aerospace

MDPI

Article

Time-Window Mismatch in Reactive Windshear Alerting During Approach and a False-Alarm Suppression Algorithm Based on a Physical-Consequence Criterion

Qianhao Liu and Dongsu Wu *

College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; lqh1008163@163.com
* Correspondence: wudongsu@nuaa.edu.cn

Abstract

Low-level windshear is among the most hazardous weather on approach, and the reactive alerting chain fixed by certification standards such as TSO-C117 is the last line of defense. This paper asks whether the fixed five-second decision window of the certified F-factor criterion leaves a systematic false-alarm space in compact-microburst encounters, and whether a criterion based on physical consequence can suppress this space without modifying certified behavior; the question is examined on a Boeing 747-400 approach-phase simulation platform, with the TSO-C117 alerting chain reproduced in the loop as the certification baseline. Replaying the standard's six reference window lengths (5–10 s) shows that false-alarm suppression and hazard detection trade off along the window axis, and none of the tested windows satisfies both. A three-channel short-window (0.5 s) neural detector with force-residual auxiliary tasks from Newton's second law therefore runs in parallel with the untouched certified chain and arbitrates the alert level. On 14,825 Monte Carlo scenarios, the false-alarm-zone alert rate drops from the 99.2% certification baseline to 8.0%, with residual triggers concentrated near the upper edge of the divergence band, while at the classifier level (a complementarity analysis, not a deployed behavior) the union of the two criteria misses none of the 6888 genuinely hazardous scenarios; re-adjudication under four third-party consequence criteria confirms the suppressed scenarios are benign. With zero real data in training, the detector covers 68.4% of 329 GPWS windshear alert events in real-flight QAR records, responding a median 14.5 s before the alert trigger. The physical-consequence criterion extends the classical predictive/reactive complementarity into the reactive layer.

Keywords: low-level windshear; reactive alerting; F-factor; false alarm; physical-consequence criterion; flight safety

Check for updates

Academic Editor: Miroslav Kelemen

Received: 28 September 2026
Revised: 4 November 2026
Accepted: 7 November 2026
Published: 10 November 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.

1. Introduction

Low-level windshear is a rapid change in wind speed and/or direction over a short distance in the lowest layer of the atmosphere, and one of the most insidious hazards in aviation: between 1964 and 1983 it caused at least 28 accidents of large transport aircraft and more than 500 fatalities; its main hazard region lies below about 1600 ft above ground, where the airspeed and altitude margins left to the crew are small and the reaction time is extremely short [1]. Among the weather phenomena that produce low-level windshear, the microburst is an intense small-scale downdraft together with its divergent outflow after ground contact [2,3]; it compresses the strongest performance loss into the shortest

encounter time. Most of the fatal accidents above are associated with it [4–7], and it therefore dominates the design requirements of alerting systems [4].

Protection against low-level windshear mainly comprises three approaches: ground-based remote sensing; airborne predictive detection using forward-looking radar or lidar; and airborne reactive systems that compute the hazard from onboard flight data after the shear has been encountered [1,4]. Among them, reactive windshear alerting is the last line of defense of the transport fleet. Its alerting logic is fixed by certification standards [8] and is based on the F-factor proposed by Bowles [4], the normalized rate of change of aircraft specific energy caused by the wind field; an F-factor that exceeds 0.1 and persists is regarded as hazardous [4]. This criterion remains the reference metric in current research [9].

The design of alerting systems has long been fixed around the F-factor: certification tests prescribe a set of must-alert wind-field scenarios and exposure times, and impose an upper bound on the false-alarm rate through statistical metrics [8], so false alarms are controlled at the design level. Throughout this paper, a false alarm denotes an alert issued by the criterion under test for an encounter whose actual physical consequence stays below the hazard level; the consequence is operationalized in Section 3.3 as the forward 5 s airspeed loss, with 5 m/s as the hazard boundary, so that the term always refers to a measurable physical outcome rather than to a signal-level discrepancy. This engineering tradeoff is reasonable: a system that frequently false-alarms on minor disturbances could neither pass certification nor retain crew trust. The question that needs discussion is therefore not whether the criterion is effective, but whether the fixed form of the alert design leaves residual space at the margin. The certification statistics are deliberately simple—a threshold applied to the filtered F-factor with fixed persistence and averaging windows—whereas the protected aircraft has its own real-time dynamic response scale, which varies with encounter geometry, intensity, and flight phase. When the aircraft encounters a compact shear whose duration is short but whose intensity suffices to cross the threshold, the decision instant of the criterion and the physical response instant of the aircraft can decouple: an alert is triggered, yet the subsequent trajectory shows no meaningful performance loss. Such residual false alarms are concentrated boundary cases rather than a general defect, and they are hard to remove by globally retuning the threshold or the method, because this would relax the detection guarantee on which the standard rests.

Machine learning entered windshear research mainly through ground infrastructure, namely, airport sensing, forecasting, and post-event analysis systems [10–12]. On the airborne-software side, deep networks have been used to estimate vertical wind and the averaged F-factor (FBAR) for radar systems, with development aligned to the emerging SAE ARP6983/ED-324 assurance standards [13,14] and EASA's Artificial Intelligence Roadmap [15]; the certified reactive decision layer itself, however, has not yet been touched by learning methods. A parallel methodological development is physics-informed machine learning [16,17], which embeds physical structure into networks as architectural constraints or auxiliary objectives and has begun to appear in aviation applications, including aeroengine sensor diagnostics [18] and airport noise prediction [19]. Together, these observations define the gap this paper addresses: the reactive alerting layer is standard-driven and effective, yet it retains a bounded false-alarm space, and a learning method based on a physical-consequence criterion is well suited to shrinking it.

The objective of this paper is to verify two claims under these conditions: the fixed decision window of the certified criterion leaves a systematic and localized false-alarm space in compact encounters; and a detector deciding on the forward physical consequence can suppress this space without degrading hazard detection and without modifying the certified chain. The experimental setup is as follows. All experiments use a six-degree-of-freedom Boeing 747-400 flight-dynamics simulation platform in the approach phase, coupled with the downburst/microburst models of Oseguera–Bowles and Bray [20,21]; the TSO-C117 criterion chain is reproduced offline as the certification baseline. Simulations are generated by Monte Carlo sampling within the parameter ranges of seven scenario families, with the aircraft fixed at a nominal mass of 260 t and an approach speed of 75 m/s, producing 710 independent simulation records (about 2.63 million flight frames). Figure 1 summarizes the overall framework: the simulation platform and the reproduced certification chain form the experimental environment (Section 2); the time-window analysis localizes the false-alarm space and predicts the divergence band (Section 3.1); the physical-consequence detector runs in parallel with the certified chain and arbitrates the alert level (Sections 3.2 and 3.3); and validation proceeds from scenario-by-scenario Monte Carlo and adversarial search to real QAR records (Sections 4 and 5). With this framework, the paper makes four contributions: (i) a time-window analysis showing that the decision window of the certified criterion and the energy-response window of the aircraft systematically decouple near the alert threshold, which explains where residual false alarms concentrate; (ii) a multi-channel short-window neural network that decides on physical consequences, in which the input is compressed by a shared encoder and three residual heads constructed axis by axis from Newton's second law provide physical intermediate representations, suppressing the false-alarm zone while preserving detection performance on hazardous encounters; (iii) scenario-by-scenario paired extended validation on 14,825 valid scenarios accumulated in two batches under the same distribution and protocol (about 58.6 million frames), in which at the classifier level the union of the two criteria detects 100% of the 6888 scenarios containing genuine hazard and the alert rate of the false-alarm zone drops from 99.2% to 8.0% (Section 4.1), together with a Bayesian optimization guided by criterion–physics divergence that independently locates the false-alarm-prone region outside the training data, showing that the divergence band and the detection performance are independent of the training data but not of the training distribution—a boundary discussed in Section 4.3; and (iv) a cross-validation with zero real data in training on 2411 real-flight QAR records (2346 flights), in which the mapped detector covers 68.4% of 329 GPWS windshear alert events, with 221 events responded to earlier than the alert trigger, at a median time gap of 14.5 s. The setup above defines the applicability range of the quantitative results reported below.

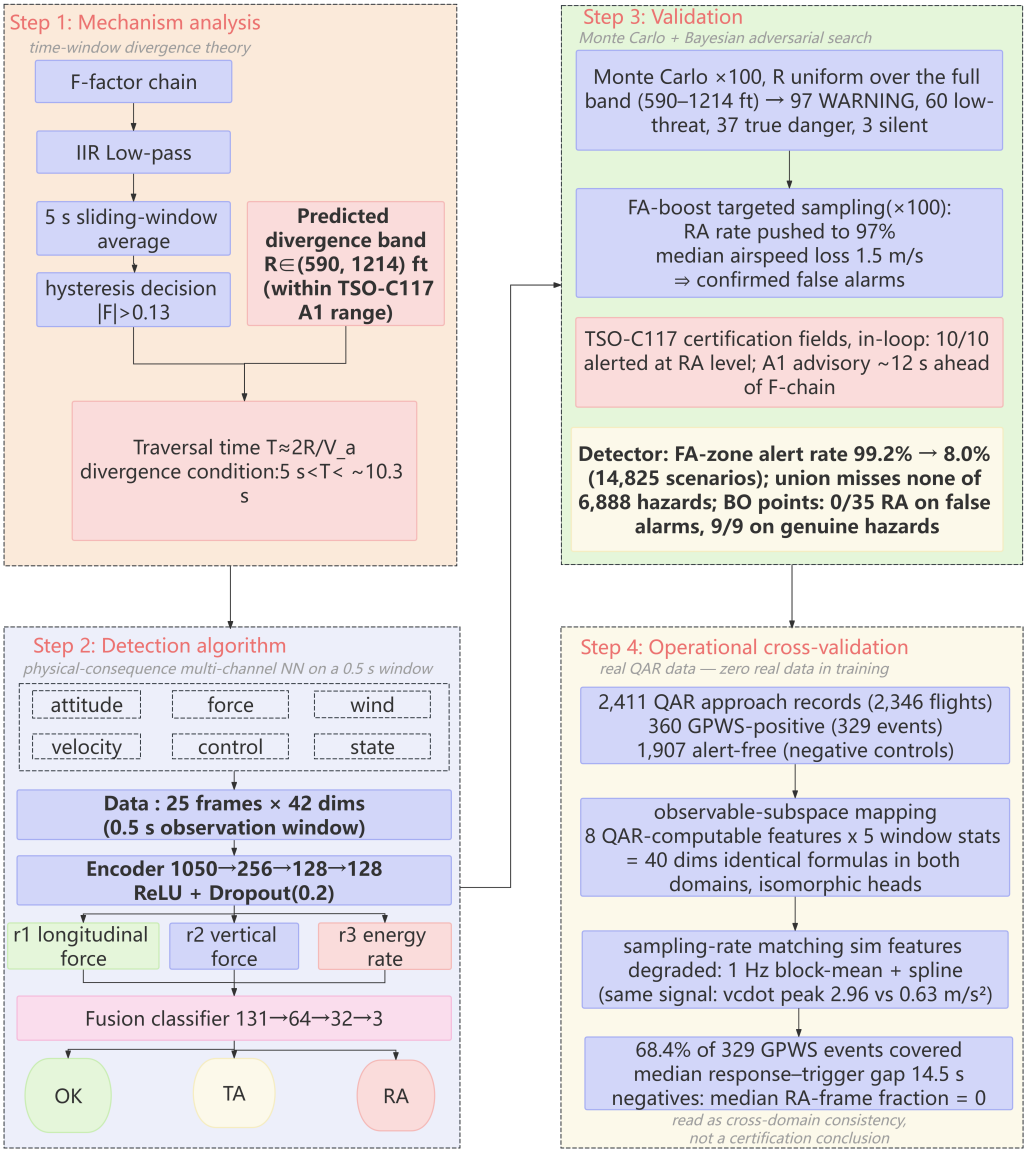


Figure 1. Overall framework. Step 1 is mechanism analysis: the scale mismatch between the time window of the certified criterion chain and the physical duration of the event yields a predicted divergence band $R \in (590, 1214)$ ft. Step 2 is the detection algorithm: a 0.5 s short-window input (25 frames × 42 dimensions) is compressed by a shared encoder, three force-residual heads provide physical intermediate representations, and the fusion classifier outputs three levels, OK/TA/RA. Step 3 is simulation validation: in-band Monte Carlo, Bayesian adversarial search, and in-loop reproduction of the ten TSO-C117 certification wind fields. Step 4 is operational cross-validation: the mapped model, trained with zero real data, covers 68.4% of 329 GPWS windshear alert events with a median response–trigger gap of 14.5 s.

2. Simulation Platform and Windshear Modeling

2.1. Simulation Platform

All experiments are carried out on a Boeing 747-400 six-degree-of-freedom nonlinear flight-dynamics simulation platform that outputs the complete aircraft state at 50 Hz. The certification baseline and the proposed detector are implemented within the same platform, so both see exactly the same dynamics frame by frame. The simulation adopts the following assumptions:

1. The aircraft is a rigid body; the earth is flat and non-rotating.

2. The aircraft mass is constant within a single simulation; simulations are flown in the approach configuration at approach speed.

3. Aerodynamic forces and moments are quasi-steady, computed from tabulated coefficients of angle of attack, sideslip angle, and control-surface deflections.

4. The atmosphere consists of a prescribed spatially varying mean wind field (Section 2.2) with superimposed stochastic turbulence; the wind field does not evolve in time during a simulation.

The equations of motion are integrated in the body-axis system. Taking the longitudinal x channel as an example, the air-relative acceleration is

$$\dot{U} = \frac{F_x}{m} - q\,W + r\,V, \tag{1}$$

where U, V, W are the components of the air-relative velocity on the body axes, F_x is the total force along the body x axis (aerodynamic, propulsive, and gravitational components), m is the aircraft mass, and q, r are body-axis angular rates; the y and z channels are analogous. The three channels of Equation (1) divide the work as follows: the x channel gives the longitudinal acceleration and is the source of the horizontal-shear term of the F-factor and of the residual head r_1; the z channel gives the vertical acceleration and is the source of the downdraft response and of the residual head r_2; the y channel carries the lateral force balance. The air-relative velocity is obtained by integrating the accelerations above; adding the local wind vector and the turbulence component gives the inertial (ground) velocity, and position is obtained by integrating the air-relative velocity in the navigation frame. The inertial (ground) velocity is the sum of the air-relative velocity and the local wind vector; differentiating this sum along the trajectory, the wind's spatial variation enters the ground-track dynamics through the wind-gradient term (the inner product of the wind gradient with the velocity), while no explicit time-derivative term appears because the prescribed field is frozen during a run (Assumption 4). The platform evaluates this prescribed field at the current aircraft position every frame, so the wind gradient acts on the ground-track dynamics whereas the aerodynamic forces in Equation (1) depend only on air-relative quantities. In other words, the prescribed wind field acts on the aircraft through the aerodynamic forces—exactly the coupling that alerting logic driven by air data and inertial measurements cares about.

