Author manuscript. Pik, E. (2026). Cross-Channel Anomaly Coupling in ADS-B: A Surveillance-Integrity Monitor for Air Traffic Management. Preprint: https://doi.org/10.2139/ssrn.6868542. Replication data and code: https://doi.org/10.5281/zenodo.20387308

Cross-Channel Anomaly Coupling in ADS-B: A Surveillance-Integrity Monitor for Air Traffic Management

Eugene Pik[a,*]

[a] *Mevocopter Aerospace, 70 Samuel Oster Ave, Vaughan, L4J 7C8, ON, Canada*

Abstract

Automatic Dependent Surveillance–Broadcast (ADS-B) telemetry underpins cooperative aircraft surveillance across the global air traffic management system. Each broadcast channel's integrity is monitored individually, yet the joint consistency of the kinematic channels is neither monitored operationally nor characterized at scale. We characterize this consistency at planetary scale and convert it into a deployable integrity monitor for air navigation service providers. We analyze one year of OpenSky Network state-vector data (calendar 2025; 117.5 million aircraft-hour aggregates over 240,001 airframes). Per-aircraft-hour indicators of severe longitudinal-acceleration (DV), vertical-acceleration (DVR), and track-angle-rate (DH) anomalies exhibit a 3.14× pairwise enrichment of DV given DVR, and they violate conditional independence most strongly in the DV–DVR relation conditioned on DH ($V_B = 1.54$), with the DV–DH pathway largely DVR-mediated. The coupling lies outside an exact within-airframe permutation null, survives eight pre-specified controls, and strengthens to 3.81× in a high-integrity stratum, inconsistent with receiver-side noise, peaking in cruise ($R_{\text{cruise}} = 4.41$), where manoeuvre is implausible. From this coupling we derive a multi-channel consistency monitor, calibrated to a controlled per-window false-positive rate, that reads the coupling structure rather than per-channel rates and so separates a correlated burden surge (coupling intact) from genuine channel decoupling, a failure mode invisible to single-channel monitoring. A sustained 50-day baseline-elevation episode (2 April–22 May 2025), absent in the calendar-aligned 2024

*Corresponding author
Email address: eugene.pik@mevocopter.com (Eugene Pik)

window, exercises the monitor as a positive control. The work establishes multi-channel kinematic-anomaly co-occurrence as structural, not coincidental, and delivers a joint-consistency diagnostic for cooperative-surveillance infrastructure.

Keywords: Automatic Dependent Surveillance-Broadcast (ADS-B), Air transport systems, Cooperative surveillance integrity monitoring, Common-cause failure coupling, Conditional independence testing, Statistical process control, Decision aid

---

Nomenclature

Acronyms

ACAS X	Airborne Collision Avoidance System X
ADS-B	Automatic Dependent Surveillance-Broadcast
ANSP	Air Navigation Service Provider
CI	Confidence Interval
CPR	Compact Position Reporting
DH	Heading-rate anomaly indicator
DV	Longitudinal-acceleration (groundspeed-divergence) anomaly indicator
DVR	Vertical-acceleration (vertical-rate) anomaly indicator
DO-260B	RTCA Minimum Operational Performance Standards for 1090 MHz ES ADS-B
FAR	Federal Aviation Regulations
GNSS	Global Navigation Satellite System
ICAO	International Civil Aviation Organization
NAC	Navigation Accuracy Category
NACv	Navigation Accuracy Category for velocity
NACp	Navigation Accuracy Category for position
NIC	Navigation Integrity Category
NOAA	National Oceanic and Atmospheric Administration
OR	Odds Ratio
RF	Radio Frequency
SIL	Source Integrity Level
SMS	Safety Management System
TBO	Trajectory-Based Operations

UTC	Coordinated Universal Time
VIF	Variance Inflation Factor

Symbols

$f_{\Delta t=1}$	Fraction of per-step pairs in an aircraft-hour with $\Delta t = 1\,\mathrm{s}$ exactly
h_i	Track angle (degrees) at message index i
N	Sample size (aircraft-hours or windows, depending on context)
$P(\cdot)$	Marginal probability
$P(\cdot \mid \cdot)$	Conditional probability
$R(\cdot \mid \cdot)$	Conditional-probability enrichment ratio (Eq. 4)
Δt	Time between consecutive state-vector messages from the same airframe
Δ_{wrap}	Shortest signed angular difference modulo 360°
V_A, V_I, V_B	Three pathway conditional-independence violation ratios (Eqs. 5–7)
v_i	Transponder-reported ground velocity (m/s) at message index i
w_i	Transponder-reported vertical rate (m/s) at message index i; predominantly barometric per DO-260B defaults but not source-bit filtered (see Section 7.5)

