the standard PPO clipping mechanism, ensuring stable convergence under the local guidance of R_η.

Concurrently, the upper-level optimization focuses on adjusting the reward parameters η. The objective is to ensure that the resultant policy $\pi_{\theta^*(\eta)}$ maximizes a global meta-objective J_{meta}:

$$\eta^* = \arg\max_\eta J_{meta}(\pi_{\theta^*(\eta)}) \tag{11}$$

In contrast to the parameterized dense reward used in the lower level, J_{meta} represents the sparse, delayed environmental feedback reflecting the ultimate flight safety criteria, such as overall survival rate and minimum altitude threshold.

To bridge the sparse meta-objective and the dense local rewards, the parameters η are iteratively updated via meta-gradients. Since the optimal policy parameters θ^* inherently depend on η, the gradient of the meta-objective with respect to η is derived using the chain rule:

$$\nabla_\eta J_{meta}(\eta) = \nabla_\theta J_{meta}(\theta) \cdot \nabla_\eta \theta(\eta) \tag{12}$$

By approximating this meta-gradient through the training trajectories, the AR-PPO framework autonomously discovers how to map sparse survival outcomes into dense aerodynamic guidance. Specifically, through the meta-gradient mechanism, the overarching discrete objective of survival is mathematically distilled into continuous, state-action-level evaluations, meaning the agent deduces the underlying physical laws, such as the absolute necessity of preserving kinetic energy, directly from the ultimate success or failure of its exploratory flights. Consequently, this empowers the agent to autonomously navigate critical physical trade-offs, such as prioritizing airspeed over altitude during vortex core penetration, without relying on handcrafted reward shaping.

2.2.2. Design of the Manual Reward Function

To evaluate the efficacy of the proposed active reward mechanism, a manual, static reward function is designed for the proximal policy optimization (PPO). This function is formulated based on classical flight control objectives during a windshear encounter, which primarily involve minimizing the deviation between the current flight states and the target go-around states.

At each time step t, the manual reward $R_{manual}^{(t)}$ is defined as a weighted sum of the multi-channel tracking errors and boundary penalties:

$$R_{manual}^{(t)} = R_{alive} - \left(\lambda_V e_V^{(t)} + \lambda_h e_h^{(t)} + \lambda_\theta e_\theta^{(t)} + \lambda_\phi e_\phi^{(t)} + \lambda_\psi e_\psi^{(t)}\right) - P_{violation} \tag{13}$$

where R_{alive} represents a constant positive reward given at each time step when the aircraft remains in the air and avoids stalling, encouraging the agent to prolong the flight duration. During a go-around, kinetic energy and potential energy must be strictly protected. The error terms are defined using the Huber loss or absolute error to prevent exploding gradients when deviations are massive:

$$e_h^{(t)} = \max(0, h_{safe} - h^{(t)})^2 \tag{14}$$

Airspeed is penalized only when it drops below the safe approach speed ($V_{ref} = 70$ m/s), and altitude is penalized only when it drops below the critical decision height (h_{safe}). The attitude channels must be regulated to the standard go-around posture; that means pitch $\theta \approx 15^\circ$, roll $\phi \approx 0^\circ$ and yaw $\psi \approx 0^\circ$. The tracking errors are defined as:

$$\begin{cases} e_\theta^{(t)} = |\theta^{(t)} - \theta_{target}| \\ e_\phi^{(t)} = |\phi^{(t)}| \\ e_\psi^{(t)} = |\psi^{(t)}| \end{cases} \tag{15}$$

A massive sparse penalty is triggered if the aircraft crashes into the ground or exceeds the critical stall angle of attack. This immediately terminates the episode.

$$P_{violation} = \begin{cases} 1000, & \text{if } h^{(t)} \leq 0 \text{ or } \alpha^{(t)} \geq 15^\circ \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

2.3. Multivariable Decoupling and Signal Reconstruction via W-MSSA

2.3.1. Denoising of Control Signals via Discrete Wavelet Transform

The control commands generated by the reinforcement learning agent, inherently contain high-frequency oscillations. These oscillations stem from the stochastic nature of the active exploration process and can cause physical damage to aircraft actuators, such as servo motors and hydraulic systems. To extract the deterministic control intent while

removing exploratory noise, the discrete wavelet transform (DWT) is employed based on multi-resolution analysis (MRA).

For a raw action signal $x(t)$, the DWT decomposes the sequence into a combination of a low-frequency approximation and high-frequency details. Using a scaling function $\phi(t)$ and a mother wavelet $\psi(t)$, the signal $x(t)$ at a maximum decomposition level J is expressed as:

$$x(t) = \sum_{k\in\mathbb{Z}} c_{J,k}\phi_{J,k}(t) + \sum_{j=1}^{J}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t) \tag{17}$$

where $c_{J,k}$ represents the approximation coefficients at level J, which capture the smooth, macro-level control intent (e.g., steady thrust increase or constant pitch maintenance). $d_{j,k}$ represents the detail coefficients at level j, which represent the localized high-frequency fluctuations caused by exploration noise. $\phi_{J,k}(t)$ and $\psi_{j,k}(t)$ represent the orthogonal basis functions generated by translating and scaling the father and mother wavelets.

To achieve precise denoising, a soft-thresholding operator is applied to the detail coefficients $d_{j,k}$. An adaptive threshold λ_j is calculated for each level to separate noise from useful control features. The modified coefficients $\tilde{d}_{j,k}$ are obtained by:

$$\tilde{d}_{j,k} = \mathrm{sgn}(d_{j,k})\max(|d_{j,k}| - \lambda_j, 0) \tag{18}$$

By applying the inverse discrete wavelet transform (IDWT) using the original $c_{J,k}$ and the processed $\tilde{d}_{j,k}$, we reconstruct the denoised control sequence $\tilde{x}(t)$. For the practical implementation in this study, the Daubechies 4 (db4) mother wavelet was selected due to its excellent orthogonality and compact support, which effectively preserves the transient aerodynamic intent. The maximum decomposition level J was set to 3. The adaptive threshold λ_j was determined using the rigorous Universal Thresholding method, $\lambda_j = \sigma_j\sqrt{2\ln(N)}$, where σ_j is the median absolute deviation of the detail coefficients, ensuring a mathematically robust separation of exploratory noise from deterministic control laws. This procedure effectively mitigates exploratory chattering while preserving the high-fidelity transient responses required for windshear escape. The resulting high-SNR (signal-to-noise ratio) signals serve as the input for the subsequent multivariable decoupling.

2.3.2. Multivariable Singular Spectrum Analysis

Subsequently, multivariate singular spectrum analysis (MSSA) is utilized to resolve the spatiotemporal coupling between the state and action variables. The classical MSSA algorithm comprises four rigorous mathematical steps: embedding, singular value decomposition (SVD), grouping, and diagonal averaging.

Construct a multi-channel trajectory matrix by stacking the lagged trajectory matrices of all $M = 9$ variables (5 states and 4 actions). To ensure consistency in the multivariable singular spectrum analysis, the action sequence is standardized to a 40-s total duration, expanded to exactly $N = 400$ data points to maintain a high-fidelity 10 Hz sampling rate. To adequately capture the aerodynamic modalities across these 400 data points without over-smoothing the critical response features, the MSSA window length L is empirically set to 40, representing a 4-s temporal moving window. Letting $K = N - L + 1$, the augmented trajectory matrix $X \in \mathbb{R}^{ML\times K}$ is formulated as:

$$\mathbf{X} = [\mathbf{X}_V^T, \mathbf{X}_h^T, \mathbf{X}_\theta^T, \mathbf{X}_\psi^T, \mathbf{X}_\phi^T, \mathbf{X}_{\delta_{th}}^T, \mathbf{X}_{\delta_e}^T, \mathbf{X}_{\delta_r}^T, \mathbf{X}_{\delta_a}^T]^T \tag{19}$$

To capture the cross-channel correlations (e.g., the aerodynamic coupling between the aileron and elevator during lateral-longitudinal leveling), we compute the block-Toeplitz cross-covariance matrix $\mathbf{C} = \mathbf{X}\mathbf{X}^T \in \mathbb{R}^{ML\times ML}$. By performing eigendecomposition on $\mathbf{C}$,

we obtain the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{ML} \geq 0$ and their corresponding eigenvectors $\mathbf{U}_i$. The i-th principal component $\mathbf{V}_i$ is computed as $\mathbf{V}_i = \mathbf{X}^T\mathbf{U}_i/\sqrt{\lambda_i}$. The matrix $\mathbf{X}$ is then decomposed into a sum of elementary rank-one bi-orthogonal matrices:

$$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i}\mathbf{U}_i\mathbf{V}_i^T = \sum_{i=1}^{d} \mathbf{X}^{(i)} \tag{20}$$

where d is the rank of $\mathbf{X}$.

The reinforcement learning outputs contain both the primary control intent and high-frequency exploratory perturbations. Partition the index set $\{1, 2, \ldots, d\}$ into P disjoint subsets $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_p$. By selecting the subset $\mathbf{I}_{main}$ corresponding to the largest eigenvalues (which represent the dominant energy of the deterministic physical operations), we isolate the core flight control modalities from the aerodynamic noise. The approximated matrix $\widehat{\mathbf{X}}$ is thus given by $\widehat{\mathbf{X}} = \sum_{i \in \mathbf{I}_{main}} \mathbf{X}^{(i)}$.