The platform records 48 state channels per frame, including attitude, body-axis velocities and their derivatives, aerodynamic and propulsive forces, atmospheric quantities, and control-surface positions; the 42 channels used as detector input are listed in Appendix A. The certified alerting baseline of Section 2.3 is also executed in the loop on the same platform.

2.2. Windshear Modeling

Windshear in the simulation is generated by two mature analytic models. Analytic models are chosen instead of numerical weather fields [22] or measured data because they allow precise control of shear intensity, scale, and position with a few physical parameters, and this controllability is a prerequisite because this paper studies exactly the relationship between the time scale of the criterion and the scale of the event. The first model is the Oseguera–Bowles axisymmetric microburst model [20], which describes the downburst as a stagnation flow in the boundary layer and gives the three-dimensional wind components in closed form from two parameters: radius and intensity. The second is the Bray method (NASA TM-85969) [21], which superposes one to five analytic downburst cells, can construct non-axisymmetric wind fields such as dual-core-in-line configurations, and additionally provides the turbulence-intensity parameters required by the Dryden stochastic turbulence model [23]. The ten TSO-C117 must-alert certification wind fields of this paper are likewise

generated by these two model families, and their characteristic-radius range covers the certification scenario set [8].

The wind field is not solved online during the simulation. For each run, the wind field is first evaluated offline on a regular three-dimensional grid and stored as a CSV file; at run time the simulator interpolates trilinearly at the current aircraft position. The grid has a horizontal spacing of 100 m and six height layers above ground at 10, 50, 100, 200, 300, and 400 m vertically, keeping the original resolution. By convention, wind components are stored in the navigation frame (NED), with downdraft as a negative vertical component; the equivalent F-factor expression of Section 2.3 switches to a "downdraft-positive" intensity convention.

If wind fields from a single parameter configuration were used, the detector might learn the parameters rather than the physics. Simulation diversity is therefore obtained through two layers of sampling: one layer samples the downburst parameters (radius R, peak vertical velocity V_{ZO}, release height H_T) within certification-relevant ranges, and the other randomizes the turbulence realization seed; aircraft-side parameters are fixed at a nominal mass of 260 t and an approach speed of 75 m/s, and the generalization boundary with respect to mass variation is discussed in Section 5.4. A total of 715 raw simulation files were generated; after cleaning and minimum-length filtering, 710 valid files (about 2.63 million flight frames) remain, organized into seven families by scenario: TSO certification wind fields, NASA Bray wind fields, single-core and dual-core microbursts, a moving-average dual-core variant, compact false-alarm-prone microbursts, and pure turbulence without microburst. Later figures and tables keep the family codes used in the evaluation code: tsosupp for TSO certification wind fields, nasa_supp for NASA Bray wind fields, dual_danger for dual-core microbursts, ma_dual for the moving-average dual-core variant, fa_boost for compact false-alarm-prone microbursts, and turb for pure turbulence; single-core microburst file names carry no family tag and comprise all remaining files outside the six families above. Files are split at the file level into 568 training files and 142 test files (522,854 frames); no wind field appears in both sets.

2.3. Certification Baseline: Offline Reproduction of the TSO-C117 Alerting Chain

The certification baseline of this paper is an engineering implementation of the reactive F-factor criterion, reproduced in the loop on the platform of Section 2.1. It is a reconstruction of the alerting logic specified by the standard, not an evaluation of any specific certified equipment; the origin of every constant (thresholds, filter, confirmation times) is itemized at the end of this section, distinguishing values mandated or referenced by the standard from engineering choices of this paper. The F-factor [4,10] compares the shear intensity of the encounter with the energy state of the aircraft; its classical form is a moving average over a distance domain of about 1 km with a threshold of 0.13. The 0.13 threshold coincides with the 8 s must-alert condition of the standard's test table ($f_{\mathrm{av},x} = 0.1311$ over an 8 s exposure [8]); the distance-domain averaging of about 1 km is the prevailing form of classical implementations and is not prescribed by the standard, and both concrete constants are engineering choices of this reproduction, as itemized at the end of this section. This value must be distinguished from the 0.1 quoted in the Introduction: 0.1 is the reference level for judging shear hazard, whereas 0.13 is the alert threshold adopted in decision implementations. Since the wind components cannot be measured directly onboard, the implementation substitutes equivalent quantities from flight parameters at the 50 Hz frame rate: the horizontal-shear term is the difference between inertial acceleration and air-relative longitudinal acceleration, and the vertical-wind term is

$$w_{\mathrm{raw}} = U_{\mathrm{aero}} \sin(\theta - \alpha) + V_d, \tag{2}$$

where U_{aero} is the air-relative speed, θ is the pitch angle, α is the angle of attack, and V_d is the downward-positive vertical speed. The raw F-factor is

$$F_{\text{raw}} = \frac{\dot{U}_{\text{I}} - \dot{U}_{\text{aero}}}{g} + \frac{w_{\text{raw}}}{U_{\text{aero}}}. \tag{3}$$

Here, the first acceleration of Equation (3) is the inertial (ground-referenced) longitudinal acceleration, distinct from the air-relative acceleration of Equation (1); their difference isolates the horizontal shear, and U_{aero} is the air-relative speed of Equation (2).

The raw value of Equation (3) is not compared against the threshold directly; it first passes a low-pass filter (first-order IIR, coefficient 0.02, with the wind-deterioration rate pre-limited to $\pm 8\ \text{m/s}^2$ to suppress transients at grid boundaries) and is then averaged over a 1 s window (50 frames, monitor channel) and a 5 s window (250 frames, detection channel), respectively. The alert decision is given by a finite-state machine acting on the 5 s average: $|F|$ exceeding 0.13 enters the suspected state, and an alert is confirmed after persistence for 2 s (100 frames); below 0.065 the machine enters a recovery state with hysteresis, and the alert is released after staying stably below the threshold for 5 s. The first 15 s of each simulation are a filter-initialization phase and do not participate in detection.

This numerical implementation reproduces the must-alert behavior required by TSO-C117 on all ten certification wind fields (Section 4) and serves unchanged as the baseline in all comparisons below. Its two fixed time scales—the 5 s decision window and the 2 s persistence confirmation—are exactly the objects of the time-window divergence analysis in Section 3.1.

TSO-C117 is a performance-based standard: it prescribes the set of must-alert wind-field scenarios and the required alerting behavior, and gives reference combinations of window length and threshold (5–10 s windows), but it does not prescribe concrete algorithms or constant values. Constraints on false alarms are given as statistical metrics rather than scenario sets: for example, nuisance caution-level alerts caused by turbulence must not exceed one per 250 system-operating hours (equivalently 3000 flight cycles at one hour per cycle). In the present reproduction, the 0.13 threshold follows the prevailing value of classical implementations; the 5 s detection window takes the lower end of the 5–10 s reference range; and the 0.065 hysteresis threshold, the 2 s confirmation duration, the IIR coefficient 0.02, and the $\pm 8\ \text{m/s}^2$ pre-limit are engineering choices of this paper. In terms of attribution, the standard mandates only the must-alert scenario set and the statistical false-alarm constraints; the moving average over a distance domain of about 1 km is the prevailing form of classical implementations, corresponding to a time window of about 13.9 s at an approach speed of 72 m/s; the concrete constant values are the engineering choices of this paper. The distance-domain moving average and the 5 s decision window are two cascaded stages that coexist in such implementations—the former smooths the raw F estimate, the latter performs the persistence judgment—and the 13.9 s equivalent length of the smoothing stage does not replace the 5 s decision scale analyzed in this paper. Taking the 5 s window at the lower end of the reference combinations has a further must-alert constraint: with the 10 s window at the other end, the window-mean peak of the A1 event with $T \approx 7.8$ s would be diluted to about 0.78 of its instantaneous value, risking falling below the threshold and violating the must-alert requirement; 5 s is therefore also approximately the longest feasible window that still accommodates the certification scenario set. Section 3.1 shows that the divergence mechanisms (dilution and stretching) do not depend on the specific values of these constants; the position of the divergence band itself shifts with the window length, and the $R \in (590, 1214)$ ft reported in this paper corresponds to the 5 s window. Certification of reactive systems prescribes only the must-alert side; there is no standard scenario set that must not be alerted. This paper uses the in-band low-threat

Monte Carlo of Section 4.2, the turbulence family of Section 2.2, and the real alert-free flight set of Section 5 as the test on this side.

This section has defined the environment of all the experiments in this paper: the Boeing 747-400 approach-phase simulation platform, the Oseguera–Bowles/Bray wind-field model family, and the offline-reproduced TSO-C117 criterion chain. The method of this paper is proposed and validated within this environment; the detection improvements reported below should be understood as one step of progress under these specific conditions, not as a general conclusion on the generic problem of reactive windshear alerting.

3. Methodology

3.1. Time-Window Divergence Analysis

We first study the temporal behavior of the certified chain. Before an alert is issued, the detection signal of Section 2.3 passes through three preprocessing stages: an IIR filter ($\alpha = 0.02$, time constant $\tau \approx 1$ s, cutoff frequency ≈ 0.16 Hz, 10–90% rise time ≈ 2.2 s), a 5 s sliding window (fill time ≈ 5 s), and state-machine persistence confirmation ($\approx$2 s). The sum of the three delays determines when the alert is actually issued, while the time scale that dilutes the detection signal is dominated by the 5 s sliding window: events shorter than the window cannot cross the threshold in the window mean and never reach persistence confirmation. This response property does not depend on the specific values of filtering and window length: we ran three window/filter configurations on three scenario groups of five cases each (pure turbulence, harmless compact microbursts, and strong dual-core microbursts) to verify it. The instantaneous-value configuration reaches a peak of $|F| = 0.164$ under turbulence, above the 0.13 threshold, and flips 20 times; the 5 s window compresses the turbulence peaks to 0.033–0.036 with zero flips, with or without filtering. However, all three configurations false-alarm 5/5 on the compact-microburst group and detect 5/5 on the dual-core group—under a short window, the peak of a compact event is still about 0.36 and stays above threshold for about 2.5 s, enough to satisfy the confirmation logic. In other words, within the structure of "fixed threshold plus persistence confirmation", none of the three representative configurations tested here passes both the turbulence-stability and the false-alarm tests. A full joint threshold–window sweep is beyond the scope of this paper, and we do not deny that such a parameter point may exist; the claim of this paper is that the scale mismatch is hard to eliminate by local adjustment of window and filtering.

Back to time stretching. When an aircraft at approach speed V_a crosses a compact microburst, the traversal time of the effective shear region is about $T \approx 2R/V_a$, where R is the characteristic radius; at a typical approach speed $V_a \approx 72$ m/s, T is determined by R alone. Throughout this paper, V_a denotes the calibrated airspeed on the approach segment, fixed at 75 m/s in the Monte Carlo experiments and taken as 72 m/s in the analytic time-scale estimates. The filter-plus-sliding-window chain applies two temporal transformations to the detection signal: one is peak dilution—events shorter than the window are compressed to about $F_0 \cdot T/5$ and may fall back below the threshold. The other is time stretching—after the event ends, the window needs about one window length to empty, so the output stays elevated for several seconds. A moving average preserves the total area of the signal: whatever the peak loses, the duration gains back, so neither effect raises the peak. False alarms occur on the stretching side: the instantaneous F peak of a compact event itself exceeds 0.13, and the long window stretches this brief exceedance into about 3.3 s of continuous above-threshold output, just exceeding the 2 s persistence-confirmation condition; after the event ends, the window output needs about one more window length to drain completely. Persistence confirmation is itself a reasonable design against turbulence flicker, but it relies on an implicit premise: the signal lasts as long as the physical shear lasts.

Stretching breaks exactly this premise: a stretched compact event and a genuine persistent shear are indistinguishable to the confirmation logic, and the alert is issued (Figure 2).

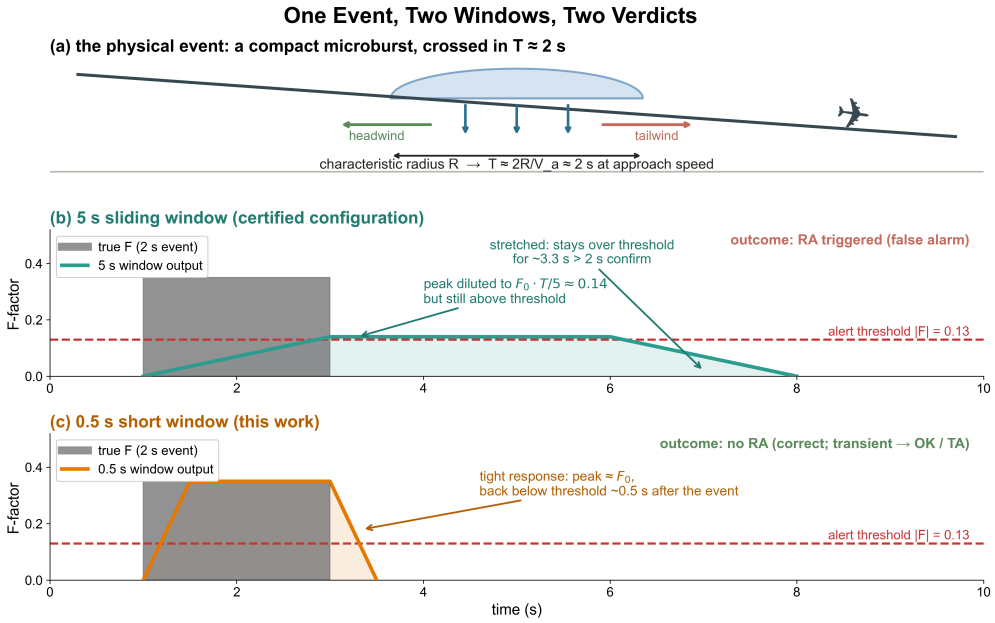


Figure 2. Decision comparison of the same compact-microburst encounter under two window lengths. (**a**) The physical event: traversal time $T \approx 2R/V_a \approx 2$ s. (**b**) The certified 5 s sliding window: the peak is diluted to about $F_0 \cdot T/5$ but still crosses the 0.13 threshold; continuous above-threshold output lasts about 3.3 s, satisfying the 2 s persistence confirmation and triggering an alert—a false alarm. (**c**) The 0.5 s short window of this work: the output tracks the event envelope and returns below threshold about 0.5 s after the event ends, correctly judged OK/TA.