1. Introduction

1.1. Background and Motivation

The transition toward 4D Trajectory-Based Operations (TBO) under global modernization frameworks such as NextGen in the United States and SESAR in Europe depends on the continuous, high-fidelity transmission of cooperative surveillance data [1, 2, 3]. Automatic Dependent Surveillance-Broadcast (ADS-B) serves as the primary data link for these predictive systems, with each transponder-equipped aircraft periodically broadcasting state vectors containing position, velocity, altitude, and heading over the 1090 MHz Mode S Extended Squitter spectrum [4, 5]. While ADS-B has substantially enhanced airspace capacity and situational awareness, its baseline architecture lacks native message authentication and encryption, leaving the broadcast network vulnerable to both unintentional degradation (radio-frequency saturation, multipath interference, packet collisions [6], antenna shadowing during dynamic manoeuvring) and intentional disruption [7, 8, 9]. As automated safety systems such as the Airborne Collision Avoidance System X (ACAS X) increasingly rely on probabilistic state-transition models that assume continuous kinematic coherence of broadcast trajectories [10, 11], the internal consistency of the broadcast kinematic channels has become operationally consequential.

Prior research has separately characterized single-dimensional degradation in ADS-B kinematic streams. Anomaly-detection studies have proposed deep neural networks [12], Gaussian mixture models [13], and Kalman-filter residuals [14] to flag corruption in position, velocity, or heading channels individually. The empirical literature has additionally documented co-occurrence of GNSS-driven position anomalies with NOTAM-reported jamming events [15, 16] and has linked aggregated kinematic-anomaly geographies to geopolitical disturbance [17]. To our knowledge, however, no prior study has quantified the joint conditional dependence structure of severe per-step anomalies in groundspeed, vertical rate, and heading at planetary scale, using consistent rate-normalized definitions and formal conditional-independence tests with statistical inference.

This gap matters because the operational implications of single-dimensional versus multi-dimensional degradation are not equivalent. An ADS-B trajectory exhibiting only a momentary heading jump may safely be discarded by a conflict-prediction algorithm and re-acquired on the next message. A trajectory in which severe heading-rate, vertical-rate, and groundspeed jumps co-occur in the same aircraft-hour is statistically inconsistent with independent single-channel noise and warrants a structurally different downstream response: aggregate fleet-health monitoring rather than tactical filtering. Air navigation service providers monitoring cooperative-surveillance infrastructure under ICAO Annex 19 Safety Management System requirements require evidence that the conditional dependence between such failures is a real, characterizable structural feature of the data product, not a coincidence of marginal failure rates.

Integrity monitoring is, by contrast, a mature discipline on the navigation side of the avionics stack. Receiver Autonomous Integrity Monitoring (RAIM), and its advanced multi-constellation successor ARAIM, provide assurance that a GNSS position solution is trustworthy by exploiting redundancy across satellite measurements to detect and, where possible, exclude inconsistent observations [18]. Cooperative surveillance has no equivalent cross-channel integrity check: the ADS-B quality indicators (Navigation Integrity Category (NIC), Navigation Accuracy Category for position and velocity (NACp, NACv), and Source Integrity Level (SIL)) certify each broadcast parameter in isolation [19], and ground surveillance-data-processing systems validate the kinematic channels independently. The diagnostic developed in this paper is, in effect, the surveillance-side analogue of a redundancy-based integrity monitor: it asks whether the kinematic channels remain *mutually consistent*, rather than whether any single channel is individually plausible.

The characterization of multi-channel failure coupling has a long methodological tradition in reliability engineering under the rubric of *common-cause failure* (CCF) analysis. Parametric families including the beta-factor and alpha-factor models, together with non-parametric copula-based approaches, each parameterize the excess joint-failure probability beyond what independent marginal failure rates would predict [20, 21, 22, 23]. To our knowledge, however, these CCF frameworks have not been applied to the multi-dimensional kinematic channels of cooperative-surveillance telemetry at planetary scale; the pairwise enrichment ratio R of Equation 4 is formally analogous to the ratio $\beta/(1-\beta)$ in the beta-factor model, reinterpreted as a conditional-probability amplification rather than a failure-fraction parameter.

1.2. Gap Identification

The existing ADS-B integrity literature suffers from four limitations that motivate this study.

Single-channel framing. Anomaly detectors and threshold-based filters report speed, altitude, and heading discrepancies separately, forcing operators to mentally integrate disparate flags without quantitative guidance on whether they co-occur more often than chance would predict.