The final step transforms the approximated matrix $\widehat{\mathbf{X}}$ back into a set of 1D multi-variable time series. For the m-th variable block in $\widehat{\mathbf{X}}$, denoted as $\mathbf{Y}^{(m)} \in \mathbb{R}^{L \times K}$ with elements $y_{i,j}$, the reconstructed one-dimensional sequence $\tilde{x}_m(n)$ at time index n ($1 \leq n \leq N$) is obtained by averaging the elements along the anti-diagonals:

$$\tilde{x}_m(n) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} y^*_{i,n-i+1} & \text{for } 1 \leq n < L^* \\ \frac{1}{L^*}\sum_{i=1}^{L^*} y^*_{i,n-i+1} & \text{for } L^* \leq n \leq K^* \\ \frac{1}{N-n+1}\sum_{i=n-K^*+1}^{N-K^*+1} y^*_{i,n-i+1} & \text{for } K^* < n \leq N \end{cases} \tag{21}$$

where $L^* = \min(L, K)$ and $K^* = \max(L, K)$, and $y^*_{i,j} = y_{i,j}$ if $L \leq K$; otherwise, $y^*_{i,j} = y_{j,i}$.

Through this rigorous W-MSSA procedure, the entangled high-dimensional DRL actions are successfully decoupled and reconstructed into pure, physically executable pilot commands.

2.4. Rule Extraction and Standard Operating Procedure Generation

Although the decoupled control signals have been filtered of high-frequency exploratory noise, they fundamentally remain continuous numerical sequences. To extract discrete rules with practical engineering value, the classification and regression tree (CART) algorithm is introduced for policy extraction [35]. This algorithm utilizes the instantaneous flight state vector $\mathbf{S}_t$ as input features and the discretized synergistic control modes as output labels.

The CART algorithm recursively partitions the state space by minimizing the Gini Impurity:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{22}$$

where D represents the flight state dataset at the current split node, K is the total number of discrete physical control modes, and p_k is the proportion of samples belonging to the k-th mode within the dataset. To ensure the conciseness and human expert readability of the extracted rules, pre-pruning techniques are employed to strictly restrict the maximum depth of the decision tree, thereby effectively preventing the model from overfitting to local high-frequency state features.

By traversing and parsing the fully trained shallow decision tree, the decision paths from the root node to the respective leaf nodes can be directly converted into structured "IF–THEN" boolean logic rules. For instance: "IF the altitude descent rate exceeds the safety threshold AND the roll angle is excessive, THEN command maximum throttle response δ_{th} AND apply opposite aileron δ_a to maintain pitch and lateral balance." These data-driven,

systematically generated physical rules mitigate the black-box limitations faced by deep reinforcement learning in the aviation control domain.

The overall training and distillation process of the proposed approach is summarized in Algorithm 1.

Algorithm 1 Explainable Reinforcement Learning Framework for Autonomous Windshear Escape

Require: Micro downburst simulation environment E, maximum iterations M
Ensure: Executable "IF-THEN" Standard Operating Procedures (SOPs)

Phase 1: AR-PPO Training (Actor-Critic Architecture)

Initialize **Actor-Critic** network parameters θ and active reward parameters η
for iteration $i = 1, 2, \ldots, M$ **do**
// Lower-Level Optimization: Policy Update
Collect transition data (S_t, A_t, S_{t+1}) using the Actor network in environment E
Compute dense active reward $R_\eta(S_t, A_t)$ using current parameters η
Update **Actor-Critic** parameters θ by maximizing lower-level objective via PPO clipping
// Upper-Level Optimization: Addressing the Sparse Reward Problem
Evaluate the updated policy to collect sparse episodic outcomes (e.g., survival/crash)
Calculate the meta-gradient based on ultimate flight safety criteria
Update reward parameters η via meta-gradients to autonomously discover optimal rewards
end for
Obtain the trained optimal policy π_{θ^*} and generate raw action sequence set $\{A\}$

Phase 2: Multivariable Decoupling and Signal Reconstruction (W-MSSA)

for each raw action sequence $x(t) \in \{A\}$ **do**
Apply Discrete Wavelet Transform (DWT) to filter high-frequency exploratory noise
Construct augmented trajectory matrix X for states and denoised actions
Compute cross-covariance matrix and perform Singular Value Decomposition (SVD)
Extract principal components representing the deterministic dynamic energy
Reconstruct the decoupled, purely aerodynamic control sequence $\{\tilde{A}\}$ via diagonal averaging
end for

Phase 3: Rule Extraction and SOP Generation

Construct the expert dataset D using flight states S and the reconstructed actions $\{\tilde{A}\}$
Apply **Sequential Pattern Mining** to dataset D to discover temporal synergistic control modes
Translate the mined sequential patterns into discrete, pilot-readable decision paths
return Human-readable "IF-THEN" Standard Operating Procedures (SOPs)

3. Results

3.1. Simulation Setup and Parameter Settings

The simulation experiments are conducted on a high-fidelity six-degree-of-freedom (6-DOF) nonlinear flight dynamics model of B737-800. To simplify the experiment while ensuring the validity of the proposed framework, this study focuses solely on scenarios where the aircraft traverses the wind field from the right side under varying wind intensities. This setup is sufficient because the control performance is identical whether crossing from the left or right, and flying directly through the middle is a relatively simpler task. During the training phase, the maximum downdraft velocity of the micro downburst core is configured to 15.0 m/s. To test the zero-shot adaptation and robustness of the learned policy, an unencountered, extreme micro downburst with a 22.0 m/s downdraft

is introduced during the testing phase. The detailed physical constraints, control action bounds, and algorithm hyperparameters are summarized in Table 1.

Table 1. Simulation environment and hyperparameter configurations.

Category	Parameter	Value
Aircraft and environment	Initial airspeed V_0 (m/s)	75.0
	Altitude h_0 (m)	500
	Micro downburst max downdraft $W_{z,max}$ (m/s)	15.0 (Train)/22.0 (Test)
	Throttle δ_{th}	[0.1, 1.0]
	Elevator δ_e (°)	$[-15.0, 15.0]/15.0$
	Rudder/Aileron δ_r, δ_a (°)	$[-20.0, 20.0]/20.0$
Active reward function parameters	Temporal discount factor γ	0.99
	Meta-objective terminal penalty R_{crash}	-100
	Meta-Objective: Terminal Reward $R_{success}$	50
Manual reward function parameters	R_{alive}	1.0
	λ_V	0.5
	λ_h	0.3
	λ_θ	0.1
	λ_ϕ	0.1
	λ_ψ	0.1
	$P_{violation}$	1000

It is crucial to note that to prevent the RL agent from exploiting unrealistic control frequencies, the simulation incorporates realistic actuator dynamics. For instance, the throttle commands are subjected to a first-order lag with a time constant of 2.5 s to simulate the spool-up delay of commercial turbofan engines. Similarly, the control surfaces are constrained by both position limits (e.g., $\delta_e \in [-15^\circ, 15^\circ]$) and mechanical rate limiters (e.g., maximum deflection rate of $15^\circ/\text{s}$).

To ensure a fair and rigorous ablation study, the static weights of the manual reward function for the standard PPO baseline were carefully calibrated through grid search to provide a strong, non-trivial benchmark.

As shown in Table 1, the longitudinal energy tracking weights λ_V and λ_h are assigned significantly higher values than the attitude tracking weights λ_θ, λ_ϕ, λ_ψ. Specifically, the airspeed penalty weight $\lambda_V = 0.5$ is deliberately set higher than the altitude penalty weight $\lambda_h = 0.3$. This static configuration reflects the fundamental aerodynamic heuristic that, during a severe micro downburst encounter, maintaining airspeed above the stall margin is more critical than strictly tracking the altitude reference.

To ensure the reproducibility of the proposed framework, all simulation experiments and reinforcement learning training were conducted on a high-performance workstation. The specific hardware configuration includes an Intel Core i9-13900K processor, an NVIDIA GeForce RTX 4090 GPU (24GB VRAM), and 64GB DDR5 RAM. The software environment is based on Ubuntu 22.04 LTS, utilizing Python 3.9 and the PyTorch 2.0 deep learning framework. Furthermore, the initial flight states for the simulation were configured to represent a standard landing approach prior to the micro downburst encounter.

Regarding hyperparameter sensitivity, the proposed framework demonstrates significant robustness compared to traditional baseline algorithms. Specifically, built entirely upon an Actor–Critic architecture, the active reward discovery mechanism was implemented explicitly to solve the sparse reward problem inherent in extreme low-altitude windshear

encounters. By autonomously discovering and dynamically adjusting the reward mapping via meta-gradients during training, this bilevel architecture intrinsically reduces the agent's sensitivity to the initial static reward weight configurations. This eliminates the necessity for exhaustive, manual hyperparameter tuning of reward functions, ensuring high practical reproducibility and generalization.