Putting the physical time scale of the microburst, $T \approx 2R/V_a$, next to the window-dominated response scale of about 5 s, the encounter space naturally divides into three regions. When T is below about 5 s, the event ends before the sliding-window output can satisfy the confirmation condition and the chain tends to stay silent; this may be called the dilution-protected region. When T exceeds about 10.3 s, a shear strong enough to alert generally accumulates an airspeed loss beyond 5 m/s, which is a genuine energy hazard under any reasonable consequence criterion; this is the agreement zone. Between them lies the divergence band 5 s $< T <$ ~10.3 s: the event pushes the full-chain output across the threshold and completes confirmation, yet the physical consequence of a considerable fraction remains mild—of 400 uniformly sampled cases within the band, 379 trigger an alert, of which 242 have a forward 5 s loss not exceeding 5 m/s, with a median of 2.8 m/s and a maximum of 4.98 m/s (Section 4.2). The signal criterion and the consequence criterion diverge systematically here. The 5 s lower bound is set by the window-dominated dilution mechanism; the upper bound of about 10.3 s is semi-empirical, calibrated by the loss distribution of a uniform in-band Monte Carlo: the loss rises monotonically with R, and all samples with $R > 1100$ ft already exceed 5 m/s (Section 4.2); the adversarial search of Section 4.3, without using this calibration information, independently converges into the band (centroid $R \approx 704$ ft, $T \approx 6.0$ s), constituting an external consistency check on the upper bound. Substituting $V_a = 72$ m/s into $T \approx 2R/V_a$ gives the radius band $R \in (590, 1214)$ ft. This band overlaps the FAA TSO-C117 certification wind fields [8], and the most compact certification wind field A1 ($R = 918$ ft, $T \approx 7.8$ s) falls inside it; the divergence band is therefore not a theoretical extrapolation but a real flight region explicitly covered by the certification standard (Figure 3).

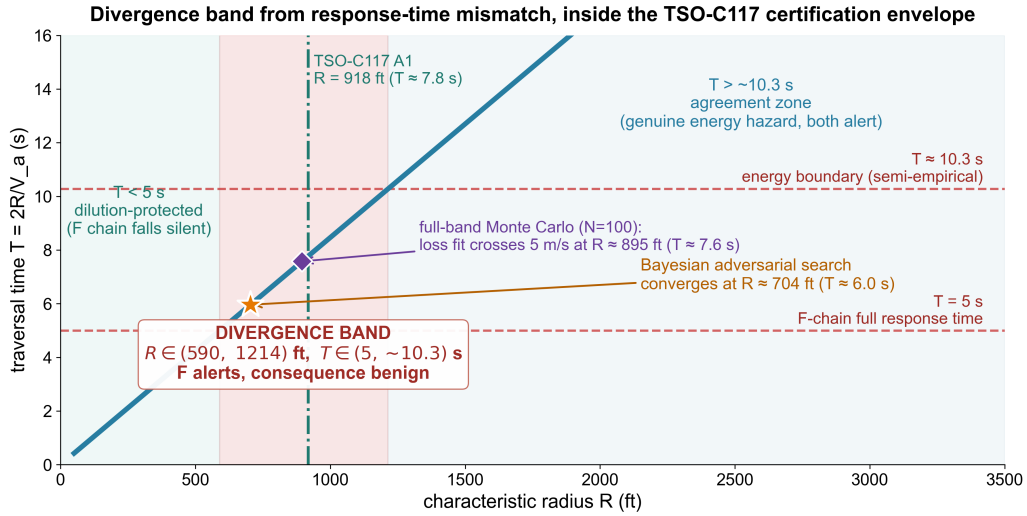


Figure 3. Relationship between the time-window divergence band and the certification envelope. At approach speed $V_a = 72$ m/s, the traversal time $T = 2R/V_a$ grows linearly with the characteristic radius R: $T < 5$ s is the dilution-protected region (the F chain tends to stay silent); $T > \sim$10.3 s is the agreement zone (genuine energy hazard, both alert); between them lies the divergence band $R \in (590, 1214)$ ft. The most compact TSO-C117 certification wind field A1 ($R = 918$ ft, $T \approx 7.8$ s) falls inside the band; the hit centroid of the Bayesian adversarial search, $R \approx 704$ ft ($T \approx 6.0$ s), is also inside.

The existence of the divergence band does not depend on the window choice, and this can be tested systematically along the standard's own criterion across the entire reference window range. The must-alert curve in Appendix A of TSO-C117 is the hyperbola $f_{\mathrm{av}} \cdot t_w \geq 1.05$, equivalent to a threshold of $1.05/t_w$ for each window length t_w; we replay this criterion offline window by window on 863 archived simulated flight records (Figure 4a; t_w takes six values from 5 to 10 s, with signal reconstruction and IIR filtering identical to the current chain). This replay archive is drawn from the same raw simulation pool as Section 2.2 and covers all seven scenario families defined there, but it applies its own minimum-length filtering and additionally retains an early development batch and 18 safe-variant runs that the stricter curation of Section 2.2 excludes, hence the different record count (863 vs. 710). The result is a consistent dilemma: at the short-window end, the 5 s window (threshold: 0.21) alerts only 45% of the dual_danger family, leaving more than half of the genuine hazards inside the dilution-protected region; the alert rates of the hazard families rise monotonically with window length, reaching 100% at 9–10 s windows (thresholds: 0.117–0.105). But along the same hyperbola, the alert rate of the fa_boost family rises in lockstep from 40% at the 5 s window to 89% at the 10 s window—suppressing false alarms and preserving detection trade off along the window axis—and no window length within the standard's reference combinations satisfies both sides. The turb family stays at zero alerts under all six window lengths, showing that this dilemma acts specifically on compact-burst signals rather than being a general signal-to-noise problem. A recomputation with the full state machine reaches the same conclusion (Figure 4b): after the threshold is lowered from the hyperbola value 0.131 to 0.10 at an 8 s window to compensate for dilution, the detection rates of all hazard families are no lower than those of the current 5 s/0.13 chain (98.4% vs. 95.2% for the dual_danger family), the turb family stays at zero, and yet the alert rate of the fa_boost family remains 97%, on par with the current chain; if the threshold stays at the hyperbola value, the detection rate of that family drops to 82%, so false-alarm suppression and detection preservation again cannot be achieved together. The position of the divergence band is set by the window length; its existence is set by the scalar structure of "fixed threshold plus persistence confirmation". This paper reports the quantitative calibration R

$\in (590, 1214)$ ft at the 5 s window; switching to another window length within the reference combinations only shifts the band, it does not remove it.

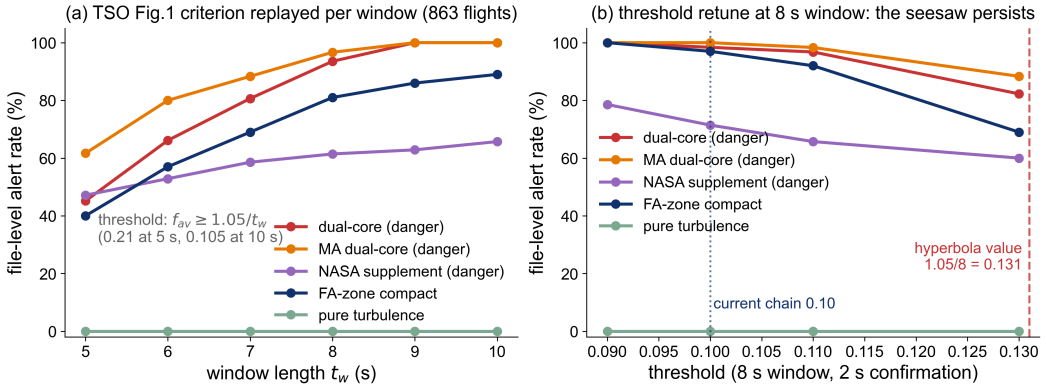


Figure 4. Offline replay test of the window–threshold dilemma on the 863-record replay archive. (**a**) File-level alert rates of the TSO-C117 Appendix A must-alert criterion $f_{\mathrm{av}} \cdot t_w \geq 1.05$ (i.e., threshold $1.05/t_w$ per window length) replayed window by window: the three hazard families rise monotonically with window length to 100% at 9–10 s, but the false-alarm-zone family (fa_boost) rises in lockstep from 40% to 89%, and the turbulence family stays at zero throughout. (**b**) Threshold sweep at a fixed 8 s window with 2 s persistence confirmation (the red dashed line marks the hyperbola reference value: 0.131): lowering the threshold from the hyperbola value 0.131 to 0.10 restores hazard-family detection (dual_danger: 82%→98%), but the false-alarm-zone alert rate rises in lockstep from 69% to 97%; the seesaw holds at every threshold point, and the turbulence family stays silent throughout.

3.2. A Multi-Channel Short-Window Detector Based on a Physical-Consequence Criterion

Section 3.1 has shown that within a scalar-threshold structure, local adjustment of window and filtering can hardly eliminate the false alarms caused by the scale mismatch; the root of the divergence is the decoupling between the signal criterion and the physical consequence of flight. The detector design therefore follows two ideas. One is to match the observation window to the event scale: a 0.5 s window (25 frames) lets the true dynamics of a compact event enter the observation unpolluted by stretching, at the cost of a lower signal-to-noise ratio and higher sensitivity to turbulence. The other is to move the decision basis from signal amplitude to physical consequence: the model predicts the severity level of the forward 5 s airspeed loss, so a harmless strong signal no longer triggers a resolution advisory.

The method of this paper differs from the PINNs paradigm of Raissi et al. [16]: PINNs solve forward or inverse problems of PDEs, whereas in this paper physics enters the network as auxiliary tasks—the model first predicts force residuals on three orthogonal axes as intermediate representations, and the classifier then decides on these representations. Physics here therefore acts only as an inductive bias, not as a hard constraint: the residual outputs are learned approximations, and the network cannot reject outputs that violate physics; whether this structure can satisfy the airworthiness-certification requirements for airborne software is outside the evaluation scope of this paper. The real value of the residual channels lies in training regularization and interpretable attribution; the ablation in Section 4.4 shows that removing them has only a marginal effect on classification performance, and this conclusion also holds at 25% and 50% reduced training-data volumes.

The three residual channels are derived from the decomposition of Newton's second law in the body-axis system and are normalized by the aircraft weight mg so that they transfer across configurations. The longitudinal force residual is

$$r_1 = \frac{1}{mg}\left| m\,(\dot{U} + q\,W - r\,V) - (T - D\cos\alpha + L\sin\alpha - mg\sin\theta)\right|, \tag{4}$$

the vertical force residual is

$$r_2 = \frac{1}{mg}\left| m\,(\dot{W} + p\,V - q\,U) - (-L\cos\alpha - D\sin\alpha + mg\cos\theta\,\cos\phi)\right|, \tag{5}$$

and the energy-rate residual is

$$r_3 = \left|\frac{\dot{V}_c}{g} - \frac{T - D}{mg}\right|, \tag{6}$$

using standard flight-mechanics notation (m is mass; U, V, W are body-axis velocities; p, q, r are the roll, pitch, and yaw rates; T, D, L are thrust, drag, and lift; α is the angle of attack; θ is the pitch angle; ϕ is the roll angle; V_c is the air-relative total speed, i.e., the magnitude of the three body-axis velocities; $\dot{V}_c$ is its time derivative). The three channels complement each other: r_1 is sensitive to horizontal shear. r_2 is sensitive to vertical shear—the downdraft core directly drives the vertical force imbalance, making it the most microburst-sensitive of the three channels. r_3 is sensitive to the energy mismatch between thrust and drag.

In network structure, an encoder (1050→256→128→128, ReLU + Dropout $p = 0.2$) compresses the 25-frame × 42-feature input into a 128-dimensional latent vector. The 42 features fall into six categories: attitude (8), velocity (11), forces and moments (10), atmosphere (4), aircraft state (6), and control surfaces (3); the channel list and the onboard measurability of each category are given in Appendix A. The attitude, velocity, and control-surface categories correspond to parameters recordable on the standard FDR/QAR bus, while the aerodynamic-force and three-dimensional wind-field categories are simulation-internal quantities that cannot be measured directly onboard—this is exactly the motivation for the observable-subspace mapping of Section 3.5. Relative to the F-factor chain, which relies on only a few channels, the multi-sensor input structurally provides redundant information paths: only a few channels depend directly on airspeed measurement, and inertial attitude, ground speed, and aerodynamic state quantities form mutually independent observation sources; quantitative behavior under sensor-loss conditions is outside the evaluation scope of this paper and is left to future work. Three independent residual heads (128→64→1, no parameter sharing) predict the residuals r_1, r_2, r_3 of Equations (4)–(6), respectively; the fusion classifier concatenates the latent vector with the three residuals into a 131-dimensional vector and outputs three-level alert probabilities (OK/TA/RA) through 131→64→32→3 and a Softmax.

For deployment, the certified chain is kept intact and runs independently, and the detector serves as a parallel physical-consequence channel; the two are fused through arbitration—when the F chain triggers an alert while the detector judges the forward physical consequence to be mild (OK/TA level), the alert is downgraded to an advisory; otherwise the original alert level is maintained. The TA and RA names borrow the two-level alerting vocabulary of TCAS [24] and in this paper denote only the severity levels of the forward energy consequence; the presentation form seen by the crew after arbitration depends on the human–machine interface design of the specific aircraft type and is outside the scope of this paper. The in-loop results of Section 4.5 show that this fusion preserves the must-alert behavior on the ten certification wind fields at the file level, with the fused system triggering RA 10/10. The precise meaning of "not touching certified detection behavior" in this paper is therefore as follows: the implementation, thresholds, and output signals of the certified chain itself are not modified, and the alert state of the F chain is preserved frame by frame; what changes is the level presented to the crew by the downstream arbitration layer, which is exactly where the reduction in the false-alarm-zone trigger rate occurs. However, a

file level of 10/10 does not imply that the level of a certification alert stays constant over its entire duration: arbitration is executed frame by frame, and the continuity of levels during the persistence of a certification alert must be rechecked under the standard's own criteria; we list this among the applicability boundaries (Section 5.4).