Interval confounding. Per-message anomaly definitions are commonly posed as absolute differences ($|\Delta v|$, $|\Delta w|$, $|\Delta h|$) between consecutive observations. At the nominal 1 Hz output rate of the OpenSky historical archive, real-world inter-message intervals nevertheless range from one second to over a minute depending on receiver coverage [5]. An identical absolute jump that is physically extreme over $\Delta t = 1\,\mathrm{s}$ is routine over $\Delta t = 30\,\mathrm{s}$, conflating data-link gaps with broadcast-channel anomalies. Rate-normalized definitions (acceleration units; degrees per second) sidestep this confound but have been used inconsistently.

Absence of formal conditional-independence inference. Where prior work has noted co-occurrence of multiple-dimensional ADS-B kinematic anomalies, it has done so descriptively, reporting tabulated joint counts or visual co-occurrence matrices, without formal hypothesis tests of conditional independence and without controls against alternative explanations such as data-link gap conflation or transponder-side data quality.

Receiver- versus broadcaster-side ambiguity. The dominant alternative explanation for any observed correlation between kinematic-anomaly channels is that the correlation lives in the receiver path, noisy reception of an otherwise coherent broadcast. Distinguishing this from a genuinely coupled broadcasting platform requires data-quality stratification that exploits the transponder's own self-reported Navigation Accuracy Category, and this analysis has not previously been performed at planetary scale.

Disconnection from the reliability-engineering CCF framework. The common-cause-failure models routinely used in nuclear, process-industry, and railway safety (Section 1.1) have not been connected to ADS-B kinematic-channel co-occurrence, leaving air-transport safety management without the inferential machinery that other safety-critical domains apply.

1.3. Contributions of This Study

This study makes five contributions to the analysis of ADS-B kinematic-anomaly consistency in the calendar-year 2025 OpenSky archive: four pieces of empirical evidence on the conditional dependence structure, and one deployable operational tool.

First, we report the joint distribution of three pre-specified rate-normalized severe-anomaly indicators (DV, DVR, DH) across 117.5 million aircraft-hour aggregates clustered by 240,001 unique airframes. We test conditional independence in all three identifiable pathways and quantify the deviation from each independence prediction with cluster-bootstrap confidence intervals.

Second, we formalize the deviation by comparing the observed conditional-independence violation against a permutation null distribution constructed by exact within-airframe random reassignment of the vertical-anomaly indicator. The observed Pathway B violation (DV $\perp$ DVR | DH) lies entirely outside the 1,000-iteration within-airframe permutation null (one-sided $p < 10^{-3}$; a standardized gap exceeding 1,200 null standard deviations from the null mean), eliminating random pairwise structure as a plausible explanation.

Third, we run eight pre-specified robustness analyses targeting the most plausible alternative explanations and confounds: three-tier data-quality decomposition, DVR-threshold sweep, seasonal robustness across four checkpoint weeks, geographic robustness across six continental airspaces, strict fixed-wing regression specification, rotary-wing counterfactual saturation grid, phase-classification symmetric-threshold sensitivity, and, decisively for the receiver-side-noise alternative, Navigation Accuracy Category for velocity (NACv), Navigation Integrity Category (NIC), and receiver-redundancy stratification of the aircraft-hour aggregates.

Fourth, we document a sustained 50-day baseline-elevation episode running from 2 April through 22 May 2025, present a per-day localization analysis across geography and airframe type, and cross-reference the episode against NOAA Storm Events records of severe convective activity in the United States to constrain candidate operational drivers.

Fifth, building on the irreducible three-way coupling, we derive a deployable multi-channel consistency monitor: a lightweight, ground-side statistical-process-control diagnostic that an air navigation service provider can run over data it already ingests, with an alert threshold calibrated to a controlled per-window false-positive rate (Section 6).

1.4. Research Questions

The analysis is organized around four research questions.

RQ1. Are severe per-step rate-normalized anomalies in groundspeed, vertical rate, and track angle conditionally dependent within aircraft-hours, beyond what their marginal failure rates would predict?

RQ2. If so, is the dependence reducible to pairwise effects, or does it persist as three-way structure under formal conditional-independence testing?

RQ3. Does the dependence survive controls for the most plausible alternative explanations, data-link interval contamination, anomaly threshold choice, seasonal and geographic heterogeneity, airframe composition, phase-classification cut-offs, and receiver-side noise indexed by transponder-self-reported NACv?

RQ4. Are there identifiable temporal episodes during which the marginal anomaly prevalences are elevated, and if so, are the elevations spatially, diurnally, or operationally structured in ways that constrain candidate operational drivers?