3.2. Experimental Results

3.2.1. Comparison in Nominal Windshear Environment

To validate the superiority of the proposed active reward updating mechanism, Figure 4 presents the episodic learning curves of the proposed AR-PPO agent and the baseline MR-PPO agent over 2000 training episodes. Both policies are initialized under identical micro downburst environments and neural network architectures. The solid lines represent the mean episodic return, while the shaded regions denote the standard deviation over multiple random seeds, reflecting the stability of the learning process.

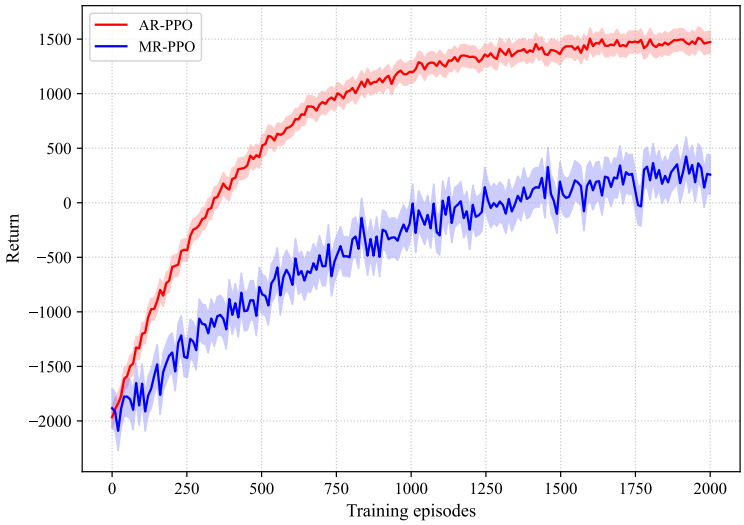


Figure 4. Convergence comparison between AR-PPO and MR-PPO.

As shown in Figure 4, the proposed AR-PPO algorithm (red line) demonstrates a rapid and stable convergence trajectory. During the early exploration phase, the AR-PPO agent quickly escapes the severe penalty region, starting near -2000. By adaptively adjusting the reward weights, the AR-PPO policy could robustly converge to a high-reward global optimum after roughly 1000 training episodes, with a conspicuously narrow shaded band indicating minimal policy variance and high training stability.

In contrast, the MR-PPO agent (blue line) exhibits sluggish learning dynamics and severe policy chattering. The static, manually tuned reward function fails to provide a consistent direction for gradient ascent during complex aerodynamic variations. Consequently, the MR-PPO agent ultimately plateaus at a profoundly suboptimal level of 500. Furthermore, the wide variance associated with the MR-PPO curve reveals that the static reward configuration causes the agent to frequently fall into local optima, repeatedly oscillating between conflicting sub-tasks without finding a globally safe escape trajectory.

Figure 5 illustrates the time-history evolution of the five multidimensional flight states, airspeed V, altitude h, pitch angle θ, roll angle ϕ and yaw angle ψ, under the nominal micro downburst encounter with $W_{z,max} = 15.0$ m/s. The shaded region from 10 s to 25 s denotes the area of the wind field, characterized by intense downdrafts and crosswinds. The state trajectories provide a definitive physical explanation for why the MR-PPO fails and the proposed AR-PPO succeeds.

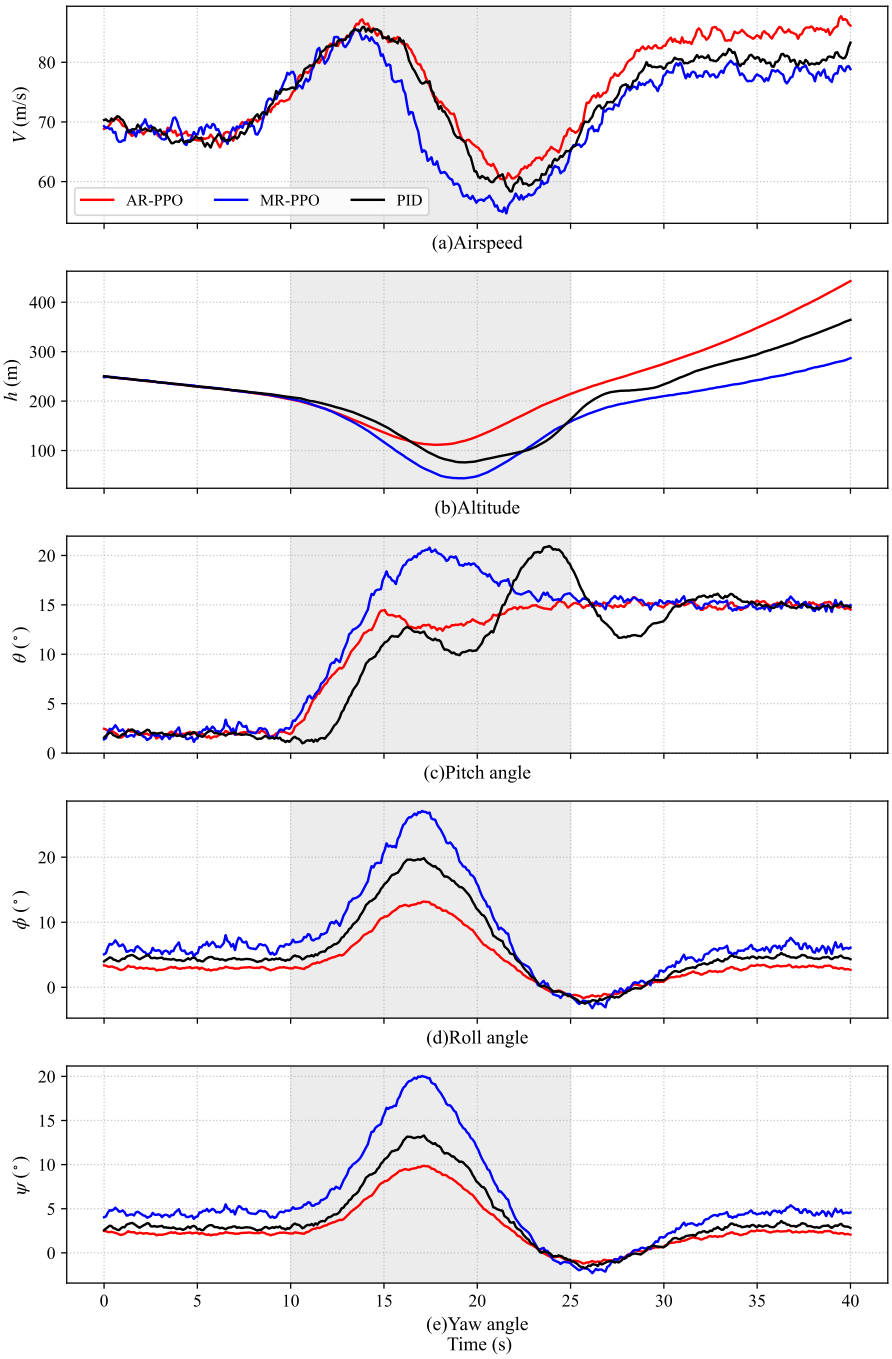


Figure 5. Evolution of flight states with different algorithms.

When encountering the severe altitude drop at $t = 10$ s, the MR-PPO agent adopts a naive, greedy strategy to minimize the altitude penalty. It aggressively pulls the stick, causing the pitch angle to overshoot dangerously above 20° around $t = 17$ s, as shown in Figure 5c. This severe aerodynamic violation triggers a massive depletion of kinetic energy, plummeting the airspeed below 56 m/s as shown in Figure 5a. Consequently, the aircraft loses vital lift, resulting in the deepest and most perilous altitude drop, bottoming out near 50 m in Figure 5b. This behavior characterizes the classic stall trap, sacrificing critical airspeed for immediate pitch, which inevitably leads to a worse altitude loss.

Guided by the active reward updating, it smoothly establishes and maintains the 13°∼15° target pitch angle without overshooting. By carefully trading altitude for airspeed, it prevents a stall, maintaining the highest minimum airspeed (≈60 m/s) inside the vor-

tex core. This physical foresight allows the proposed agent to execute the safest altitude recovery. The conventional PID controller (black line), while safer than MR-PPO, exhibits a typical integral phase lag in establishing the pitch attitude (peaking only after $t = 20$ s), significantly delaying the overall altitude recovery. Furthermore, a direct comparison between the proposed algorithm and the conventional PID controller in Figures 4 and 5 highlights the fundamental advantage of the proposed framework. While the PID controller is purely reactive, suffering from a severe integral phase lag that delays critical energy recovery and pitch establishment, the proposed agent acts with instantaneous precision.

To trace the root causes of the flight state evolutions and the algorithmic convergence behaviors, Figure 6 illustrates the continuous multivariable control commands generated by the three strategies. The subplots detail the throttle δ_{th}, elevator δ_e, aileron δ_a, and rudder δ_r inputs. The analysis of these actuator dynamics provides a definitive physical explanation for the observed stall trap and lateral instabilities.