3.3. Data and Training

The dataset is the 710-file collection of Section 2.2, split at the file level into 568 training files and 142 test files (522,854 frames), with no wind field appearing in both sets. Labels are generated from physical ground truth rather than from the criterion under test: frames whose forward 5 s airspeed loss exceeds 5 m/s are labeled RA, those exceeding 2.5 m/s are labeled TA, and the rest are OK. The 2.5 m/s boundary is half the RA consequence level—an average $\bar{F} \approx 0.05$ over the 5 s horizon, half of the 0.1 hazard reference—so the TA band collects encounters that are energetically noticeable yet clearly below the hazard level. This anchoring is also consistent with the standard's own test procedures, which require the system to remain silent under sustained average shear intensities of $f_{\mathrm{av},x} = 0.02$ and 0.04 lasting 20 s [8]. The 5 m/s value is numerically equivalent to the 0.1 F-factor reference level—losing 5 m/s within 5 s corresponds to an average $\bar{F} \approx 0.10$; the two are numerically equivalent but structurally different: the label takes the true energy consequence within a per-frame forward window and does not pass through the transformations of filtering, sliding-window averaging, and persistence confirmation, so the original temporal structure of the event is fully preserved. The learning target defined this way is decoupled from the F-factor chain and anchored directly to the energy consequence of the encounter. This decoupling concerns the labels and the decision criterion: the 42-dimensional input still contains the two F-derived channels FFactor and WindDetRate, and a retraining ablation removing these two channels is reported in Section 4.4.

Training uses AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-4}), batch size 4096, 60 epochs, random seed 42, gradient clipping 1.0, with the learning rate halved after the validation loss plateaus (PyTorch 2.11, CUDA). The joint loss $L = L_{\mathrm{cls}} + \lambda L_{\mathrm{phys}}$ ($\lambda = 1.0$) combines weighted cross-entropy with Smooth-L1 residual regression (residual ground truth pre-clipped to $[0, 2.0]$); against label imbalance—OK frames account for about 92.8% of the training set—inverse-frequency weighted sampling is used so that the batch distribution is approximately balanced.

3.4. Bayesian Optimization Design for Adversarial Validation

To test whether the divergence band of Section 3.1 is a structural property of the encounter space rather than a product of the training distribution, we run an adversarial search independent of the training data [25,26]. A single simulation takes 30–60 s, so exhaustive enumeration is infeasible; we adopt Bayesian optimization (Gaussian-process surrogate, Matérn-5/2 kernel, EI acquisition function) to search for high-divergence combinations in the space $R \in [600, 5000]$ ft, $V_{ZO} \in [25, 85]$ ft/s, $H_T \in [800, 3500]$ ft—encounters in which the F-factor chain alerts while the physical consequence stays mild, judged by a forward 5 s loss not exceeding 5 m/s (Section 4). The search starts from 50 freshly labeled seeds drawn uniformly over the whole domain, of which 8 fall inside the predicted band and only 1 hits the divergence target, so convergence is not constructed by the initial sampling; 50 BO iterations follow, with a random search of the same budget as the baseline. The seed set and the random baseline also serve as out-of-band controls, so that alert rates can be compared across radius regions.

3.5. Observable-Subspace Mapping and Sampling-Rate Matching for Real Data

Section 5 extends the validation to QAR data of real flights, which requires a methodological preparation step. In the 42-dimensional input of the full-state detector, simulation-internal aerodynamic and propulsive quantities—thrust, drag, lift, moments, and the three-dimensional wind field—predominate, whereas QAR records only standard flight parameters, so the detector cannot be ported directly; this is a constraint of data observability, not of model capability. The treatment in this paper is to map the detector onto the QAR-observable subspace: eight features that can be computed by exactly the same formulas in simulation and in real data are selected—true airspeed, vertical speed, indicated airspeed and its 0.4 s difference, normal load factor, and radio altitude—plus two quantities derived from the vertical-wind estimate $w_h = -V_d - U_{\text{aero}} \cdot \sin(\theta - \alpha)$: a dead-band F-factor $|w_h|/U_{\text{aero}}$, set to zero when its value is below 0.01; and a wind-detection rate, the 0.4 s negative difference of the exponentially smoothed w_h, clipped to ± 3 m/s^2. Five statistics (mean, standard deviation, minimum, maximum, trend) of a 1 s sliding window over these eight features form a 40-dimensional input; the network and the three physical residual heads remain isomorphic, with only the input dimension reduced accordingly. Using statistics instead of derivatives also makes the mapping naturally robust to sparse sampling: on the same batch of QAR data, the peak F-factor computed by the classical definition is only 0.107, already below the 0.13 alert threshold, so derivative-type criteria are unusable under this sampling condition.

The second prerequisite is sampling-rate matching. QAR records are aggregated by the second (equivalent to 1 Hz) and then restored by spline interpolation, and their feature scales differ systematically from the 25 Hz simulation data: for the same physical signal, the median airspeed-difference peak is 2.96 m/s^2 on the simulation side and only 0.63 m/s^2 on the QAR side. The simulated training features are therefore degraded by the same second-level aggregation plus spline interpolation before the statistics are computed—this is a pure sampling-theory operation that introduces no real data. The mapped model is still trained only on the 568 simulation files, with label definitions, thresholds, and the file-level split identical to the full-state model; all real data appear only in the validation of Section 5. The processing chain is causal with respect to the alert timeline: every feature is a trailing 1 s window computed from the current and the 24 preceding samples only, the 0.4 s differences look backward, and the window statistics are aligned to the window end, so no future information enters the detector input at any stage. The spline interpolation used to restore a uniform sampling grid is an offline reconstruction applied identically to the detector input channels and to the GPWS alert timeline; it preserves the relative timing between the two, and the lead times reported in Section 5 are therefore measured quantities, not look-ahead artifacts. On real records the three residual heads of Section 3.2 can no longer be evaluated against ground truth, because the force terms of Equations (4)–(6) are not recordable onboard; on the QAR side they therefore act only as a structural prior carried by the trained weights, and only the classification output is read.

4. Results and Discussion

4.1. Large-Scale Extended Monte Carlo Validation

The 142 files of the held-out test set offer limited resolution for rate estimation, and boundary-hugging conclusions such as zero triggering in the false-alarm zone especially need larger samples. This section extends the validation to Monte Carlo scenarios from the same distribution as the training data but entirely resampled, in two batches run under an identical protocol: an initial batch of 3500 scenarios, and—in response to the reviewers' request for larger samples and seed-sensitivity analyses—a further 11,400 scenarios sampled anew under the same distribution, including 400 scenarios of a dedicated family drawn

uniformly inside the divergence band itself. The seven scenario families of Section 2.2 regenerate wind fields by random sampling within their respective parameter ranges, and the sub-threshold gust scenarios (headwind/tailwind bands of 2–5 m/s amplitude) serve as a second benign family beyond turbulence. The aircraft is fixed at a nominal mass of 260 t and an approach speed of 75 m/s (the sensitivity to both is quantified separately in Section 4.7); the detector is not retrained or adjusted in any way. The F-factor chain and the detector run in parallel on every scenario, and the file-level decisions of both methods and the forward 5 s airspeed-loss ground truth are recorded scenario by scenario. Of the 14,900 scenarios, 75 are removed (simulation timeouts or airspeed losses beyond the physically possible range, judged as numerical-integration divergence rather than genuine encounters), leaving 14,825 valid samples—about 58.6 million flight frames, roughly 22 times the frame count of the training data. Table 1 gives the file-level paired results for the nine families.

Table 1. Scenario-by-scenario paired results of the extended validation on 14,825 scenarios (file level; intervals are Wilson 95% confidence intervals).

Scenario Family	N	F-Chain Alert %	Detector RA Trigger %	Cases with True RA	F-Chain Detected	Detector Detected
False-alarm zone (compact)	3198	99.2 [98.8, 99.4]	8.0 [7.1, 9.0]	51	51	51
Divergence band (targeted)	400	94.8 [92.1, 96.5]	42.5 [37.7, 47.4]	137	137	137
Single-core microburst	1772	71.5 [69.4, 73.6]	69.1 [66.9, 71.2]	1144	1103	1142
NASA single-core supplement	1755	72.1 [70.0, 74.2]	74.8 [72.7, 76.8]	1225	1160	1225
Dual-core microburst	1800	96.4 [95.5, 97.2]	71.8 [69.7, 73.8]	1185	1185	1183
MA dual-core	1800	100.0 [99.8, 100.0]	85.1 [83.3, 86.6]	1439	1439	1431
Pure turbulence	1400	0.0 [0.0, 0.3]	0.0 [0.0, 0.3]	0	—	—
Sub-threshold gust	900	0.0 [0.0, 0.4]	0.0 [0.0, 0.4]	0	—	—
TSO certification family	1800	99.6 [99.2, 99.8]	98.7 [98.0, 99.1]	1707	1707	1707

On the hazard side, the six hazard-containing families give 6888 scenarios containing true RA frames. The detector detects 6876 of them (99.83%); the 12 misses are all boundary encounters whose maximum forward losses lie between 5.01 and 5.47 m/s, and on every one of them the F chain alerts in the same scenario, so the encounter is still covered after arbitration; the F chain misses 106 (98.5% on its own), and all 106 misses are independently detected by the detector. The two criteria are strictly complementary on brand-new samples: as a classifier-level complementarity analysis (not a deployed behavior—the deployment of Section 3.2 only ever downgrades, never adds, an alert), the file-level detection rate of their union is 100% over all 6888 hazardous scenarios. Neither side misses any of the 1707 certification-family scenarios containing genuine hazard; the must-alert property is preserved at large sample size. The detector achieves 100% file-level detection on four families—NASA supplement, the certification family, the false-alarm zone, and the targeted divergence band—reproducing the held-out conclusion of zero missed hazardous files at more than 100 times the sample size.

The F-chain-only column of Table 2 concentrates in the false-alarm zone (2915 cases) and in the sub-RA encounters of the dual-core and MA dual-core families (442 and 261 cases). The latter two contain no true RA frames under the 5 m/s criterion; the F chain alerts them at 90% and 100%, respectively, while the detector triggers only 18% and 28%—the systematic over-alerting of the certified criterion on sub-RA encounters is confirmed at large sample size, shares the same time-window mismatch mechanism as the false-alarm zone, and the detector gives consistent partial suppression of every manifestation of this mechanism.

On the benign side, both criteria stay silent throughout the 1400 pure-turbulence cases, and the detector does not output a single frame at TA level or above. The 900 sub-threshold gust cases have a median forward loss of 1.71 m/s and a maximum of 4.12 m/s, none crossing the 5 m/s hazard threshold, and both the F chain and the detector stay at zero RA alerts throughout this family as well; the detector outputs TA-level decisions on 420 cases, which

under the arbitration semantics of Section 3.2 correspond to advisory-level cues, not alerts. On the 3198 false-alarm-zone cases, the F chain alerts at 99.2% and the detector triggers RA at 8.0%; the main false-alarm-suppression conclusion holds at nearly 250 times the number of held-out false-alarm-zone files, but the trigger rate is not the zero suggested by the small held-out sample. Table 3 and Figure 5 show where the residual triggers lie: of the 256 residual RA triggers, none lies below $R = 704$ ft, and 246 lie in the upper-edge stratum with R above 895 ft. The median forward loss in this stratum has already reached 4.06 m/s, close to the 5 m/s hazard threshold, and all 51 genuine hazards of the false-alarm zone fall in this stratum and are all detected by the detector. The residual triggers concentrate on the boundary where the divergence band transitions to genuine hazard; they belong to the inherent ambiguity of the criterion near the threshold, not to a failure at the core of the false-alarm zone. The zero triggering on the 13 held-out files should be read as small-sample fluctuation: at the 6.5% no-hazard trigger rate measured in this section, the probability of observing zero triggers on 13 files is about 42%. The detector additionally outputs TA-level (advisory) decisions without RA on 2445 of the 3198 false-alarm-zone cases; under the arbitration semantics of Section 3.2 these correspond to alerts downgraded to advisories rather than complete silence, consistent with operational expectations.

Table 2. Cross-distribution of the paired four quadrants against physical ground truth (file counts).

False-alarm zone	True RA	51	0	0	0
False-alarm zone	No true RA	205	2915	0	27
Divergence band (targeted)	True RA	137	0	0	0
Divergence band (targeted)	No true RA	33	209	0	21
Single-core microburst	True RA	1101	2	41	0
Single-core microburst	No true RA	46	118	37	427
NASA single-core supplement	True RA	1160	0	65	0
NASA single-core supplement	No true RA	30	76	58	366
Dual-core microburst	True RA	1183	2	0	0
Dual-core microburst	No true RA	109	442	0	64
MA dual-core	True RA	1431	8	0	0
MA dual-core	No true RA	100	261	0	0
Pure turbulence	No true RA	0	0	0	1400
Sub-threshold gust	No true RA	0	0	0	900
TSO certification family	True RA	1707	0	0	0
TSO certification family	No true RA	69	17	0	7

Table 3. Stratified statistics of the 3198 false-alarm-zone cases by characteristic radius.

***R* Stratum (ft)**	**N**	**F-Chain Alerts**	**Detector RA**	**Detector TA Only**	**Median Loss (m/s)**	**Max Loss (m/s)**	**Cases with True RA**
(680, 760]	963	936	0	646	1.91	2.95	0
(760, 840]	916	916	0	736	2.83	3.74	0
(840, 895]	639	639	10	629	3.59	4.24	0
(895, 950]	680	680	246	434	4.06	7.62	51

4.2. Monte Carlo Validation of the Divergence Band

To test the core prediction of Section 3.1, we generated 400 independent microburst scenarios covering the entire predicted band under the Bray wind-field model: R sampled uniformly within [590, 1214] ft, $V_{ZO} \in [68, 78]$ ft/s, $H_T \in [1000, 1400]$ ft, with the wind field, configuration, and physical records of each case archived individually. Table 4 gives the results of the F-factor chain: 379 trigger WARNING, 21 stay silent. The adjudication criterion is the same as the detector labels—a forward 5 s airspeed loss not exceeding 5 m/s. Of the 379 triggers, 242 are low threat, with a median loss of 2.81 m/s, an interquartile range

of 1.38–3.86 m/s, and a maximum of 4.98 m/s; 137 are genuine hazards, with a median loss of 10.73 m/s and a maximum of 15.44 m/s. In the low-threat group, the only quantity exceeding its limit is the F-factor itself, peaking at 0.14–0.26; the decoupling between the alert basis and the physical consequence is directly visible here. All 21 silent cases lie at the lower in-band edge ($R \leq 704$ ft, 81% of them below 640 ft), with losses not exceeding 0.99 m/s and F peaks of 0.12–0.14, sitting exactly near the threshold and exhibiting the soft-boundary shape of the 5 s window's dilution-protected region.