1.5. Scope and Non-Attribution Statement

The analytical framework is *descriptive, non-causal, and non-attributional*. All analyses characterize the internal statistical consistency of reported ADS-B data as observed at the receiver. They do not infer (i) equipment malfunction or avionics hardware faults; (ii) specific root causes such as GNSS spoofing, RF jamming, or cyberattack; (iii) pilot behaviour or air navigation service provider performance deficiencies; or (iv) regulatory non-compliance or certification failures. The terms “anomaly” and “failure” refer throughout to deviations of the broadcast record from the rate-normalized threshold definitions of Section 3.2, not to any inferred operational event.

The findings are positioned as a *strategic infrastructure health indicator* for post-operational review, capacity planning, and certification oversight. They are explicitly not a real-time tactical input for ACAS X collision avoidance or for TBO conflict prediction, both of which require sub-second update latencies beyond the granularity of the hourly aggregations analyzed here.

None of the per-channel quality indicators (NACp, NACv, NIC, SIL) or the existing channel-by-channel integrity monitors discussed in Section 1.1 asks whether the kinematic channels are mutually consistent, whether a velocity anomaly, a vertical-rate anomaly, and a track-angle anomaly co-occur more often than independent channel faults would produce. This paper treats that joint consistency as a first-class integrity signal: one that is measurable from data already broadcast and reducible to a lightweight monitor. The five contributions enumerated above accordingly fall into two thrusts: the empirical characterization of the coupling, and the deployable monitor derived from it.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on multi-dimensional anomaly detection and conditional dependence inference in surveillance data. Section 3 details the data architecture, the rate-normalized anomaly definitions and their physical and regulatory anchoring, the cluster-bootstrap and cluster-robust inference framework, and the catalogue of pre-specified robustness analyses. Section 4 reports the headline pairwise dependence, the three-way decomposition, the permutation null test, the eight robustness analyses, and the Spring 2025 localization diagnostic. Section 5 interprets the mechanism implied by the conditional dependence structure and discusses the operational implications and the constraints on candidate drivers of the Spring 2025 elevation. Section 6 develops a deployable multi-channel consistency monitor from the finding and characterizes its operating point. Section 7 acknowledges the methodological constraints, and Section 8 summarizes the contributions and outlines outstanding work.

2. Literature Review

The expansion of crowdsourced cooperative-surveillance archives over the past decade has enabled extensive empirical research into ADS-B reliability. However, the majority of existing literature approaches kinematic inconsistency through single-metric lenses or localized detection algorithms. We organize the relevant prior work into three threads: single-metric anomaly detection, multi-metric co-occurrence reporting, and conditional dependence methods in adjacent sensor domains.

2.1. Single-Metric Anomaly Detection in ADS-B

Numerous studies have proposed anomaly-detection frameworks utilizing deep neural networks [12], Gaussian mixture models [13], and Kalman-filter residuals [14], with related two-phase methods ranking abnormalities in archived flight-operations data [24], to identify discrete GNSS-spoofing events or per-message trajectory deviations. While these models offer sensitive flagging of individual-channel corruption, they treat each kinematic dimension independently, frequently rely on heavily pre-filtered training datasets bounded to specific airspace sectors and short temporal windows [25], and discard incomplete or malformed packets prior to analysis [26, 27]. This pre-filtering introduces a documented survivorship bias that masks the systemic observability of the network at scale [28]. Architecturally, these detectors are predominantly *supervised* models trained to recognize specific injected-attack classes (spoofing, replay, message-injection, or trajectory-saturation) against labelled or simulated attack corpora, an approach continued in recent deep convolutional and recurrent ADS-B detectors [25, 29]; their decision boundary is consequently tied to the attack taxonomy present in the training data.

A complementary thread of work uses ADS-B as an instrument for diagnosing upstream GNSS health. Position-channel jumps have been correlated with NOTAM-reported jamming events and used to construct empirical anomaly geographies [15, 16], and aggregated kinematic anomalies have been linked to specific GNSS-disturbance episodes and to broader patterns of regional infrastructure stress [17, 30]. These studies, however, focus on single-dimensional (position) anomalies and do not characterize the joint multi-dimensional structure that is the subject of the present work.

2.2. Multi-Metric Co-occurrence in Surveillance Data

The concept of multi-dimensional kinematic-consistency scoring remains underdeveloped in the ADS-B literature. Composite reliability indices are well established in adjacent surveillance domains; the Navigation System Integrity parameter in GNSS monitoring combines multiple satellite ranging errors into a single protection-level metric [19, 31]. ADS-B itself emits transponder-self-reported quality categories (Navigation Accuracy Category for velocity (NACv), Navigation Accuracy Category for position (NACp), Navigation Integrity Category (NIC), and Source Integrity Level (SIL)) into the message-level records of the 1090 MHz extended squitter [19, 32]. The empirical literature has, however, primarily used these categories to filter data prior to anomaly analysis rather than as variables in a joint statistical model.