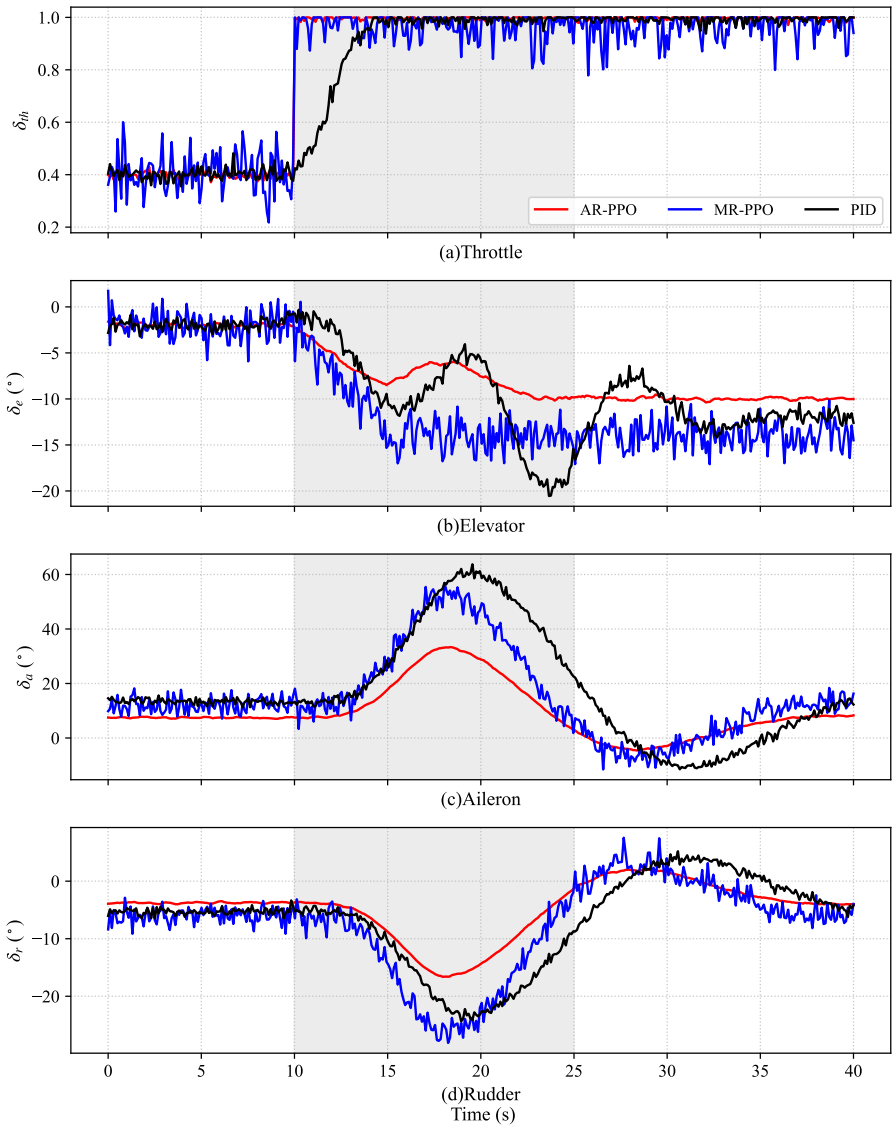


Figure 6. Evolution of actions with different algorithms.

The top two states δ_{th} and δ_e completely demystify the altitude loss and airspeed depletion seen in the MR-PPO. Upon windshear detection at $t = 10$ s, both the proposed AR-PPO and MR-PPO agents correctly command immediate maximum take-off/go-around

(TOGA) thrust ($\delta_{th} \rightarrow 1.0$). In contrast, the conventional PID controller exhibits a dangerous integral lag, severely delaying the energy recovery process.

Driven by the static manual reward to aggressively minimize altitude loss, the MR-PPO agent executes a violent, saturated pull-up maneuver of $\delta_e < -15^\circ$. This exact action causes the pitch overshoot of exceeding 20° observed in Figure 5c, triggering the kinetic energy depletion. Furthermore, the extreme high-frequency chattering in the MR-PPO's elevator command directly accounts for the massive variance and instability shown in the Figure 4 learning curve.

Conversely, the proposed AR-PPO agent executes a smooth, measured elevator deflection of approximately -6° to -10°, perfectly balancing the trade-off between pitch establishment and airspeed preservation, yielding the optimal recovery trajectory.

The bottom two subplots (δ_a and δ_r) validate the realism of the simulation environment and the robustness of the proposed framework. When penetrating the highly asymmetric vortex core ($15\text{ s} < t < 25\text{ s}$), the standard PPO and PID controllers exhibit massive control saturation and severe oscillations, commanding aileron deflections approaching 60°. This aggressive over-control directly induces the hazardous roll and yaw deviations seen in Figure 5d,e.

The control action analysis conclusively proves that the proposed active-reward mechanism effectively suppresses the high-frequency policy chattering typical of standard DRL algorithms. By optimizing the control surface deflections physically, it inherently avoids the stall trap and lateral divergence, leading to the highly stable algorithmic convergence presented in Figure 4.

3.2.2. Robustness Analysis

To ensure the multivariable physical dynamics are highly interpretable without compromising the raw high-fidelity data, the evaluation provides detailed textual signposts for the state and action trajectories. In the nominal micro downburst encounter, the multivariable aircraft states (Figure 5) visually capture the critical airspeed-altitude trade-off during the vortex core penetration phase (10 s < t < 25 s). By tracing these state trajectories, readers can clearly observe the AR-PPO agent intentionally relaxing the pitch angle to preserve kinetic energy (airspeed margin) at the expense of a controlled potential energy (altitude) drop. Correspondingly, Figure 6 details the specific continuous control actions executing this behavior, explicitly demonstrating how the learned policy actively avoids the aggressive, saturation-inducing actuator commands typical of the static-reward baseline.

Furthermore, this robust physical logic is validated across broader operational envelopes. As presented in the multivariable state comparison under varying windshear intensities (Figure 7), it becomes increasingly evident that the agent consistently applies this energy trade-off strategy across different severity levels. The synchronized analysis of these three figures provides a transparent visualization of how the agent circumvents the conventional stall trap, rendering the extreme recovery trajectories physically verifiable and easily interpretable.

The ultimate validation of the proposed method lies in the severe, unseen windshear scenario from Figure 7k–o. As established in the nominal analysis in Figures 5 and 6, the MR-PPO agent is inherently flawed by its static reward, which blindly prioritizes altitude preservation through aggressive elevator deflection. Under the extreme downdraft of the severe Micro downburst, this flaw is fatally magnified. The MR-PPO agent wildly pulls the pitch angle beyond 20°, as shoen in Figure 7m, instantly stalling the aircraft. Consequently, the airspeed plummets, and the aircraft completely loses aerodynamic lift, resulting in a catastrophic crash.

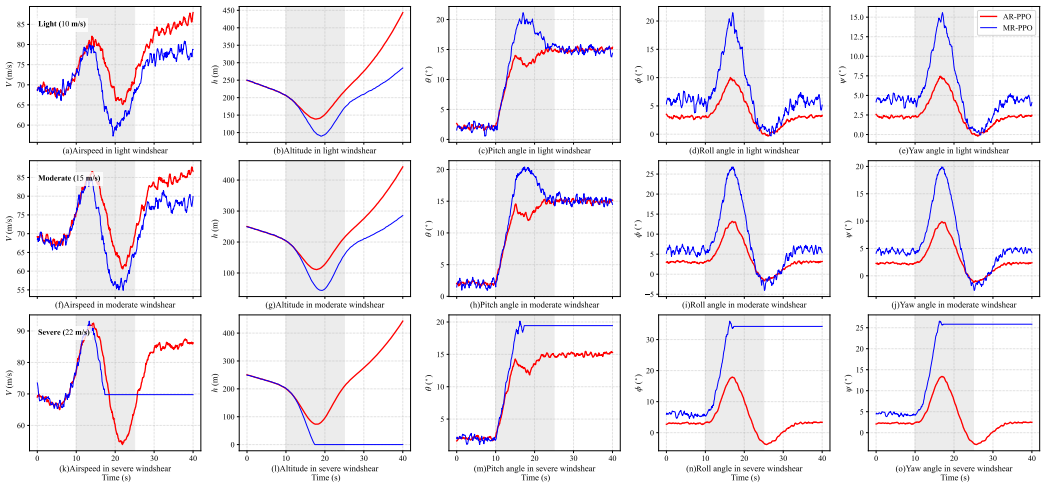


Figure 7. Robustness and generalization comparison under varying windshear intensities.

Conversely, the AR-PPO agent successfully survives this zero-shot extreme test. Relying on its dynamically updated reward weights, it decisively sacrifices some altitude, dropping to approximately 70 m in Figure 7l, to maintain a sub-critical pitch angle seen in Figure 7m and preserve kinetic energy. By keeping the airspeed above the critical stall threshold seen in Figure 7k, the AR-PPO agent secures enough lift to execute a breathtaking recovery.

The robustness disparity is equally striking in the lateral-directional channels. As the intensity of the asymmetric vortex increases from light to severe, the MR-PPO agent loses lateral control authority. In the severe scenario of Figure 7n,o, the MR-PPO agent experiences severe lateral divergence, with roll angles exceeding 30° and yaw angles exceeding 25°, indicating a complete departure from controlled flight before crashing.