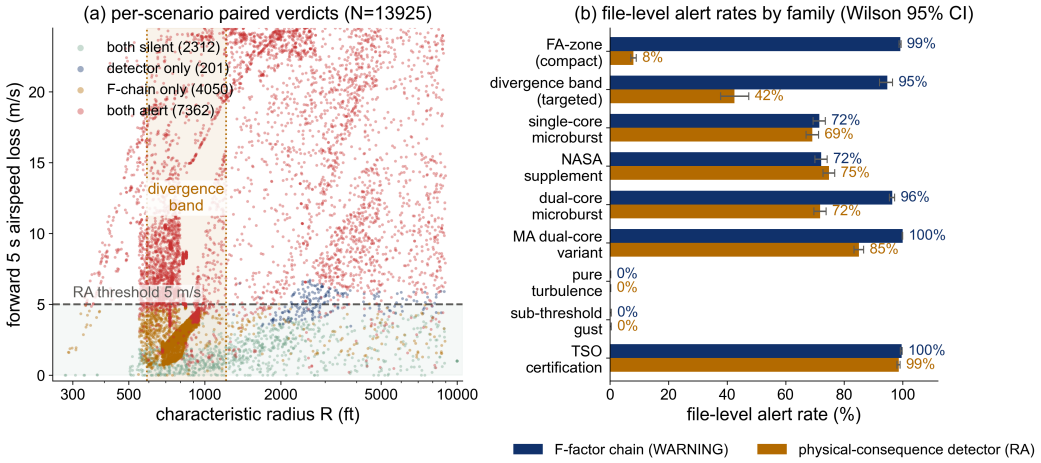


Figure 5. Paired results of the extended validation on 14,825 cases. (**a**) Scenario-by-scenario distribution of characteristic radius versus forward 5 s airspeed loss, colored by the paired-decision quadrants: within the false-alarm zone (fa_boost sampling region $R \in [680, 950]$ ft), F-chain-only points concentrate below the 5 m/s loss, and both-alert points appear along the rising-loss direction at the upper edge; large-radius scenarios of the hazard families generally cross the 5 m/s threshold and are dominated by both-alert, and the 201 detector-only alerts split into 106 genuine hazards missed by the F chain and 95 sub-threshold encounters. (**b**) Paired file-level trigger rates of the nine scenario families (error bars are Wilson 95% confidence intervals): the false-alarm zone shows the largest contrast between the two methods (99.2% vs. 8.0%), the pure-turbulence and sub-threshold-gust families stay silent on both sides throughout, and the certification family triggers nearly fully on both sides.

Table 4. Monte Carlo validation of F-factor alerts (N = 400, R sampled uniformly within [590, 1214] ft, $V_{ZO} \in [68, 78]$ ft/s, $H_T \in [1000, 1400]$ ft).

Outcome	Count	Rate
F-factor chain triggers WARNING	379	94.8%
Confirmed false alarm (low threat)	242	60.5%
True positive	137	34.2%
F-factor chain silent	21	5.2%

The in-band loss distribution simultaneously calibrates the two boundaries of the divergence band. The loss rises monotonically with R (correlation coefficient 0.90), and the linear fit crosses 5 m/s at $R \approx 883$ ft ($T \approx 7.5$ s); grouped by R, none of the 207 cases with $R < 900$ ft exceeds the limit, all 72 cases with $R > 1100$ ft exceed it, and the group with the largest R (beyond 1150 ft, near the 1214 ft upper bound) already has a median loss of 12.7 m/s. In other words, the lower half of the divergence band ($T \lesssim 7.5$ s) is the core false-alarm region; the closer to the upper bound, the more an alert tends to be a genuine hazard, and beyond the upper bound one enters the agreement zone—the semi-empirical upper bound of about 10.3 s in Section 3.1 is calibrated from this. On the fa_boost family sampled directionally at the band center, the false-alarm rate reaches 99.2%, reported together with

the 60.5% of uniform in-band sampling—the former corresponds to the worst case, the latter to the band average. The forward 5 s loss of the false-alarm-zone family has a median of 3.01 m/s and a maximum of 7.62 m/s: 1.6% exceed the 5 m/s adjudication threshold and only 0.03% exceed a 50%-loosened 7.5 m/s, while a 50%-tightened 2.5 m/s—the TA level—is exceeded by 64.1%; most encounters in this family are energetically noticeable yet clearly below the hazard level, consistent with the two-level TA/RA alerting structure of Section 3.2.

4.3. Bayesian Adversarial Search

Of the 50 BO iterations, 35 hit the divergence target, a discovery rate of 70%; the false-alarm points converge to $R \in [602, 1247]$ ft with a centroid of $R \approx 704$ ft, $T \approx 6.0$ s, consistent with the theoretical band (Figure 6). The random search of the same budget finds zero, but this control does not constitute independent evidence for the existence of the divergence band: a random hit requires the radius to fall in the band and the intensity to cross the threshold simultaneously; as a reference, of the 50 whole-domain uniform initial seeds of BO, only 8 fall in the band and 1 hits; at this hit rate, the probability of zero discoveries in 50 points is about 36%, which is not surprising, and the independent existence evidence is carried by the uniform in-band Monte Carlo of Section 4.2. The value of the BO point set lies elsewhere: it is the only adversarial sample set in this paper sampled outside the training data, and the search process reads no training file; its convergence region overlaps the sampling region of the fa_boost family in the training set ($R \in [680, 850]$ ft), so strictly speaking it is independent of the training data, not of the training distribution. The detector's RA trigger rate on these 35 false-alarm points is zero—18 stay silent throughout and 17 reach only TA level; the same search also encounters nine encounters whose physical consequence exceeds the limit while the F chain does not alert, and the detector triggers RA on all nine. False-alarm suppression and hazard discrimination hold outside the training data as well. The out-of-band control further shows that the F-factor alert rate rises with R—11.8%, 48.6%, and 87.5% in the band, the transition region, and the far region, respectively—but confirmed false alarms concentrate in the predicted band, accounting for 5.9% of in-band samples, and disappear entirely in the far region: the divergence is band-specific, not a general property of strong encounters. Note that the two false-alarm rates have different denominators: the 5.9% here takes all in-band samples as the denominator, with V_{ZO} uniform over the whole [25, 85] ft/s domain and most samples too weak to trigger an alert; the 60.5% of Section 4.2 comes from in-band scenarios with V_{ZO} locked to the medium-high intensity range [68, 78] ft/s.

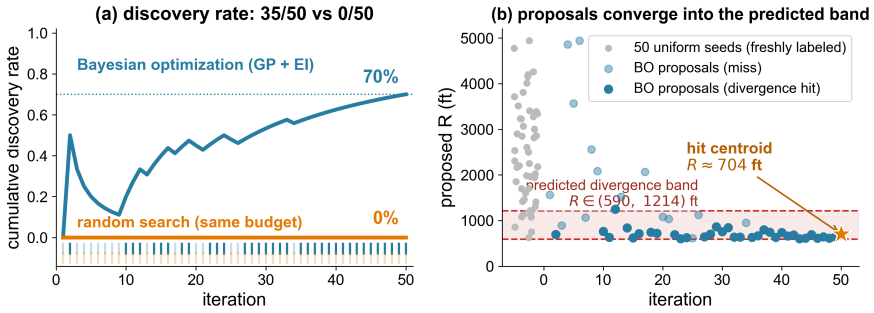


Figure 6. Convergence behavior of the Bayesian adversarial search. (**a**) Cumulative discovery rate: 35 divergence targets hit in 50 iterations (70%), zero for the same-budget random search. (**b**) Distribution of GP-proposed radii over iterations: gray points are the 50 whole-domain uniform initial seeds (plotted at iteration 0), blue points are BO proposals; hitting proposals (dark) converge into the predicted divergence band, with hit radii in [602, 1247] ft—individual points slightly exceed the 1214 ft upper bound, consistent with its semi-empirical nature—and a hit centroid of $R \approx 704$ ft.

For the revision, all 50 BO iterations were additionally re-flown from their archived (R, V_{ZO}, H_T) parameters with per-frame physics records retained, so that the search results could be re-adjudicated under the third-party consequence criteria of Section 4.6 rather than only under the detector's own label definition. The re-flight reproduces the F-factor alert on 48 of the 50 iterations (the remaining two sit at the soft boundary of the dilution-protected region), and the re-run forward losses correlate with the original BO records at 0.996, confirming that the archived search results are reproducible. The third-party re-adjudication itself is reported in Section 4.6.

4.4. Detection Performance and Ablation

Figure 7 compares file-level trigger rates on the 142 held-out files across the seven scenario families, taking WARNING-level alerts for the F chain and RA-level outputs for the detector, with Wilson 95% confidence intervals whose effective sample size is the file count. On the 13 files of the fa_boost false-alarm family, the F-chain WARNING trigger rate is 92.3% [66.7, 98.6] and the detector RA trigger rate is 0%, with not a single frame triggered among 53,313 frames; ten in-loop configurations newly generated from the same region also confirm this, with an average RA-frame fraction of 0.03%. On the 26 files of the dual-core hazard family, the two rates are 92.3% [75.9, 97.9] and 53.8% [35.5, 71.2], respectively. This family has 12 files that contain true RA frames, and the detector detects all of them, giving a file-level recall of 12/12; the other 2 files on which the detector triggers RA, like the remaining 12 untriggered files, contain no true RA frames under the 5 m/s criterion and are sub-RA encounters rather than RA-level events. The higher trigger rate of the F chain here comes partly from over-alerting exactly this kind of sub-RA encounter—the same mechanism as the false-alarm zone. On the 21 pure-turbulence files, both channels stay silent. The 24 files of the certification family are Monte Carlo variants generated from the parameters of the ten certification wind fields as the parent population via bootstrap resampling plus parameter jitter; on this family the two methods tie at 87.5%, i.e., 21/24 files, but the untriggered files differ: of the 3 files on which the F chain stays silent, 2 are benign variants in weak configurations containing no true RA frames, and 1 contains 196 true RA frames yet is missed; the 3 files on which the detector stays silent all contain no true RA frames, and one of them is exactly the file that the F chain false-alarmed on—so, on this family the detector additionally corrects one miss and one false report of the F chain. The 10/10 of Section 4.5 is the in-loop reproduction of the ten certification wind fields in their standard configurations, not the same set as the 24 Monte Carlo variant files here.

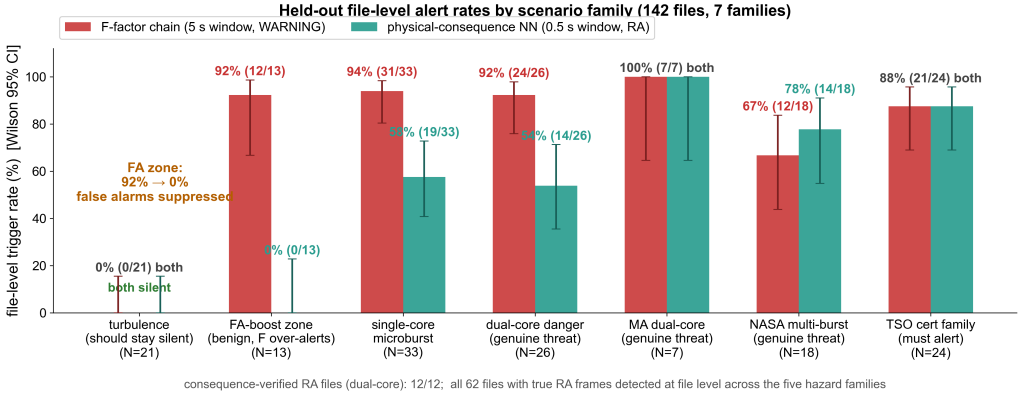


Figure 7. File-level trigger-rate comparison on 142 held-out files across seven scenario families: the F-factor chain (5 s window, WARNING level) versus the physical-consequence detector (0.5 s window, RA level). Error bars are Wilson 95% confidence intervals with the file count as the effective sample size; the mapping between family codes and names is given in Section 2.2.

On the 522,854-frame held-out set, the overall accuracy is 97.6%, and the balanced accuracy—the equal-weighted mean of the per-class recalls—is 96.9%, with OK, TA, and RA at 0.977, 0.967, and 0.963, respectively, and an RA precision of 0.844. The fraction of OK frames falsely reported as RA is 0.82%, i.e., 3949 of 480,827 frames, rising to 2.3% when TA-level false reports are included—the false-report cost of the short window concentrates at the TA advisory level rather than the RA decision level. Per-family frame-level RA recalls are as follows: dual-core: 98.9%, ma_dual: 98.6%, nasa_supp: 97.0%, tsosupp: 97.1%, single-core microburst: 91.3%; the predicted RA-frame fraction on fa_boost and turbulence is 0.0% for both. The frame-level gap concentrates in the single-core family, whose 33 test files contain 4353 true RA frames, and all 16 of its files containing true RA are detected at the file level; the five hazard families together have 62 files containing true RA frames (12 dual-core, 7 ma_dual, 14 nasa_supp, 13 tsosupp, 16 single-core), and none is missed—file-level recall 100%. The correlation coefficients between predicted and true residuals are 0.931/0.943/0.949 ($r_1/r_2/r_3$); the attribution of RA frames is dominated by the vertical channel at 52%, followed by the energy-rate channel at 41% and the longitudinal channel at 7% (Figure 8).