Co-occurrence of multi-dimensional anomalies has been reported descriptively in two prior strands of work. First, several studies have tabulated joint counts of position, velocity, and altitude flags within receiver-coverage samples or restricted airspace sectors [33, 27, 5]. Second, hardware-failure characterizations have noted that transponder corruption is rarely confined to a single channel [34]. Neither thread has posed the co-occurrence as a formal conditional-independence question with inferential controls, and neither has tested whether the apparent coupling survives removal of data-link-interval contamination or restriction to high-integrity message pairs.

2.3. Conditional Dependence Methods in Sensor Networks

Conditional-independence inference has a long methodological history in biostatistics, econometrics, and graphical-model learning, including the use of cluster-robust standard-error corrections for non-independent sampling [35, 36, 37] and within-cluster permutation tests [38, 39]. Applications to surveillance and sensor data have nevertheless remained sparse: anomaly detection in cooperative-surveillance networks has been treated overwhelmingly as a classification problem rather than as a distributional-consistency problem, despite the obvious clustering of observations within airframes and the resulting need for cluster-robust inference.

In reliability engineering specifically, the quantification of excess joint-failure probability beyond what independent marginal rates would predict has been formalized under the rubric of *common-cause failure* (CCF) and *dependent failure behaviour* analysis; a recent systematic review covers the field across nuclear, process-industry, and infrastructure domains [20]. Parametric CCF families include the beta-factor model, recently applied to safety-critical redundant infrastructure such as railway train-control on-board systems [21], and the alpha-factor model, recently extended to multi-group component assignments [22]. Non-parametric approaches based on copula functions model the joint failure distribution without restrictive assumptions on the marginals [23]. To our knowledge, however, none of these reliability-engineering frameworks has previously been applied to the multi-dimensional kinematic channels of cooperative-surveillance telemetry at planetary scale; the pairwise enrichment ratio R of Equation 4 is formally analogous to the ratio $\beta/(1-\beta)$ in the beta-factor model, reinterpreted as a conditional-probability amplification rather than a failure-fraction parameter.

The present study extends formal conditional-independence inference to the crowdsourced ADS-B surveillance archive, with a planetary-scale unit of analysis, an explicit pre-specified robustness catalogue, and an exact within-cluster permutation test that distinguishes structural multi-channel coupling from random pairwise concentration. In contrast to the supervised attack-class detectors above, the present approach is *unsupervised* and label-free: it characterizes the joint distributional consistency of the kinematic channels rather than learning a decision boundary against an enumerated attack set, and therefore requires neither labelled attack data nor an attack taxonomy fixed in advance. Table 1 summarizes how it differs from the prior threads along the dimensions that matter for a structural multi-channel claim.

Table 1: Positioning of the present study against prior approaches to ADS-B kinematic-anomaly analysis and to dependent-failure modelling (Section 2). "CI test" denotes a formal conditional-independence hypothesis test; "Rx/Tx control" denotes discrimination of receiver-side from broadcaster-side contributions.

Approach	**Channels**	**CI test**	**Scale**	**Rx/Tx control**	**Monitor**
Single-channel detectors [12, 13, 14]	One per model	No	Sector / window	No	No
GNSS-anomaly geographies [15, 17]	Position only	No	Regional	No	No
Multi-metric co-occurrence [33, 34]	Several (descriptive)	No (counts)	Varies	No	No
Reliability-engineering CCF [21, 22, 23]	Generic components	Parametric model	Component / system	N/A	No
This study	**Three (jointly)**	**Yes (exact permutation)**	**Planetary, 117.5M a-h**	**Yes (NACv/NIC/recv.)**	**Yes (two-index SPC)**