The AR-PPO agent, however, demonstrates profound decoupling capabilities. Even under the extreme 22 m/s asymmetric crosswind, it tightly bounds the roll and yaw oscillations, keeping peak roll under 18°, safely maintaining the non-zero DC bias required for the wing-low approach technique.

While Figures 4–7 established the algorithmic superiority and zero-shot robustness of the AR-PPO agent against mathematical baselines, Figure 8 bridges the simulation-to-reality gap. It presents a high-fidelity comparative validation between the autonomous AR-PPO agent and empirical flight data recorder (FDR) data from a human pilot facing an identical micro downburst encounter.

The most striking difference between the AI and the human pilot is observed in the pitch angle response seen in Figure 8c. When the windshear warning is triggered ($t = 10$ s), the human pilot exhibits an inherent cognitive and physiological reaction delay of approximately 1 to 1.5 s. Once initiated, the human pilot's pitch control is characterized by a stepped, discrete pull-up followed by severe high-frequency oscillations, ranging from 12° to 18°. This phenomenon, known in aviation as pilot-induced oscillation (PIO), occurs because the human pilot struggles to perfectly synchronize the elevator inputs with the highly unsteady aerodynamic forces of the vortex core.

The human pilot's delayed reaction and oscillatory pitch control directly translate into a deeper and more dangerous altitude loss. As shown in Figure 8b, the human pilot's trajectory sags to a minimum altitude of approximately 75 m. The AR-PPO agent, utilizing its predictive active-reward mechanism, optimally trades energy without hesitation, securing a minimum altitude of approximately 100 m. By reducing an additional 25 m of altitude clearance compared to the human pilot, the AI agent provides a significantly wider safety margin against controlled flight into terrain (CFIT).

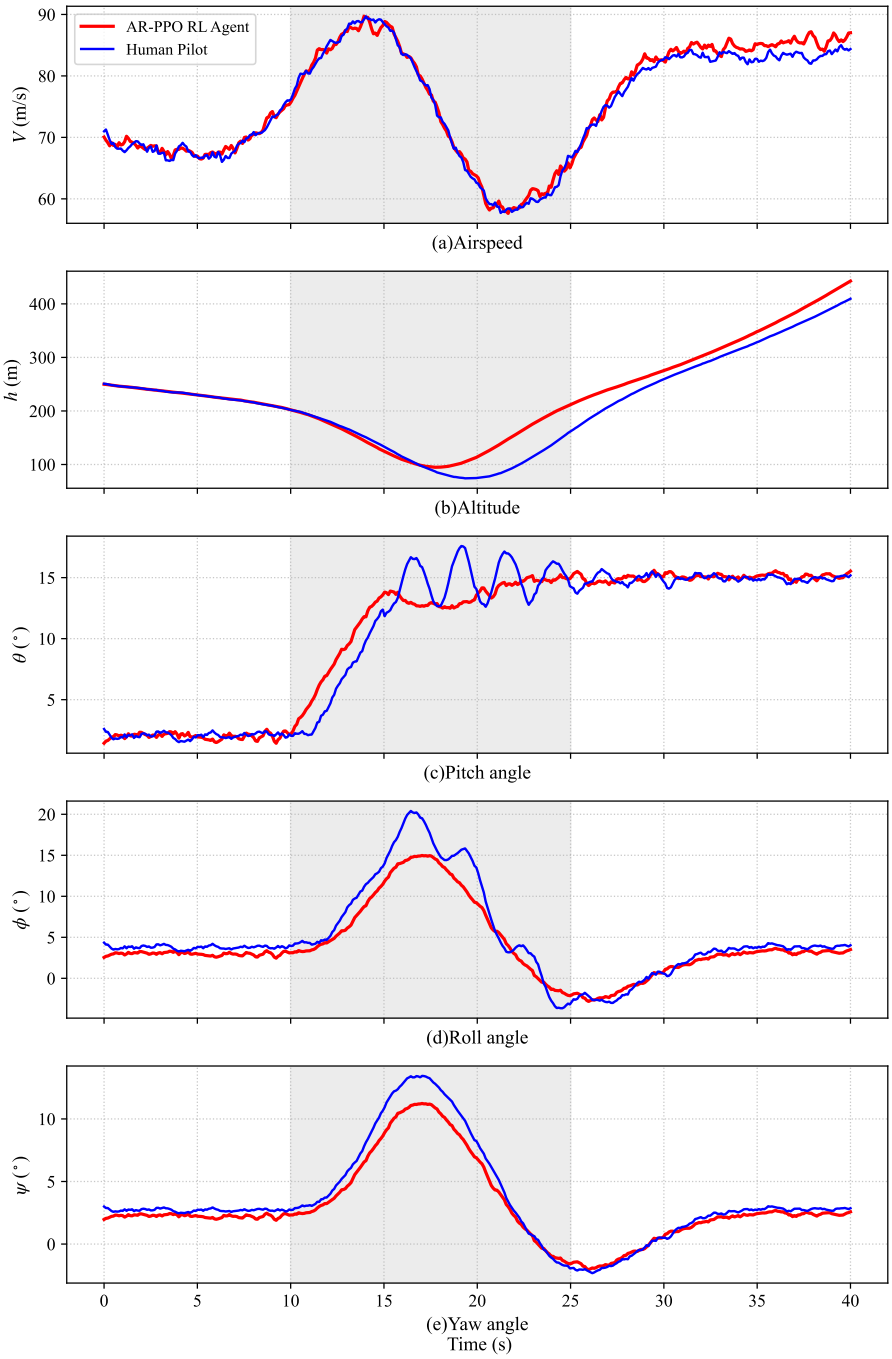


Figure 8. Flight simulator validation comparing the multidimensional state trajectories of the proposed AR-PPO agent against empirical human pilot data.

Conversely, the AR-PPO agent acts with instantaneous precision. Benefiting from the smooth, continuous control policy (as previously verified in Figure 6), the agent establishes the target 15° go-around pitch angle earlier, faster, and with near-zero overshoot or oscillation.

Under extreme windshear, a human pilot faces an overwhelming cognitive load: they must simultaneously push the throttle, pull the stick, and fight the asymmetric crosswind using ailerons and rudders. This split attention inevitably degrades lateral-directional precision. As evident in Figure 8d,e, the human pilot suffers from larger transient deviations in both roll (exceeding 20°) and yaw, exceeding 12.5°. The AR-PPO agent, inherently capable of processing high-dimensional multi-axis inputs simultaneously, executes perfectly

decoupled control. It smoothly rejects the lateral gusts, bounding the peak roll and yaw deviations significantly tighter than the human pilot.

Connecting these empirical findings with the robustness matrix, it is evident that human pilots, much like static reward functions, are susceptible to delays, over-control, and task-saturation during extreme weather anomalies. The proposed AR-PPO framework transcends these limitations. By actively exploring and adapting to the dynamic environment, it achieves a level of anticipatory, decoupled, and optimally bounded flight control that surpasses standard empirical human performance.

To uncover the mechanical root causes behind the state trajectory discrepancies observed in Figure 8, Figure 9 details the continuous multivariable control commands, throttle, elevator, aileron, and rudder, executed by both the AR-PPO agent and the empirical human pilot during the identical simulator test. This microscopic view perfectly illustrates the physiological boundaries of human manual flight compared to the optimal precision of the proposed AI framework.