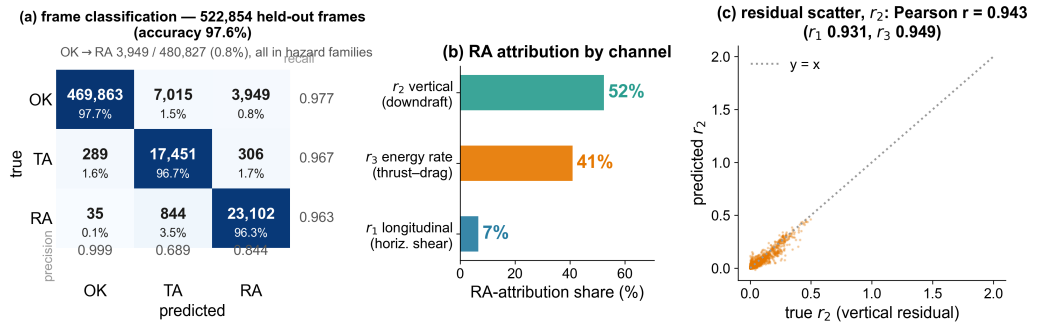


Figure 8. Classification performance and residual attribution on the 522,854-frame held-out set. (**a**) Three-level confusion matrix: overall accuracy: 97.6%; the fraction of OK frames falsely reported as RA is 0.8% (3949 frames, all in hazard families). (**b**) Residual-channel attribution of RA frames: the vertical channel r_2 accounts for 52%, the energy-rate channel r_3 for 41%, and the longitudinal channel r_1 for 7%. (**c**) Scatter of predicted versus true residuals for the r_2 channel, Pearson $r = 0.943$ (0.931 and 0.949 for r_1 and r_3); the dotted diagonal is the $y = x$ reference of perfect prediction.

Two ablation groups separate the roles of window length and the physics auxiliary tasks. Widening the input window to the F-chain scale of 250 frames reduces the tsosupp recall from 97.1% to 94.4% and dual-core from 98.9% to 93.6%, while the false-alarm-zone silence is unchanged—a long input window harms the detection of compact short-duration events and contributes nothing to false-alarm suppression. This shows that the value of the short window lies in resolution and scale matching; the suppression of false alarms does not come from the window length. Removing the residual heads leaves performance essentially unchanged, showing that their role is training regularization and interpretable attribution, consistent with the inductive-bias positioning of Section 3.2. A third ablation removes the two F-derived input channels, FFactor and WindDetRate, leaving 40 input channels, testing whether the detector's decision is decoupled from the F-factor criterion not only at the label level but also at the input level. With these channels removed, the overall accuracy is 97.4% and the RA recall 96.0%; per-family recalls stay within 0.9 percentage points of the main model (dual-core: 98.0%, tsosupp: 96.3%), and the predicted RA-frame fraction on the false-alarm-zone and turbulence families remains 0.0%. The detector therefore does not rely on recomputing the F signal internally: suppression of the false-alarm zone survives the removal of every input channel derived from the F-factor chain. Full metrics of the three ablation variants are given in Table 5.

Another ablation group tests whether the residual heads bring regularization benefits when training data are scarce: the training files are randomly reduced to 25% and 50%, with all other settings identical to the main model, and two variants with and without residual heads are trained at each data volume. At the 25% level, the variant with residual heads has a 0.7-percentage-point higher RA recall (94.0% vs. 93.3%); at the 50% level the relation reverses (95.0% vs. 95.4%); the false-alarm-zone silence is unchanged at both levels. The directions are inconsistent and the magnitudes lie within single-seed fluctuation; the residual heads have no defensible data-efficiency benefit, and this paper keeps them for training regularization and interpretable attribution, not for classification performance. This result also supports locating the physicality of the network in the consequence labels and the architectural prior, rather than treating the residual heads as a source of performance.

Table 5. Ablation comparison on the held-out set. (thr5 is the main model; per-family columns are frame-level RA recall. thr5 and norest5 use a 25-frame input window: of 715 raw files, 710 remain valid after cleaning and minimum-length filtering, split 568/142, with 522,854 test frames. The 250-frame input window of w250t5 requires at least 260 frames after cleaning; in the independent pipeline 704 files are valid, split 563/141, and its test frame count is 497,656 by Σ(cleaned frames − 249)—not the same file set as 522,854, so the two cannot be directly subtracted.)

Variant	Change vs. Main Model	Overall acc.	RA Precision	RA Recall	Dual-Core	ma_dual	nasa_supp	tsosupp	Single-Core	FA/Turbulence RA Frames
thr5 (main model)	—	97.6%	84.4%	96.3%	98.9%	98.6%	97.0%	97.1%	91.3%	0.0%
w250t5	input window 25→250 frames	97.9%	84.8%	93.8%	93.6%	97.1%	96.2%	94.4%	87.6%	0.0%
norest5	physics residual heads removed	97.7%	83.4%	96.2%	98.2%	98.0%	96.0%	97.0%	93.4%	0.0%
nof5	FFactor/WindDetRate channels removed (42→40 inputs)	97.4%	80.7%	96.0%	98.0%	97.7%	96.9%	96.3%	91.9%	0.0%

To quantify training randomness, the main model was retrained under three further seeds (0, 1, 2) in addition to the original seed 42, with the file split and all hyperparameters identical; only the weight initialization and the mini-batch sampling order change. Across the four seeds, the overall accuracy spans 97.4–97.9% (mean: 97.6%), the RA recall spans 96.1–96.6% (mean: 96.3%), and the RA precision spans 82.2–85.5% (mean: 84.0%); per-family RA recalls vary by at most 1.2 percentage points (nasa_supp: 96.1–97.3%), and the predicted RA-frame fraction on the false-alarm-zone and turbulence families is 0.0% for every seed. All quantitative conclusions of this section, including false-alarm-zone silence, therefore fluctuate by no more than about half a percentage point under retraining, with RA precision the most sensitive metric (3.3 points).

4.5. In-Loop Behavior on Certification Wind Fields and Operational Cost

All ten TSO-C117 certification wind fields alert in the loop: the F-factor chain triggers WARNING 10/10, the physical-consequence channel triggers RA 10/10 (RA-frame fractions: 5.0–51.3%, maximum forward 5 s losses: 7.1–42.0 m/s), and the fused system triggers RA 10/10. On A1, the detector outputs TA-level probability about 12 s before the encounter and escalates to RA at the peak, earlier than the alert confirmation of the F chain. In C1/C2/D1, the aircraft touches down in the no-recovery mode—an inherent property of these wind fields without crew intervention, not an alerting failure.

A frame-by-frame continuity recheck of the arbitration levels during the persistence of certification alerts was also carried out for this revision. In the raw arbitration, the presented level can briefly oscillate within one alert episode when the detector output crosses the RA threshold intermittently: over the 1732 alert episodes of the certification-supplement family, 466 (26.9%) contain at least one downgrade–restore oscillation, with the shortest WARNING segment lasting 0.08 s. A monotone episode rule eliminates this by construction: within one F-chain alert episode the presented level is latched and may only escalate, never downgrade, until the episode ends. Under this rule, 98.1% of certification-family episodes remain at WARNING level throughout, and 97.9% of false-alarm-zone episodes are presented as

advisory-only from start to end, with zero oscillation in both families. The rule changes only the presentation of levels within an episode, not the underlying per-frame decisions, so all rates reported in this paper are unaffected.

Weighting false reports and misses by relative cost, $R(k) = P_{\mathrm{FA}} + k \cdot P_{\mathrm{miss}}$ (k being the cost of missing a genuine hazard relative to a false report), the held-out set gives $P_{\mathrm{FA}} = 0$ in the false-alarm zone versus 92.3% for the F chain, with frame-level miss rates of 1.1–8.7% whose upper end is the single-core microburst family, and 100% file-level recall of files with genuine hazard. The cost of missing one genuine microburst far exceeds that of one unnecessary go-around [27]. Because P_{FA} here is at the file level while P_{miss} is at frame level, and the frame-level miss rate of the baseline F chain is not calibrated, we state the conclusion restrictively: the detector dominates on the P_{FA} term within this simulation evaluation domain, while a like-for-like comparison of the P_{miss} term requires first calibrating the baseline miss rate at frame level; the strict validity range of any dominance claim under $R(k)$ is set by the crossover k^* obtained after that calibration, which this paper does not undertake. The value of k in real operations depends on operator policy and must additionally account for the residual RA activity in the alert-free set of Section 5 and the downgrade risk of Section 5.3; this paper draws no conclusion on the operational value of k. Frame-level misses concentrate on marginal frames whose downdraft and F amplitudes are about half those of detected frames—a blind spot inherent to the short window. The network has about 350k parameters and about 350k multiply-accumulates (MACs) per frame, and it already runs online in the loop at the 50 Hz step in the C-implemented simulation loop; this compute scale leaves ample margin on modern airborne processors, and the embedded latency on the target platform will be measured once the platform is determined.

4.6. Re-Adjudication Under Third-Party Consequence Criteria

The adjudication used so far—a forward 5 s airspeed loss beyond 5 m/s—is homologous with the detector's training labels, so a false-alarm conclusion measured under it is not fully independent of the training target. To close this circularity, every valid scenario of the extended validation was re-adjudicated under four further consequence criteria that share no definition with the training labels: (C2) the transient excess altitude loss within a forward 5 s window, relative to the nominal 3° glidepath sink rate; (C3) the transient specific-energy loss over the same window; (C4) the minimum radio altitude within a forward 5 s window (ground proximity); and (C5) the sustained post-encounter specific-energy deficit, measured against the glidepath extrapolated from the encounter start and evaluated at the earlier of 5 s after the encounter end or the first crossing of 100 m AGL, so that the terminal level-off does not contaminate the measure. C1 denotes the homologous airspeed-loss criterion for reference.

On the 3198 false-alarm-zone scenarios the picture is consistent. Under C1, 1.6% (51 cases) are hazardous. The transient criteria fire almost universally: 98.2% of cases exceed 40 m of transient excess altitude loss (C2) and all cases exceed 20 m of transient specific-energy loss (C3). This is not a defect of the scenarios but the physics of the encounter: a microburst traverse necessarily trades altitude and energy before any crew or system response can act, so a transient-excursion criterion cannot separate a benign compact encounter from a hazardous one—the divergence band lives precisely in this window. The sustained and proximity criteria, which separate consequence from mere excursion, are unanimously benign: no case descends below 60 m AGL within the window (C4), and no case retains a post-encounter specific-energy deficit beyond 20 m (C5; median −14.3 m, i.e., the aircraft typically recovers to above the extrapolated glidepath once the shear has passed). Under every criterion that survives the encounter itself, the false-alarm-zone scenarios are

benign, and the suppression conclusion of Section 4.1 does not depend on the homologous label definition.

The same re-adjudication was applied to the 50 re-flown Bayesian-search iterations of Section 4.3, which the search selected as divergent (F chain alerts, detector silent). Of these, 48 reproduce the F-factor alert on re-flight; 12 exceed 5 m/s under C1, 42 exceed 40 m under C2, 1 descends to 2.9 m AGL under C4, and 6 retain a sustained deficit beyond 20 m under C5; 44 of the 50 satisfy all sustained and proximity benignity conditions. The search therefore finds predominantly benign divergent encounters, but not exclusively so: an adversarial search optimizing for criterion divergence occasionally lands on genuinely hazardous encounters, and the one case reaching 2.9 m AGL is a reminder that the band boundary is soft. We report this composition as measured, rather than claiming that every discovered divergence point is harmless.

4.7. Sensitivity to Approach Speed and Aircraft Mass

All results above are measured at the nominal approach point (75 m/s, 260 t). Two one-factor sensitivity batches quantify what happens off this point, each comprising 100 scenarios sampled uniformly inside the divergence band ($R \in [590, 1214]$ ft, $V_{ZO} \in [68, 78]$ ft/s, $H_T \in [1000, 1400]$ ft), with the band family of Section 4.1 (400 cases at 75 m/s, 260 t) as the baseline. Table 6 gives the results.

Table 6. One-factor sensitivity inside the divergence band (100 scenarios per configuration; baseline row from the 400-case band family).

Configuration	Actual Speed (m/s)	Mass (t)	F-Chain Alert %	Detector RA %	True-Hazard %	Median Loss (m/s)
Baseline (band family)	76.6	260	94.8	42.5	34.3	3.81
Slow approach	69.1	260	100.0	100.0	23.0	2.02
Fast approach	86.7	260	91.0	16.0	44.0	4.54
Light	76.6	240	94.0	0.0	36.0	3.98
Heavy	76.6	280	99.0	0.0	34.0	3.96

Approach-speed shifts behave as the time-scale analysis predicts: a slower crossing lengthens the traversal time $T = 2R/V_a$ and moves encounters out of the band toward the agreement zone, so the detector's RA rate rises to 100% at 69 m/s; a faster crossing shortens T toward the dilution-protected region and the RA rate falls to 16%. The F chain is far less speed-sensitive (91–100%), because its 5 s window already integrates over the whole encounter at these speeds. The genuine-hazard fraction under C1 moves only modestly with speed (23–44%), and the speed-driven change in the detector rate is in the conservative direction wherever the two disagree.

Mass is a harder boundary. At ±20 t the detector falls silent throughout the band, including on the 34–36% of scenarios that are genuine hazards under C1: the training corpus contains only 260 t airframes, and the network has learned mass-specific feature scales. This is consistent with the jitter probe reported in Section 5.4 (detection collapses beyond about ±5 t), now quantified inside the band itself. Mass is an input channel of the network, so the deficiency is a data-coverage problem rather than an architectural one, and it is directly remediable: we retrained the detector from scratch with 684 additional scenarios whose masses are sampled uniformly in 230–290 t appended to the training set, keeping the all-260 t test set unchanged (overall accuracy: 97.2%, per-family RA recalls: 93.6–98.5%, false-alarm and turbulence families still 0.0%). On the same ±20 t band configurations, hazard detection recovers from 0/70 to 70/70; the RA trigger rate is 43% at 240 t, while at 280 t the retrained model alerts on every scenario—detection is restored at the cost of discrimination, and the residual failure direction is conservative. Under the arbitration semantics of Section 3.2, an over-triggering detector never adds alerts, it merely refrains from downgrading, so the

deployed system falls back to the certified chain's behavior, which itself alerts at 94–99% here. Approach-speed sensitivity is essentially unchanged by this retraining (12% versus 16% at 87 m/s), since only mass was varied. Until a mass-jittered model is validated at the same depth as the main model, the detector should still be operated as a per-mass configuration. The F chain, being a normalized signal criterion, is essentially mass-insensitive (94–99%).