3. Methods

3.1. Data Source and Population

We analyze the calendar year 2025 (1 January 2025 through 31 December 2025 UTC) of OpenSky Network state-vector data [4, 5]. The OpenSky historical archive resamples raw 1090 MHz Mode S Extended Squitter messages to nominal 1 Hz aircraft state vectors, partitioned in Trino by hourly Unix timestamp. We extracted state vectors with `onground = false` and a valid six-character ICAO 24-bit aircraft identifier, excluding seven hex patterns corresponding to known anonymized, reserved, or test-allocation blocks: the five exact literals {`000000`, `FFFFFF`, `000001`, `123456`, `ABCDEF`} and the two prefix patterns {`0000%`, `FFFF%`} applied as SQL `LIKE` clauses against the OpenSky `state_vectors_data4` table. The two prefix patterns each cover 256 24-bit addresses (`000000`–`0000FF` and `FFFF00`–`FFFFFF`); these ranges are dominantly unassigned/reserved per the ICAO 24-bit aircraft address allocation scheme defined in Annex 10 Volume III, Part I, Chapter 9 and its Appendix ("Aircraft Addressing System") [40], with `FFFFFF` reserved for the all-call broadcast address, but a small minority of valid assignments may fall within them. The fraction of the active 2025 airframe population in these ranges is expected to be negligible given the size of the active block allocations elsewhere in the 24-bit space, but the exclusion is acknowledged as a minor source of selection in the population definition. Stage 2.1 extraction with the temporal-validity floor of Section 3.8 yields 117,542,773 aircraft-hour aggregates; after the low-latency screen of Section 3.5 (which removes 42 aircraft-hours, a fraction 3.6×10^{-7}) the headline analysis sample comprises:

- **Dataset 3A**: 117,542,731 aircraft-hour aggregates, clustered by 240,001 unique airframes.
- **Dataset 3B**: 791,862,868 non-overlapping 300-second tumbling windows derived from the same underlying state vectors.

Each airframe is enriched with manufacturer-issued metadata (aircraft type designator, engine type, airframe category) merged on `icao24` from a separately curated registry built from the OpenSky aircraft database [4], which provides the base `icao24`, type-designator, registration, and `icaoAircraftClass` fields, augmented with the FAA Aircraft Registry [41] and Transport Canada CARS [42] for tail-number provenance, and the Rikgale ICAO Aircraft Type designator mapping [43, 44] for typecode-to-engine and typecode-to-airframe-category resolution. Manufacturer-issued metadata covers 77.3% of unique airframes in the 2025 sample. For the calendar-aligned year-over-year comparison reported in Section 4.15 (Table 7), the 50-day window of Spring 2024 (2 April through 22 May 2024) is separately re-extracted from the same `state_vectors_data4` schema using the identical SQL pipeline, hex exclusions, on-ground filter, and rate-normalized flag definitions; details of this re-extraction are given in Section 3.11.

3.2. Kinematic Anomaly Indicators

We define three binary indicators for kinematic discontinuities in the broadcast trajectory.

3.2.1. Per-Step Rate-Normalized Definitions

For consecutive state vectors at times t_i, t_{i-1} from the same aircraft, with $\Delta t = t_i - t_{i-1}$:

$$\mathrm{DV}_i = \mathbf{1}\left[\frac{|v_i - v_{i-1}|}{\max(\Delta t,\, 0.1)} > 10\ \mathrm{m/s^2}\right], \tag{1}$$

$$\mathrm{DVR}_i = \mathbf{1}\left[\frac{|w_i - w_{i-1}|}{\max(\Delta t,\, 0.1)} > 8\ \mathrm{m/s^2}\right], \tag{2}$$

$$\mathrm{DH}_i = \mathbf{1}\left[\frac{|\Delta_{\mathrm{wrap}}(h_i, h_{i-1})|}{\max(\Delta t,\, 0.1)} > 15^\circ/\mathrm{s}\right], \tag{3}$$

where v denotes the transponder-reported ground velocity, w denotes the transponder-reported vertical rate (the OpenSky `vertrate` column of `state_vectors_data4`, predominantly barometric per DO-260B defaults but not filtered on the DF17 BDS 0,9 source bit; see Section 7.5), h denotes the track angle, Δ_{wrap} returns the shortest signed difference modulo 360°, and the $\max(\Delta t, 0.1)$ safety floor prevents division by zero in the rare case of duplicate timestamps in the 1 Hz resampled OpenSky state-vector product. Per the OpenSky Trino documentation, the `time` column of `state_vectors_data4` is rounded to integer seconds in the historical archive, so $\Delta t = \mathrm{time}_i - \mathrm{time}_{i-1}$ takes integer values ≥ 1 for all non-degenerate pairs and the floor activates only at $\Delta t = 0$. Per-step pairs are restricted to $0 < \Delta t \leq 60\,\mathrm{s}$ by the Stage 2.1 SQL filter (`dt_seconds > 0 AND dt_seconds <= 60`); the safety floor therefore never activates on any per-step pair that enters the rate computation, and the 0.1 constant is defensive code that cannot affect any reported indicator. Aircraft-hour-level indicators (DV, DVR, DH) are the per-step disjunctions over the hour: each indicator fires if any per-step pair within the hour exceeds the corresponding threshold.