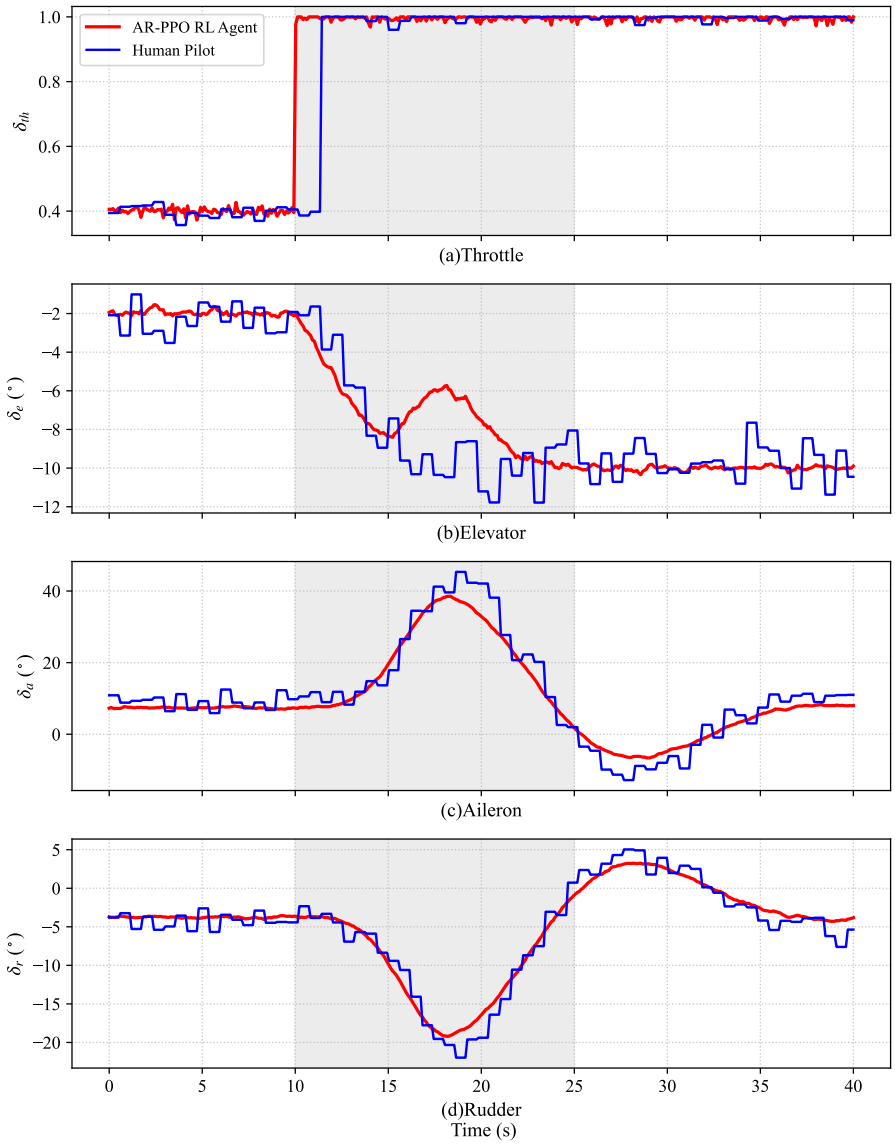


Figure 9. Actuator-level comparison between the proposed AR-PPO agent and the empirical human pilot.

The human pilot's inputs exhibit a distinct staircase or zero-order hold characteristic. This reflects the human cognitive loop of perceiving instruments, processing information,

making a decision, and applying manual force to the yoke/rudder pedals. This biological processing inherently results in discrete, stepped control actions. Furthermore, Figure 9a explicitly captures the human reaction latency, while the AR-PPO agent commands full TOGA thrust instantly at the 10-s mark, the human pilot's throttle input is delayed by a full second. This critical 1-s delay is the primary catalyst for the deeper altitude loss observed in the human pilot's trajectory in Figure 8b.

Figure 9b reveals the physical origin of the human pilot's PIO seen previously in Figure 8c. Following the initial delay, the panicked human pilot over-compensates by pulling the elevator past -10° in discrete, aggressive steps. In contrast, the AR-PPO agent executes a smooth, continuous elevator curve that peaks gently around -8°. More importantly, demonstrating its predictive active-reward mechanism, the AR-PPO agent smoothly relaxes the elevator, moving towards -6°, directly inside the vortex core (15–20 s) to trade altitude for airspeed, completely avoiding the human pilot's over-control.

Figure 9c,d demonstrate control behavior under severe asymmetric crosswinds. The human pilot's stepped inputs in the aileron and rudder highlight the immense cognitive load required to manually coordinate lateral leveling while simultaneously fighting a longitudinal Micro downburst. The human inputs overshoot significantly, aileron exceeding 40°. The AR-PPO agent, unburdened by human cognitive bandwidth limits, processes the high-dimensional state space instantaneously. It commands perfectly smooth, perfectly decoupled lateral corrections, keeping the aileron tightly bounded below 40° while maintaining the required DC bias for the crab/wing-low approach.

Through a rigorous progression of algorithmic ablation, nominal kinematic tracking, zero-shot severe weather stress-testing and empirical flight simulator validation, this study conclusively demonstrates the superiority of the proposed AR-PPO framework. By replacing static penalty mechanics with an active-exploration reward strategy, the proposed AI agent effectively transcends both the mathematical limitations of conventional DRL baselines and the physiological limitations of human pilots, offering a robust, life-saving autonomous control solution for extreme low-altitude windshear encounters.

3.3. Design for Physical SOP Extraction

To overcome the inherent black-box nature of DRL and meet the stringent explainability and safety requirements for airworthiness certification, a post-training policy distillation pipeline is designed. The experimental objective is to translate the high-frequency, continuous neural network outputs into human-readable, rule-based SOPs. The experiment is structured into three sequential phases.

First, an ensemble of successful recovery trajectories is generated by the well-trained AR-PPO agent under varying Micro downburst intensities. Then, DRL policies intrinsically exhibit high-frequency stochastic chattering which are mechanically detrimental to actuators and physiologically impossible for human pilots to execute. Next, the Wavelet transform provides multi-resolution noise filtering, while MSSA extracts the principal deterministic dynamic components, isolating the pure aerodynamic control laws from algorithmic noise.

With the W-MSSA-reconstructed actions serving as the expert target, a CART algorithm is employed for policy distillation. To ensure the extracted rules are actionable, the CART model's inputs are strictly limited to observable flight deck instruments, such as airspeed and altitude. The tree depth is mathematically constrained to prevent overfitting, thereby extracting a generalized, hierarchical set of "If–Then" physical SOP rules.

To visually validate the efficacy of the proposed physical SOP extraction pipeline, Figure 10 illustrates the transformation from raw neural network commands to human-executable discrete rules within a unified coordinate system. This multilayered visualiza-

tion explicitly demonstrates how the AI's complex maneuvering intent is distilled into an actionable flight manual.

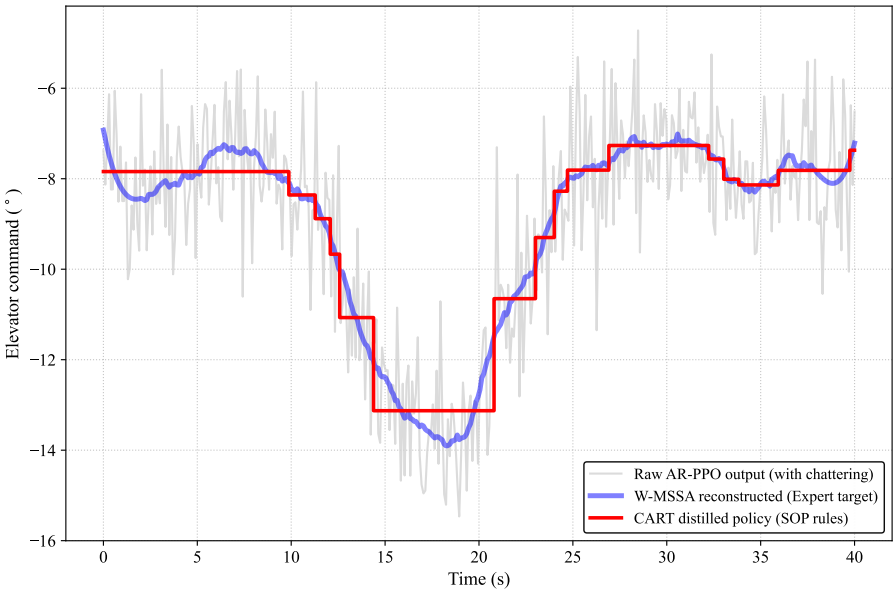


Figure 10. Signal reconstruction and policy distillation fidelity.

The underlying gray line represents the raw AR-PPO elevator command. It is heavily corrupted by high-frequency algorithmic chattering, a common artifact of neural network exploration that renders direct implementation unfeasible.

The W-MSSA algorithm acts as a robust filter against this noise, successfully extracting the continuous, deterministic expert trajectory (thick light blue line). This reconstructed signal clearly reveals the true aerodynamic intent of the AR-PPO agent: rather than panicking and pulling the stick, the agent executes a decisive, smooth pitch-down maneuver (pushing the elevator to approximately -14° between 15 and 20 s) to trade altitude for critical airspeed inside the severe vortex core.

Finally, the CART policy distillation captures the "zero-order hold" cognitive characteristic of human manual flight. Human pilots cannot execute continuous gradients; they operate in discrete, stepped maneuvers. The resulting distilled policy perfectly replicates this human physiological trait. More importantly, the red line maintains a tight, high-fidelity adherence to the W-MSSA expert target without overfitting to the local algorithmic noise.

This high congruence proves that the extracted explicit "If–Then" rules successfully retain the optimal aerodynamic recovery logic. By bridging the critical gap between DRL optimality and human execution, this framework provides a highly transparent, certifiable, and pilot-trainable solution for extreme low-altitude windshear encounters.

As illustrated in Figure 11, the intricate AI control strategies are successfully decoded into pilot-readable Standard Operating Procedures (SOPs).

Figure 11 presents the hierarchical decision tree extracted via the CART policy distillation, which seamlessly translates abstract reinforcement learning intelligence into actionable aerodynamic guidance. The logical flow of the extracted SOP can be analyzed from the root to the terminal leaves.

Triggering mechanism (Root Node). The root node serves as the primary anomaly detector. It continuously monitors the aircraft's energy state, triggering the wind shear recovery protocol only when a severe energy loss is detected (e.g., deceleration rate $\dot{V} < -2.0\ \mathrm{m/s^2}$ combined with a low altitude threshold $h < 300$ m).