5. Cross-Validation on Real QAR Data

5.1. Validation Protocol and Data

The validation of the previous sections was completed inside the simulation loop; this section moves to quick access recorder (QAR) data of real flights and cross-validates the mapped model of Section 3.5, whose training and hyperparameter selection used no real data at all. Statistical-learning and machine learning analyses of QAR data have been systematically reviewed in aviation safety [28,29]. The validation set comes from 2411 approach-phase QAR records, of which 65 rows are exact duplicates; after removal they correspond to 2346 flights, and the statistics below are deduplicated by flight. A total of 360 records triggered the onboard GPWS windshear alert during descent, involving 354 flights and forming the positive side; the QAR discretes do not distinguish alert levels. Taking the alert onset as the center, a window from 90 s before to 45 s after is taken and aligned by onboard timestamps, giving 405 candidate events, of which 76 are removed because of data gaps around the trigger instant, leaving 329 usable events involving 328 flights. The negative control is 1907 records with no windshear alert at any time, involving 1853 flights; another 144 records (139 flights) whose alerts appear only outside the descent phase satisfy neither side's definition and are excluded from both sets.

Real flight has no windshear annotation in the certification sense, and the GPWS alert is an operational reference rather than ground truth: alerts may be triggered by non-shear components or by the handling process, and unalerted encounters may also genuinely exist. The coverage rates below should therefore be read as a measure of agreement between two independent systems, not as detection rates themselves.

5.2. Consistency Check on the Simulation Side

The cost of the mapping is first checked inside the simulation domain: on the same held-out set of 142 files and 522,854 frames, the metrics of the full-state model and the two mapped variants are listed in Table 7. The observable-subspace mapping is nearly lossless: the overall accuracy is 97.6% for the full-state model, 97.5% for the native-sampling mapped variant, and 97.8% for the sampling-rate-matched version. All three models have zero RA-frame fraction on the false-alarm and turbulence families—the false-alarm-suppression capability is fully preserved in the eight-dimensional subspace. The recall drop concentrates in the dual-core hazard family, from 98.9% to 92.6%; resolving this family's secondary inner core depends on the moment channels of the full state, which is the main cost of the observability contraction; the recall of the certification family tsosupp instead rises to 98.9%.

Table 7. Three-model comparison on the same held-out set of 522,854 frames; per-family columns are frame-level RA recall.

Model	Input	Overall acc.	RA Recall	Dual-Core	ma_dual	nasa_supp	tsosupp	FA/Turbulence RA Frames
Full-state model (Section 4.4)	42-dim full state	97.6%	96.3%	98.9%	98.6%	97.0%	97.1%	0.0%
Mapped model (native sampling)	8 features × 5 statistics	97.5%	96.4%	93.7%	96.8%	96.4%	97.7%	0.0%
Mapped model (rate-matched)	Same as above	97.8%	96.3%	92.6%	97.9%	95.8%	98.9%	0.0%

5.3. Results on Real Data

On the 329 GPWS-positive events, 68.4% (225 events) show RA-level output within 10 s around the trigger instant; the control variant without sampling-rate matching reaches 54.1%, so rate matching alone contributes +14.3 percentage points and lengthens the median time gap between the first sustained alert and the first GPWS trigger from 4.0 s to 14.5 s. Of the 225 covered events, 221 respond earlier than the GPWS trigger; the median time gap is computed on these 221 events, with an interquartile range of 3.8–59.9 s (Figure 9). Two properties of this statistic are worth stating. First, the record window extends 90 s before the trigger, so a response beginning earlier than the record start would be right-censored; in the data no event reaches that boundary (maximum lead 89.0 s), and the median is therefore unaffected by censoring. Second, the coverage rate is insensitive to the choice of coincidence window: it is 66.3% at ±3 s, 68.1% at ±5 s, 68.4% at both ±10 s and ±15 s, and 69.6% at ±30 s. This time gap does not constitute an early-alerting capability in the operational sense—under the deployment of Section 3.2, the detector participates in arbitration only after the certified chain alerts and does not alert independently—what it shows is that the physical signal of energy loss appears in the onboard-observable quantities earlier than the crossing of the certification threshold—the physical consequence on which arbitration relies is already stably readable before the alert instant. This gap is consistent in magnitude with the about 12 s of the A1 wind field in the in-loop reproduction of Section 4.5. In Figure 9a, the average RA probability of covered events crosses 0.5 about 4 s before the trigger, peaking at 0.83 at 1.4 s before the trigger; uncovered events stay near zero throughout, peaking at no more than 0.02—the two groups are nearly separable at the probability level, meaning that non-coverage is not a weak response but the absence of discriminating structure in the observable subspace for these events. Grouping by whether a distinguishable airspeed-acceleration signal exists in the 5 s before the trigger, with the signal criterion $|\dot{v}|$ peak above 0.5 m/s^2, the two groups have almost identical coverage: 68.4% for the 320 events with signal, and 6 of the 9 events without signal are covered. This shows that coverage does not depend on a single feature at the instant before the trigger, but on the accumulated shape of the window statistics.

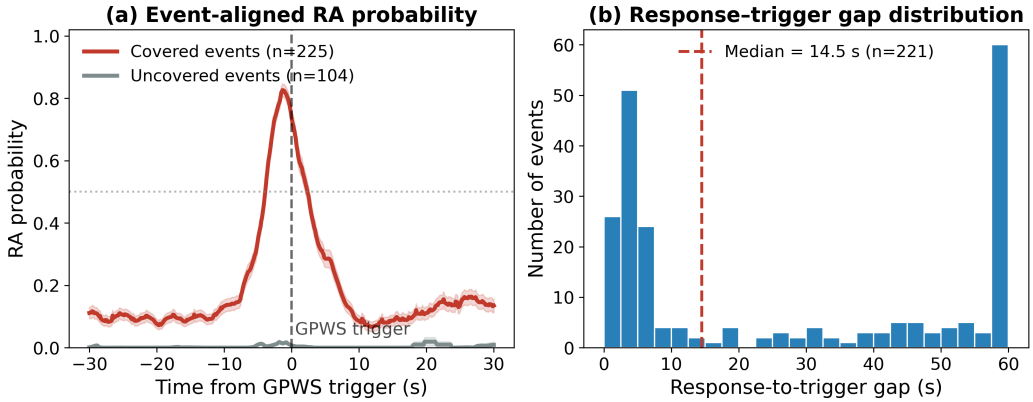


Figure 9. Behavior of the mapped model on real QAR-positive events. (**a**) Average RA probability aligned to the GPWS trigger instant, with the shaded band showing the standard error of the mean: covered events cross 0.5 about 4 s before the trigger, peaking at 0.83 at 1.4 s before the trigger; uncovered events stay below 0.02 throughout. (**b**) Distribution of the response-to-trigger gap for the 221 early-responding events, median 14.5 s; 60 s on the horizontal axis is the plotting upper limit, and long-tail events are not expanded in the figure.

Under the arbitration semantics of Section 3.2, the 104 uncovered events correspond to alerts that would be downgraded, and their severity needs separate examination. The

median radio altitude at trigger of this group is 658 m, higher than the 510 m of the covered group; the peak airspeed acceleration in the 5 s before the trigger is almost the same as the covered group, a median of 1.28 versus 1.36 m/s^2, so the apparent intensity at the trigger instant is not weaker; taking the airspeed decrease in the 5 s after the trigger as the consequence proxy, the median is 4.11 m/s, below the covered group's 4.94 m/s, but 32.7% still exceed 5 m/s. The distribution around this median is broad: the interquartile range is 2.13–5.72 m/s, the 90th percentile reaches 7.05 m/s, and the maximum is 9.52 m/s; counted against the label structure of Section 3.3, 31 of the 104 events stay below 2.5 m/s, 39 fall in the TA-equivalent band of 2.5–5 m/s, and 34 exceed 5 m/s. The 34 severe events are mostly high-altitude triggers (median radio altitude: 608 m; only four below 300 m), where the recovery margin is larger; this stratification mitigates the risk but does not remove it. The downgraded events therefore mostly have mild consequences, but they are not all harmless. QAR's second-level sampling systematically attenuates peaks, so the proxy values above underestimate the true losses; whether all downgraded events correspond to mild consequences requires higher-rate data or ground-truth verification, and this paper makes no such claim.

The negative control comprises 1853 flights after deduplication, of which 194 are removed for incomplete feature caches, leaving 1659 in the statistics. The median RA-frame fraction is 0; the rate-matched variant has a mean of 3.5%, and 43.6% of flights have RA-level output lasting more than 1 s; the control variant without rate matching gives 1.1% and 23.0%, respectively. The difference between the two variants has the same origin as the +14.3 percentage points on the positive side: rate matching shifts the sensitivity upward as a whole, and the same setting that raises coverage from 54.1% to 68.4% also raises the sustained RA activity on alert-free flights from 23.0% to 43.6%—two faces of the same sensitivity shift. Among flights with sustained output, about 60% are accompanied by significant airspeed-disturbance signals in the corresponding periods (53–59% under different statistical-window conventions), and the median airspeed-disturbance intensity during sustained-RA periods is 5.99 m/s^2, versus 1.42 m/s^2 in matched quiet periods; airspeed disturbance is homologous with the detector input, so this coincidence is not independent evidence that the residual activity is harmless, only that the residual output is not random noise. These flights never triggered GPWS, and the residual activity more likely corresponds to bumps or weak shear below the alert threshold—disturbances that the Dryden turbulence underlying the benign simulation families does not generate: as a stationary Gaussian process with a fixed spectrum it produces continuous, statistically uniform small-amplitude fluctuations, whereas real approach-phase disturbances are intermittent, localized events (gust patches, weak shear layers, thermals); in the absence of ground truth, this paper makes no further false-alarm attribution for it.

Finally, two typical events. In the covered case, a flight triggers GPWS at a radio altitude of 451 m, with a peak airspeed acceleration of 0.87 m/s^2 in the 5 s before the trigger; the model first gives sustained RA-level output 12.2 s before the trigger and maintains it until the event ends, a time gap close to the median of the whole set; the vertical motion and airspeed change of this event are clearly distinguishable in the mapped subspace—a typical transferable event. In the uncovered case, another flight triggers GPWS at a radio altitude of about 864 m, with a peak airspeed acceleration of only 0.46 m/s^2 in the 5 s before the trigger, and the model's RA-probability peak over the entire 135 s window is only 0.02—no shear structure indeed appears in the observable subspace. Such high-altitude trigger events represent the main shape of the uncovered part: the signal is either weaker than the sampling resolution of QAR, or the channel it acts on is not within the eight-dimensional subspace.

5.4. Applicability Boundaries and Future Work

Before discussing the applicability boundaries, it must be made clear that the false-alarm suppression of this paper is not free of cost; the costs concentrate in four places. First, frame-level misses concentrate on marginal frames whose intensity is about half that of detected frames—a blind spot inherent to the 0.5 s short window; the held-out frame-level miss rate is 1.1–8.7%, with the single-core microburst family at the upper end. Second, residual triggers and misses remain at large sample size: the false-alarm-zone trigger rate is 8.0% in the 14,825-case extended validation rather than the zero suggested by the small held-out sample, and on the hazard side the detector alone misses 12 boundary cases (all caught by the F chain; Section 4.1); extreme values of small samples should be read as statistical fluctuation. Third, arbitration downgrade carries tail risk: among the 104 real-flight events whose alerts would be downgraded, about one third still exceed 5 m/s in the post-trigger 5 s airspeed-decrease proxy, and this proxy is systematically underestimated by second-level sampling; this paper makes no claim about the harmlessness of downgraded events. Fourth, sensitivity and observability each have two faces: the same sampling setting that raises QAR-side coverage from 54.1% to 68.4% simultaneously raises sustained RA activity on alert-free flights from 23.0% to 43.6%, and mapping to the onboard-observable subspace reduces the dual-core family recall from 98.9% to 92.6%. These costs do not negate the value of false-alarm suppression; what they delimit is exactly the boundary within which this value holds.

First, the boundaries specific to this section. Three explanations for the uncovered events cannot be ruled out: the energy consequences of the real events are weaker than the 5 m/s/5 s RA criterion; residual attenuation of the signal by QAR sampling; and differences between the QAR fleet's aircraft types and the simulation platform; without ground truth, the three cannot be separated further. The results of this section are therefore evidence of cross-domain consistency, not a certification conclusion on operational performance.

More generally, the results of this paper hold within the simulation environment defined in Section 2, which this section has extended to real operational data; these applicability boundaries also point to the directions of future work. The detector is trained and evaluated on the Boeing 747-400 simulation platform of Section 2.1; the QAR validation of this section has crossed to real data, but the fleet types differ from the platform and there is no windshear ground truth in the certification sense; rechecking on QAR of a fleet of the same aircraft type as the platform, and building a ground-truth reference with flight-test or high-confidence meteorological-reconstruction data, is the most direct next step. The simulated wind fields all come from the Oseguera–Bowles/Bray analytic model family; other shear phenomena, such as fronts, sea-land breeze, and wake vortices, are not covered, and the same time-window analysis can be generalized to these wind fields to test whether the divergence band still exists. The baseline chain is a representative offline reproduction of the TSO-C117 alerting logic; certified airborne equipment may differ in filter and window details, and Section 3.1 shows that the divergence is not specific to the parameter point, but an interface-level comparative validation against a concrete airborne implementation is still necessary. For the label definition, the hazard criterion takes the forward 5 s airspeed loss with the 5 m/s RA threshold; Section 4.2 shows the main conclusions are robust to ±50% changes in the threshold, and finer consequence criteria linked to crew recovery capability are worth exploring. On data scale, the 710 simulation files do not support generalization claims beyond the sampled envelope; enlarging the sampling envelope and the aircraft-type coverage is a direct extension direction. Training and all validation were completed at the 260 t nominal mass, and paired experiments show that detection performance is sensitive to mass offsets beyond about ±5 t, and Section 4.7 shows that mass-jittered retraining over 230–290 t restores hazard detection at ±20 t (70/70), with discrimination at +20 t still to

be improved for fleet deployment. All scenarios are approach-phase encounters; shear encounters in other flight phases, such as takeoff, cruise, and holding, are not covered, and their time-window characteristics need to be tested separately on platform models of the corresponding phases. The two boundaries of the divergence band likewise depend on the platform: the 5 s lower bound is set by the window-dominated dilution mechanism, and the upper bound of about 10.3 s is calibrated from the in-band loss distribution of this platform and may shift with aircraft type and approach speed; multi-scale fusion weighted by the estimated event time scale is a natural improvement path.