We acknowledge that disjunctive hourly aggregation makes the marginal hour-level prevalence $P(\mathrm{DV} = 1)$ a non-linear function of the per-hour valid-pair count N_{pair} (specifically $1 - (1 - p)^{N_{\mathrm{pair}}}$ under per-pair independence); two aircraft-hours with identical per-pair anomaly rates but different N_{pair} will exhibit different hour-level marginals. Four distinct lines of evidence refute this confound as an explanation for the headline coupling result:

(i) the Navigation Accuracy Category for velocity (NACv) stratification of Section 4.10, computed on aircraft-hour aggregates whose stratification variable is the within-hour minimum of the message-level NACv, exhibits a *stronger* pairwise enrichment in the high-NACv stratum ($3.367\times$, NACv ≥ 3) than in the low/unknown-NACv stratum ($3.003\times$);

(ii) the conditional enrichment ratio $P(\mathrm{DV} \mid \mathrm{DVR})/P(\mathrm{DV})$ is mildly less sensitive to a common N_{pair} multiplier than its constituent marginals (the ratio attenuates by a factor of ≈ 2 across the empirical $N_{\mathrm{pair}} \in [60, 3599]$ range under plausible per-pair rates, while the marginals themselves vary by a factor of ≈ 8); this is a partial mitigation rather than an invariance;

(iii) the within-cluster permutation null of Section 3.9.5 preserves the empirical per-cluster $(N_{\mathrm{pair}}, \mathrm{DV}, \mathrm{DH})$ joint distribution and still places the observed V_B entirely outside the null (one-sided $p < 10^{-3}$; standardized gap $> 1{,}200$ null SD); and

(iv) the explicit cluster-robust logistic regression of Section 4.11 with log(total_pairs) as a covariate and a DVR $\times$ log(total_pairs) interaction yields a density-adjusted DVR odds ratio at the data median (8.01) statistically indistinguishable from the density-unadjusted Model A value (8.111), with the interaction coefficient (OR $= 0.946$, $[0.941, 0.951]$) indicating a small density-dependent attenuation in the *opposite* direction from the confound prediction.

Argument (iv) is the load-bearing counter-argument; arguments (i), (ii), and (iii) provide independent corroboration.

3.2.2. Physical and Regulatory Justification of Thresholds

The three thresholds were pre-specified in the project configuration file [45] before any analysis was performed and were not adjusted subsequently.

Longitudinal acceleration ($DV > 10$ $m/s^2 \approx 1.02\,g$).. Transport-category aircraft are certified to a limit manoeuvring load factor of $2.5\,g$ per 14 CFR Part 25.337 [46], but *sustained longitudinal* accelerations during routine cruise and climb rarely exceed $\sim 3\,\mathrm{m/s^2}$ outside of takeoff roll. We emphasize that the longitudinal axis is distinct from the vertical axis in routine turbulence: moderate-to-severe turbulence routinely produces transient $\pm 1\,g$ vertical accelerations but rarely produces $1\,g$ longitudinal accelerations, because the latter would require a one-second velocity change of order $10\,\mathrm{m/s}$ ($\approx 19\,\mathrm{kt}$). A per-step rate-normalized DV at the $10\,\mathrm{m/s^2}$ threshold is therefore unambiguously outside the routine flight envelope on the longitudinal axis specifically, and is consistent with sensor disagreement, transponder upset, GNSS-velocity-solution discontinuity, measurement-quantization artefact, or a genuine upset-recovery manoeuvre. The threshold is not designed to discriminate between these causes; it is designed to isolate the high-rate tail of the per-step longitudinal-acceleration distribution.

Vertical acceleration ($DVR > 8$ $m/s^2 \approx 0.82\,g$).. This threshold has a twofold justification. First, $0.8\,g$ vertical acceleration approaches the severe-upset boundary used in FAA Loss-of-Control accident analyses. Second, the relevant measurement-resolution floor is set by the broadcast vertical-rate field itself, not by the barometric-altitude encoding: the DO-260B Airborne Velocity message (BDS 0,9) encodes vertical rate in a 9-bit subfield with a least-significant-bit resolution of $64\,\mathrm{ft/min}$ ($\approx 0.33\,\mathrm{m/s}$) [19]. A consecutive-message vertical-rate jump exceeding $8\,\mathrm{m/s^2}$ over $\Delta t \leq 1\,\mathrm{s}$ implies a reported vertical-rate change of at least $8\,\mathrm{m/s}$ ($\approx 1{,}575\,\mathrm{ft/min}$, roughly $25\times$ the encoding LSB), far above the quantization floor; such a jump reflects a genuine large excursion in the broadcast vertical-rate channel rather than encoder round-off. (The $25\,\mathrm{ft}$ least-significant-bit of the barometric-altitude encoding governs the altitude field, which is not the field on which DVR is computed.)