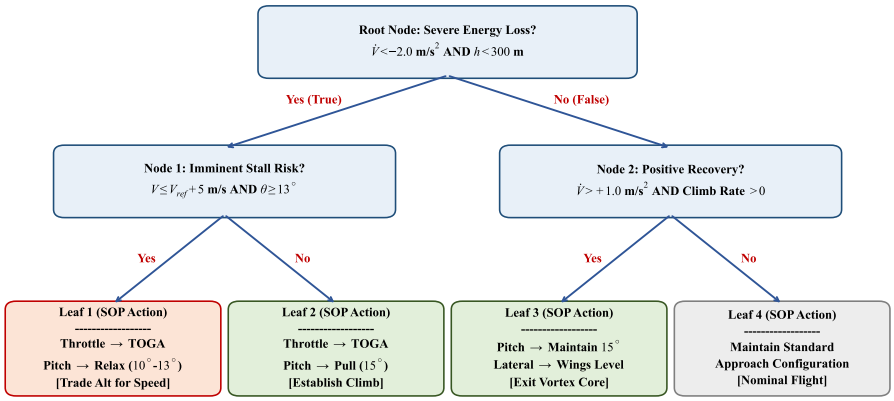


Figure 11. Extracted physical SOP via CART policy distillation.

Critical risk assessment (Node 1). If the aircraft is caught in the severe downdraft (Root = True), the tree immediately evaluates the imminent stall risk at Node 1. The decision boundary relies on the critical airspeed margin $V \leq V_{ref} + 5$ m/s and the current pitch attitude $\theta \geq 13°$.

The vortex core strategy (Leaf 1 and Leaf 2). If an imminent stall risk is identified (Node 1 = Yes), the tree directs the policy to Leaf 1, which embodies the most critical, counter-intuitive physical law discovered by the agent: relaxing the pitch to trade altitude for airspeed. This prevents a catastrophic stall during the vortex core penetration. Conversely, if the airspeed remains within a safe margin (Node 1 = No), the policy shifts to Leaf 2, commanding a standard pull-up maneuver to establish a positive climb.

Recovery and nominal flight (Node 2, Leaf 3 and 4). Once the aircraft exits the core downdraft and registers a positive recovery trend $\dot{V} > +1.0$ m/s^2 and positive climb rate, the tree navigates to the right branch. Leaf 3 instructs the pilot to maintain the current pitch and keep the wings level to safely exit the wind shear zone, eventually transitioning back to the nominal approach configuration (Leaf 4).

Ultimately, Figure 1 demonstrates that the proposed framework does not merely output a rigid numerical trajectory; rather, it distills a highly interpretable, physics-informed decision-making paradigm. This guarantees that the AI-generated strategies can be safely integrated into human-in-the-loop aviation operations.

4. Discussion

The experimental results demonstrate that the proposed AR-PPO framework fundamentally resolves the limitations of conventional DRL and manual flight operations in extreme windshear environments. A pivotal finding of this study is the AI agent's autonomous discovery of the "altitude-for-airspeed" trade-off during vortex core penetration. Traditional MR-PPO agents and human pilots often succumb to the "stall trap," greedily pulling the elevator to preserve altitude, which rapidly depletes kinetic energy and leads to catastrophic lift loss.

In contrast, guided by the meta-gradient mechanism, the AR-PPO agent prioritizes airspeed over altitude, proving that the proposed bilevel optimization architecture empowers the agent to deduce fundamental aerodynamic laws from sparse survival outcomes rather than blindly following static reference tracking.

Furthermore, the comparative analysis against empirical human pilot data highlights the physiological boundaries of manual control. The human cognitive loop introduces dangerous reaction latencies and pilot-induced oscillations (PIO), particularly when managing coupled longitudinal and lateral-directional dynamics under severe asymmetric

crosswinds. The AR-PPO agent overcomes this by processing high-dimensional multi-axis inputs instantaneously, ensuring tightly bounded, perfectly decoupled control.

From a regulatory and airworthiness certification perspective, existing certification frameworks for microburst avoidance systems strictly require deterministic guarantees and exhaustive formal verification. This fundamentally clashes with the statistical and stochastic nature of deep neural networks. A significant philosophical challenge remains: regulatory agencies, such as EASA and the FAA, demand absolute predictability and traceability in safety-critical systems. It is important to note that most existing explainable reinforcement learning (XRL) methods primarily focus on feature attribution to interpret the internal cognitive intent of the neural networks, providing post-hoc statistical interpretations without guaranteeing strict bounds on physical execution. To overcome this hurdle, the aviation AI must transition from statistical explainability to programmatic and deterministic verification. In contrast to standard cognitive XRL methods, this study utilizes wavelet analysis, multivariate singular spectrum analysis (MSSA), and sequential pattern mining explicitly as the core tools for explainable reinforcement learning. This unique policy distillation pipeline focuses on physical execution, reconstructing continuous, black-box trajectories into transparent, discrete "IF–THEN" physical SOPs. By explicitly translating uncertifiable neural weights into verifiable operational rules, direct algorithmic performance comparisons with standard cognitive XRL methods become inapplicable. More importantly, this pipeline ensures that the advanced recovery paradigms discovered by the AI are actionable, physically bounded, and aligned with the stringent predictability requirements necessary for eventual regulatory acceptance and integration into evidence-based training (EBT) for human pilots. Despite these promising advancements, this study has certain limitations that must be acknowledged. First, the validation relies entirely on a simulation-based environment utilizing a 6-DOF B737-800 flight dynamic model, which cannot fully replicate the hardware-in-the-loop latencies of actual flight operations. Second, the deterministic vortex-ring model used for the micro downburst inherently lacks the stochastic turbulence intensity (e.g., Dryden or von Karman models) and atmospheric uncertainties commonly observed in real-world conditions. Therefore, future research will focus on integrating stochastic atmospheric variations to enhance the fidelity of the windshear model. Ultimately, we plan to validate these extracted SOPs through human-in-the-loop flight simulator experiments, utilizing the distilled rules for quantitative pilot training analysis to further enhance the certification readiness of AI-driven flight control systems.

5. Conclusions

This paper presents a robust, explainable data-driven reinforcement learning framework for autonomous windshear escape, effectively addressing the unreadability and instability of conventional DRL in aviation applications. The main conclusions are as follows.

Autonomous reward discovery. The proposed AR-PPO framework utilizes bilevel optimization and meta-gradients to autonomously map sparse episodic outcomes into dense aerodynamic guidance, successfully mitigating gradient interference and avoiding the heuristic "stall trap".

Robust zero-shot generalization. By actively prioritizing kinetic energy preservation during severe vortex core penetration, the AR-PPO agent achieves superior recovery performance and precise multivariable decoupling. It significantly outperforms static-reward algorithms and empirical human pilots, demonstrating exceptional zero-shot adaptability in unseen extreme micro downbursts (22.0 m/s).

Transparent policy distillation. By employing W-MSSA for noise filtering and signal reconstruction, alongside CART for decision boundary extraction, the black-box neural network strategies are successfully distilled into human-readable SOPs. This transparent

representation ensures that the AI-generated maneuvers are physically executable and verifiable, establishing a solid foundation for human-in-the-loop operations and airworthiness certification.

Author Contributions: Data Curation, Y.W.; Formal Analysis, Y.Z. and Z.G.; Funding Acquisition, Z.G.; Investigation, Y.Z. and Y.W.; Methodology, Y.Z. and Y.W.; Project Administration, Y.W.; Resources, Y.W.; Software, Y.Z.; Supervision, Z.G.; Validation, Y.Z. and Z.G.; Visualization, Y.Z. and Y.W.; Writing—Original Draft, Y.Z. and Y.W.; Writing—Review and Editing, Y.W. and Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Civil Aviation Administration of China (ASSA2024/121), the National Natural Science Foundation of China (U2333202, 52272351) and the Aeronautical Science Foundation of China (2022Z066052002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Shandong Airlines and are available from Yitan Wang with the permission of Shandong Airlines.

Acknowledgments: The authors acknowledge the funding supported by Zhenxing Gao and the data provided by Shandong Airlines.