In addition to the items above, several boundaries of evidence granularity must be stated. The adjudication criterion for false alarms is homologous with the detector's training labels—both are a forward 5 s airspeed loss not exceeding 5 m/s—so the false-alarm conclusions measured under it are not fully independent of the training target; this revision closes that gap with the four third-party consequence criteria of Section 4.6, under which the false-alarm-zone scenarios remain unanimously benign on every sustained and proximity measure. A richer third-party check—trajectory-keeping ability under a standard escape profile [30,31]—remains open. Second, the held-out results on the fa_boost family in Section 4.4 are an in-distribution evaluation—the test and training files are generated from the same parameter box—and a recheck under an out-of-distribution split remains to be done; the input ablation removing the two F-derived channels FFactor and WindDetRate is now reported in Section 4.4 (Table 5), under which false-alarm-zone silence is preserved, so the statement "decoupled from the F-factor criterion" now covers the labels, the decision criterion, and the input channels. Finally, the frame-by-frame continuity recheck of arbitration levels during the persistence of a certification alert is reported in Section 4.5, and multi-seed repetition of the main model is reported in Section 4.4: across four seeds the frame-level metrics fluctuate by at most about half a percentage point, tightening the earlier single-seed estimate that referenced the 0.7 percentage points of the reduced-data ablation. The joint threshold–window sweep and the convergence test under finer wind-field grids remain future work.

6. Conclusions

This paper has proposed and validated a time-window divergence analysis of reactive windshear alerting: the root cause of the F-factor's systematic false alarms in compact microbursts is the mismatch between the response time scale of the alerting chain and the physical duration of the event—the moving average stretches a brief threshold crossing into a persistent one, and persistence confirmation therefore fails. The resulting predicted band $T \in (5, \sim 10.3)$ s ($R \in (590, 1214)$ ft) overlaps the TSO-C117 certification envelope, and the most compact certification wind field A1 ($R = 918$ ft) falls inside it, showing that the divergence band is not a theoretical extrapolation but a real flight region explicitly covered by the certification standard.

Two independent lines of evidence support this analysis. Monte Carlo simulation gives in-band uniform/targeted false-alarm rates of 60.5%/99.2%; Bayesian optimization guided by criterion–physics divergence converges to the same region with a 70% discovery rate (centroid $R \approx 704$ ft); the detector triggers RA on none of the 35 false-alarm points found by the search and triggers on all nine genuine-hazard points encountered in the same search, showing that the divergence band is a structural property of the encounter space and that false-alarm suppression and hazard discrimination hold outside the training data. The three-channel short-window detector built on the physical-consequence criterion moves the decision basis from signal amplitude to the physical consequence of flight: measured by the 5 m/s/5 s consequence criterion of this paper, the detector's RA trigger rate in the false-alarm zone on the held-out set is 0 (13 files), versus a 92.3% WARNING trigger rate of

the F chain; frame-level recalls on the five hazard families are 91.3–98.9%, file-level recall of hazardous files is 100%, and all ten TSO-C117 certification wind fields alert in the loop. The held-out conclusions are reproduced in scenario-by-scenario paired extended validation on 14,825 valid scenarios accumulated in two batches under the same distribution and protocol (about 58.6 million frames): at the classifier level, all 6888 scenarios containing genuine hazard are detected by the union of the two criteria, the false-alarm-zone alert rate drops from 99.2% to 8.0%, and the residual triggers concentrate in the transition from the upper edge of the divergence band to genuine hazard. Re-adjudication under four third-party consequence criteria that share no definition with the training labels confirms that the false-alarm-zone scenarios are unanimously benign on every sustained and proximity measure, and seed repetition shows the main frame-level metrics fluctuating by at most about half a percentage point; the boundary cases are equally explicit—the detector is configured per aircraft mass, and its detection rate collapses outside the ± 5 t neighborhood of the training mass, a data-coverage gap that mass-jittered retraining closes at ± 20 t at the cost of discrimination at $+20$ t (Section 4.7). After mapping the detector onto the eight-dimensional onboard-recordable observable subspace with sampling-rate matching, the method also holds on 2411 real QAR records (2346 flights): trained without any real data, it covers 68.4% of 329 GPWS windshear alert events, of which 221 respond earlier than the GPWS trigger with a median gap of 14.5 s; the median RA-frame fraction on alert-free flights is 0 with a mean of 3.5%, and 43.6% of alert-free flights have RA-level output lasting more than 1 s. At the same time, the cost side must be stated: under the arbitration deployment semantics of Section 3.2, the 104 uncovered events (31.6%) correspond to alerts that would be downgraded, of which about one third exceed 5 m/s in the post-trigger 5 s airspeed decrease; the operational acceptability of this risk transfer depends on ground-truth calibration and operator policy and must be weighed together with the false-alarm-suppression benefit. Within the policy of Section 3.2 a downgraded alert is not silenced but retained at advisory level, so even the worst-case outcome of arbitration leaves the crew with a caution-level cue; the choice between this advisory-preserving policy and stricter configurations is an operator-level decision that requires labeled operational data.

These results do not mean that the certified criterion fails: the F-factor chain remains effective within its design envelope, and what this paper targets is a bounded residual space near the alert threshold. No single window or criterion can be optimal at all time scales; the predictive/reactive two-layer architecture of operational systems has been built on exactly this complementarity since the late 1980s, and a learned consequence criterion extends this complementarity into the reactive layer. The applicability boundaries of this paper and the corresponding improvement directions have been discussed in Section 5.4.

Author Contributions: Conceptualization, Q.L. and D.W.; methodology, Q.L.; software, Q.L.; validation, Q.L.; formal analysis, Q.L.; investigation, Q.L.; resources, Q.L. and D.W.; data curation, Q.L.; writing—original draft preparation, Q.L.; writing—review and editing, Q.L. and D.W.; visualization, Q.L.; supervision, D.W.; project administration, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The simulation data presented in this study are available on request from the corresponding author. The operational QAR data are not publicly available due to confidentiality agreements with the data provider.

Acknowledgments: This work was supported by COMAC Shanghai Aircraft Design and Research Institute and Shanghai Aircraft Flight Test Engineering Co., Ltd. During the preparation of this manuscript, the authors used Kimi (Moonshot AI) for English translation and language polishing of part of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. The 42 State Channels of the Detector Input

The detector input is taken from the 48 state channels recorded per frame by the platform, with 6 items removed (to give 42): the frame number (`Frame`), time (`TimeSec`), the ground flag (`OnGround`), and the three alert-label channels (`ShearAlert`, `EnergyAlert`, `PinnsAlert`). Names follow the simulator's bus variable names and are typeset in typewriter font as raw identifiers, distinct from mathematical symbols. By onboard measurability, the 42 channels are listed in two categories; this distinction is exactly the basis of the observable-subspace mapping of Section 3.5.

A total of 27 are standard bus parameters measurable or recordable onboard: attitude (8): `Q`, `QDot`, `P`, `R`, `Pitch`, `Roll`, `Alpha`, `Beta`; velocity (11): `U`, `W`, `V`, `Uaero`, `Waero`, `Vn`, `Ve`, `Vd`, `Vc`, `VcDot`, `Mach`; atmosphere (1): `Rho`; aircraft state (4): `Pz`, `GForce`, `AGL`, `Mass`; control surfaces (3): `De`, `Elevator`, `ThrottleRatio`.

A total of 15 are simulation-internal quantities not directly measurable onboard: aerodynamic forces and moments (10): `UDot`, `WDot`, `VDot`, `Lift`, `Drag`, `Thrust`, `SideForce`, `XForce`, `ZForce`, `Pmt`; three-dimensional wind field (3): `WindN`, `WindE`, `WindD`; alerting-computer internals (2): `WindDetRate`, `FFactor`.

`UDot`, `WDot`, `VDot` can in principle be approximately reconstructed from the specific force of inertial measurements, but in this platform's records they are taken from the resultant-force output of the dynamics model and are therefore classified as simulation-internal quantities; `FFactor` and `WindDetRate` are intermediate quantities of the criterion chain reproduced in Section 2.3 and are F-derived channels, so the statement "decoupled from the F-factor criterion" in the main text refers only to the labels and the decision criterion and does not cover these two input channels (see Section 3.3).

References

1. International Civil Aviation Organization. *Manual on Low-Level Wind Shear*, 1st ed.; Doc 9817 AN/449; ICAO: Montréal, QC, Canada, 2005.
2. Fujita, T.T. *The Downburst: Microburst and Macroburst*; University of Chicago Press: Chicago, IL, USA, 1985.
3. Wilson, J.W.; Wakimoto, R.M. The discovery of the downburst: T.T. Fujita's contribution. *Bull. Am. Meteorol. Soc.* **2001**, *82*, 49–62. [CrossRef]
4. Bowles, R.L. Reducing windshear risk through airborne systems technology. In *Proceedings of the 17th Congress of the International Council of the Aeronautical Sciences (ICAS), Stockholm, Sweden, 9–14 September 1990*; Paper ICAS-90-1.9.3; American Institute of Aeronautics and Astronautics: Washington, DC, USA, 1990.
5. Fujita, T.T.; Caracena, F. An analysis of three weather-related aircraft accidents. *Bull. Am. Meteorol. Soc.* **1977**, *58*, 1164–1181. [CrossRef]
6. National Transportation Safety Board. *Aircraft Accident Report: Delta Air Lines, Inc., Lockheed L-1011-385-1, N726DA, Dallas/Fort Worth International Airport, Texas, August 2, 1985*; NTSB/AAR-86/05; NTSB: Washington, DC, USA, 1986.
7. National Transportation Safety Board. *Aircraft Accident Report: Pan American World Airways, Inc., Boeing 727-235, N4737, Kenner, Louisiana, July 9, 1982*; NTSB/AAR-83/02; NTSB: Washington, DC, USA, 1983.
8. Federal Aviation Administration. *Technical Standard Order TSO-C117b: Airborne Windshear Warning and Escape Guidance Systems for Transport Airplanes*; FAA: Washington, DC, USA, 2019.
9. Wang, Q.; Sun, H.; Zhang, K.; Wang, Q. Horizontal scale study of low-level wind shear based on F-factor. In Proceedings of the International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), Wuhan, China, 26–28 July 2024; SPIE: Bellingham, WA, USA; Volume 13395, p. 127. [CrossRef]

10. Khattak, A.; Zhang, J.; Chan, P.-W.; Chen, F.; Almaliki, A.H. Aviation safety at the brink: Unveiling the hidden dangers of wind-shear-related aircraft-missed approaches. *Aerospace* **2025**, *12*, 126. [CrossRef]
11. Khattak, A.; Zhang, J.; Chan, P.-W.; Chen, F.; Almaliki, A.H. A new frontier in wind shear intensity forecasting: Stacked temporal convolutional networks and tree-based models framework. *Atmosphere* **2024**, *15*, 1369. [CrossRef]
12. Khattak, A.; Chan, P.-W.; Chen, F.; Almaliki, A.H. A hybrid VMD-TPE-TCN framework for wind shear prediction near an airport runway. *Atmosphere* **2025**, *16*, 381. [CrossRef]
13. Hou, F.; Sun, G. A hybrid deep learning architecture for enhanced vertical wind and FBAR estimation in airborne radar systems. *Aerospace* **2025**, *12*, 679. [CrossRef]
14. SAE International G-34 Committee on Artificial Intelligence in Aviation; EUROCAE WG-114. *Development and Assurance Considerations for Aeronautical Systems and Equipment Implemented with Machine Learning*; ARP6983/ED-324, draft, target publication date 31 December 2026; SAE International: Warrendale, PA, USA, 2026.
15. European Union Aviation Safety Agency. *EASA Artificial Intelligence Roadmap 2.0: A Human-Centric Approach to AI in Aviation*; EASA: Cologne, Germany, 2023.
16. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [CrossRef]
17. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [CrossRef]
18. Li, H.; Gou, L.; Li, H.; Liu, Z. Physics-guided neural network model for aeroengine control system sensor fault diagnosis under dynamic conditions. *Aerospace* **2023**, *10*, 644. [CrossRef]
19. Zhu, D.; Peng, J.; Ding, C. A neural network with physical mechanism for predicting airport aviation noise. *Aerospace* **2024**, *11*, 747. [CrossRef]
20. Oseguera, R.M.; Bowles, R.L. *A Simple, Analytic 3-Dimensional Downburst Model Based on Boundary Layer Stagnation Flow*; NASA TM-100632; NASA Langley Research Center: Hampton, VA, USA, 1988.
21. Bray, R.S. *A Method for Three-Dimensional Modeling of Wind-Shear Environments for Flight Simulator Applications*; NASA TM-85969; NASA Ames Research Center: Moffett Field, CA, USA, 1984.
22. Proctor, F.H. Numerical simulations of an isolated microburst. Part I: Dynamics and structure. *J. Atmos. Sci.* **1988**, *45*, 3137–3160. [CrossRef]
23. U.S. Department of Defense. *Military Specification: Flying Qualities of Piloted Airplanes*; MIL-F-8785C; DoD: Washington, DC, USA, 1980.
24. RTCA. *Minimum Operational Performance Standards for Traffic Alert and Collision Avoidance System II (TCAS II)*; RTCA DO-185B; RTCA: Washington, DC, USA, 2008.
25. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; de Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **2016**, *104*, 148–175. [CrossRef]
26. Moss, R.J.; Kochenderfer, M.J.; Gariel, M.; Dubois, A. Bayesian safety validation for failure probability estimation of black-box systems. *J. Aerosp. Inf. Syst.* **2024**, *21*, 533–546. [CrossRef]
27. Federal Aviation Administration. *Pilot Windshear Guide*; Advisory Circular AC 00-54; FAA: Washington, DC, USA, 1988.
28. Wang, F.; Zhang, Y.; Wang, Y.; Liu, T.; Wei, Z.; He, H. QAR data-driven flight anomaly detection and risk warning: A review of statistical learning and machine learning methods in aviation safety. *Aerospace* **2026**, *13*, 693. [CrossRef]
29. Jasra, S.K.; Valentino, G.; Muscat, A.; Camilleri, R. A comparative study of unsupervised machine learning methods for anomaly detection in flight data: Case studies from real-world flight operations. *Aerospace* **2025**, *12*, 151. [CrossRef]
30. Psiaki, M.L.; Stengel, R.F. Optimal flight paths through microburst wind profiles. *J. Aircr.* **1986**, *23*, 629–635. [CrossRef]
31. Miele, A.; Wang, T.; Melvin, W.W. Optimal take-off trajectories in the presence of windshear. *J. Optim. Theory Appl.* **1986**, *49*, 1–45. [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.