Conflicts of Interest: Yitan Wang was employed by Shandong Airlines Co., Ltd. The remaining author declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AEDRL	Active exploration deep reinforcement learning
AR-PPO	Active-reward proximal policy optimization
CART	Classification and regression tree
DRL	Deep reinforcement learning
IL	Imitation learning
IRL	Inverse reinforcement learning
PIRL	Programmatically interpretable RL
PPO	Proximal policy optimization
RL	Reinforcement learning
SILVER	Shapley value-based interpretable policy via explanation verification
XRL	Explainable reinforcement learning
DWT	Discrete wavelet transform
IDWT	Inverse discrete wavelet transform
MRA	Multi-resolution analysis
MSSA	Multivariate singular spectrum analysis
SNR	Signal-to-noise ratio
SVD	Singular value decomposition
W-MSSA	Wavelet analysis combined with multivariate singular spectrum analysis
6-DOF	Six-degree-of-freedom
CBTA	Competency-based training and assessment
CFIT	Controlled flight into terrain
EBT	Evidence-based training
FDR	Flight data recorder
PIO	Pilot-induced oscillation
SOP	Standard operating procedure
TOGA	Take-off/go-around

References

1. Razzaghi, P.; Tabrizian, A.; Guo, W.; Chen, S.; Taye, A.; Thompson, E.; Bregeon, A.; Baheri, A.; Wei, P. A survey on reinforcement learning in aviation applications. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108911. [CrossRef]
2. Lombaerts, T.; Looye, G.; Ellerbroek, J.; Martin, M.R.Y. Design and piloted simulator evaluation of adaptive safe flight envelope protection algorithm. *J. Guid. Control Dyn.* **2017**, *40*, 1902–1924. [CrossRef]
3. Harms, M.; Lim, J.; Rohr, D.; Rockenbauer, F.; Lawrance, N.; Siegwart, R. Towards robust optimization-based autonomous dynamic soaring with a fixed-wing UAV. *IEEE Robot. Autom. Lett.* **2026**, 7620–7627. [CrossRef]
4. Liu, F.; Dai, S.; Zhao, Y. Learning to Have a Civil Aircraft Take Off under Crosswind Conditions by Reinforcement Learning with Multimodal Data and Preprocessing Data. *Sensors* **2021**, *21*, 1386. [CrossRef] [PubMed]
5. Park, S.; Fanjoy, A.; Golubev, V.V. Application of reinforcement learning for autonomous dynamic soaring. In Proceedings of the AIAA Scitech 2025 Forum, Orlando, FL, USA, 6–10 January 2025. [CrossRef]
6. Chen, L.; Lu, J.; Yin, Y.; Huang, J.; Xiang, Y.; Liu, H. Learning step-level dynamic soaring in shear flow. *arXiv* **2026**, arXiv:2604.12413. [CrossRef]
7. Ma, Q.; Wu, Y.; Shoukat, M.U.; Yan, Y.; Wang, J.; Yang, L.; Yan, F.; Yan, L. Deep reinforcement learning-based wind disturbance rejection control strategy for UAV. *Drones* **2024**, *8*, 632. [CrossRef]
8. Zhang, Y.; Chai, S.; Zhang, Y.; Huang, D.; Ge, Q. Cascaded TD3-PID hybrid controller for quadrotor trajectory tracking. *arXiv* **2026**, arXiv:2604.13505. [CrossRef]
9. Ren, Y.; Zhu, F.; Sui, S.; Yi, Z.; Chen, K. Enhancing quadrotor control robustness with multi-proportional–integral–derivative self-attention-guided deep reinforcement learning. *Drones* **2024**, *8*, 315. [CrossRef]
10. Nigam, R.; Choi, J.; Parikh, N.; Li, M.Z.; Tran, H.T. Survey of inverse reinforcement learning in aviation and future outlooks. *J. Aerosp. Inf. Syst.* **2026**, *23*, 305–321. [CrossRef]
11. Song, L.; Guo, Q.; Channa, I.A.; Wang, Z. A survey of maximum entropy-based inverse reinforcement learning: Methods and applications. *Drones* **2025**, *17*, 1632. [CrossRef]
12. Zare, M.; Kebria, P.M.; Khosravi, A.; Nahavandi, S. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Trans. Cybern.* **2024**, *54*, 7173–7186. [CrossRef] [PubMed]
13. Kayal, A.; Pignatelli, E.; Toni, L. The impact of intrinsic rewards on exploration in reinforcement learning. *Neural Comput. Appl.* **2025**, *37*, 16269–16303. [CrossRef]
14. Pathak, D.; Agrawal, P.; Efros, A.A.; Darrell, T. Deep reinforcement learning-based wind disturbance rejection control strategy for UAV. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017. Available online: https://proceedings.mlr.press/v70/pathak17a.html (accessed on 4 August 2026).
15. Burda, Y.; Edwards, H.; Storkey, A.; Klimov, O. Exploration by random network distillation. *arXiv* **2017**, arXiv:1810.12894. [CrossRef]
16. Saglam, B.; Kozat, S.S. Deep intrinsically motivated exploration in continuous control. *Mach. Learn.* **2023**, *112*, 4959–4993. [CrossRef]
17. Oh, J.; Farquhar, G.; Kemaev, I.; Calian, D.A.; Hessel, M.; Zintgraf, L.; Singh, S.; van Hasselt, H.; Silver, D. Discovering state-of-the-art reinforcement learning algorithms. *Nature* **2025**, *648*, 312–319. [CrossRef] [PubMed]
18. Lu, R.; Shao, Z.; Ding, Y.; Chen, R.; Wu, D.; Su, H.; Yang, T.; Zhang, F.; Wang, J.; Shi, Y.; et al. Discovery of the reward function for embodied reinforcement learning agents. *Nat. Commun.* **2025**, *16*, 11064. [CrossRef] [PubMed]
19. Zhao, D.; Xu, H.; Zhang, X. Active exploration deep reinforcement learning for continuous action space with forward prediction. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 6. [CrossRef]
20. Russo, A.; Proutiere, A. Model-free active exploration in reinforcement learning. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023. [CrossRef]
21. Eysenbach, B.; Gupta, A.; Ibarz, J.; Levine, S. Diversity is all you need: Learning skills without a reward function. *arXiv* **2019**, arXiv:1802.06070. [CrossRef]
22. Zhao, R.; Gao, Y.; Abbeel, P.; Tresp, V.; Xu, W. Mutual information state intrinsic control. *arXiv* **2021**, arXiv:2103.08107. [CrossRef]
23. As, Y.; Sukhija, B.; Treven, L.; Sferrazza, C.; Coros, S.; Krause, A. ActSafe: Active exploration with safety constraints for reinforcement learning. *arXiv* **2025**, arXiv:2410.09486. [CrossRef]
24. Bush, T.; Chung, S.; Anwar, U.; Garriga-Alonso, A.; Krueger, D. Interpreting emergent planning in model-free reinforcement learning. *arXiv* **2025**, arXiv:2504.01871. [CrossRef]
25. Qian, Y.; Nguyen, S.; Chen, C.; Zhou, Q.; Zhao, L. Interpret policies in deep reinforcement learning using SILVER with RL-guided labeling. *arXiv* **2025**, arXiv:2510.19244. [CrossRef]
26. Gokhale, G.; Karimi Madahi, S.S.; Claessens, B.; Develder, C. Distill2Explain: Differentiable decision trees for explainable reinforcement learning in energy application controllers. In Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems, Singapore, 4–7 June 2024. [CrossRef]

27. Wen, Y.; Li, S.; Zuo, R.; Yuan, L.; Mao, H.; Liu, P. SkillTree: Explainable skill-based deep reinforcement learning for long-horizon control tasks. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025. [CrossRef]
28. Zhang, M.; Miao, Z.; Nan, X.; Ma, N.; Liu, R. Explainable reinforcement learning for the initial design optimization of compressors. *Biomimetics* **2025**, *10*, 497. [CrossRef] [PubMed]
29. Engelhardt, R.C.; Lange, M.; Wiskott, L.; Konen, W. Sample-based rule extraction for explainable reinforcement learning. In *Machine Learning, Optimization, and Data Science*; Nicosia, G., Ojha, V., La Malfa, E., La Malfa, G., Pardalos, P., Di Fatta, G., Giuffrida, G., Umeton, R., Eds.; Publishing House: Cham, Switzerland, 2023; pp. 330–345. [CrossRef]
30. Shawon, R.U.; Liu, S.; Siddiqua, A. Explainable reinforcement learning for multi-agent systems. In Proceedings of the 2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI), Athens, Greece, 3–5 November 2025. [CrossRef]
31. Verma, A.; Murali, V.; Singh, R.; Kohli, P.; Chaudhuri, S. Programmatically interpretable reinforcement learning. In Proceedings of the 35th International Conference on Machine Learning (PMLR), New Orleans, LA, USA, 10–15 July 2018. Available online: https://proceedings.mlr.press/v80/verma18a.html (accessed on 4 August 2026).
32. Champion, K.; Lusch, B.; Kutz, J.N.; Brunton, S.L. Data-driven discovery of coordinates and governing equations. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22445–22451. [CrossRef] [PubMed]
33. Stevens, B.L.; Lewis, F.L.; Johnson, E.N. *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2015. Available online: https://pure.psu.edu/en/publications/aircraft-control-and-simulation-dynamics-controls-design-and-auto/ (accessed on 4 August 2026).
34. Schultz, T.A. Multiple vortex ring model of the DFW microburst. *J. Aircr.* **1990**, *27*, 163–168. Available online: https://arc.aiaa.org/doi/pdf/10.2514/3.45913?casa_token=IgxKdQQfL0gAAAAA:A-ajzBNu0EgZTBXTKFGz8nHssQ-n34gwqiob3WCPYnf-2dVXMK8B9yYb9fr5wLbQU-lbPUMi (accessed on 4 August 2026). [CrossRef] [PubMed]
35. Akbar, H.; Arshad, I.; Rahman, S.; Aamir, F.; Fatima, M.; Khan, H.U.R. TRUST-X: Real-time explainable decision-making framework for reinforcement learning-based UAV control. In Proceedings of the 2025 20th International Conference on Emerging Technologies (ICET), Peshawar, Pakistan, 18–19 November 2025. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.