Track-angle rate ($DH > 15°/s$).. The standard rate of turn under instrument flight rules is $3°/\mathrm{s}$ [47]. The chosen threshold of $15°/\mathrm{s}$ corresponds to “rate 5”, five times the standard rate, a turn rate achievable only in unusual attitudes, aerobatic flight, or transponder upset. For comparison, the maximum sustained bank angle in routine commercial operations corresponds to approximately $3.5°/\mathrm{s}$ at typical cruise speeds.

3.2.3. Why Rate-Normalized, Not Raw Differences

A natural alternative would be to flag absolute jumps in velocity, vertical rate, and heading without dividing by Δt. We reject this formulation. At the nominal 1 Hz output rate of OpenSky state vectors, $\Delta t = 1\,\mathrm{s}$ is modal but not universal: real-world Δt ranges from 1 to over 60 s depending on receiver coverage. A jump of 10 m/s in velocity is physically extreme over $\Delta t = 1\,\mathrm{s}$ (equivalent to $10\,\mathrm{m/s}^2$) but routine over $\Delta t = 30\,\mathrm{s}$ (equivalent to $0.33\,\mathrm{m/s}^2$). The position-derived alternative, differencing latitude and longitude to compute a ground speed, is methodologically unreliable in 1 Hz ADS-B because Compact Position Reporting (CPR) decoding and message-fragmentation artefacts can produce spurious large position deltas [4, 48]. We therefore use the transponder-reported velocity, vertical rate, and track angle exclusively, normalized by the observation interval, throughout the analysis.

3.3. Three Extraction Tracks

For each aircraft-hour we computed three parallel sets of indicators differing in how the per-pair time-difference distribution was handled.

Track 1 (Unfiltered).. All per-step pairs with $0 < \Delta t \leq 60\,\mathrm{s}$. The maximum-permissible-jump formulation: any large discontinuity, regardless of measurement interval, triggers the flag. Reported in sensitivity tables only; not the primary track.

Track 2 (Rate-Normalized; primary).. The rate-normalized definitions of Section 3.2, applied to all per-step pairs with $0 < \Delta t \leq 60\,\mathrm{s}$. By construction the indicators have units of acceleration ($\mathrm{m/s}^2$, $^\circ$/s), making them dimensionally invariant under variable receiver-coverage resampling and therefore comparable across geographies. This is the primary track for all reported results.

Track 3 (Phys-Restricted).. The same rate-normalized definitions, additionally restricted to pairs with $\Delta t \leq 5\,\mathrm{s}$. Acts as an additional filter against interval-confounded anomalies in which the rate-normalized denominator is large enough to mask a physically improbable jump.

Each aircraft-hour therefore carries nine indicator columns: $\{\text{Track}\} \times \{\text{DV}, \text{DVR}, \text{DH}\}$.

3.4. Data Quality Stratification

The fraction of per-step pairs within an aircraft-hour with $\Delta t = 1\,\mathrm{s}$ exactly, denoted $f_{\Delta t=1}$, summarizes receiver-coverage quality. We define three quality tiers:

- **PureDQ**: $f_{\Delta t=1} = 1.000$ (every pair sampled at the minimum interval).
- **HighDQ**: $0.999 \leq f_{\Delta t=1} < 1.000$.
- **LowDQ**: $f_{\Delta t=1} < 0.999$.

PureDQ aircraft-hours cannot, by construction, exhibit any apparent anomaly attributable to a long Δt between consecutive observations. The three-tier decomposition (Section 3.10) tests whether the conditional dependence of interest is monotonically attenuated as Δt-contamination increases. Results show that it is not.

3.5. Latency-Tier Screening Check

We define the latency tier of an aircraft-hour by its median per-step Δt: Low ($\leq 5\,\mathrm{s}$), Normal ($5 < \Delta t \leq 15\,\mathrm{s}$), High ($> 15\,\mathrm{s}$). The Dataset 3A temporal-validity floor of Section 3.8 (at least 60 valid per-step pairs) already excludes most pathological-coverage records, and as a screening check we verified that 99.99996% of aircraft-hours (117,542,731 of the 117,542,773 passing the temporal-validity floor) fall in the Low tier. We retain only Low-tier aircraft-hours throughout Section 4 for dimensional consistency with the rate-normalized indicators of Section 3.2; the filter is not substantively restrictive.

3.6. Pipeline Architecture

The extraction pipeline runs against the OpenSky Trino interface in five stages. Stage 1 builds an enriched aircraft metadata table cross-referenced across the FAA, Transport Canada, and Rikgale registries. Stage 2.1 produces Dataset 3A by aggregating per-step rate-normalized flags to aircraft-hour granularity, retaining median latitude/longitude, the latency tier, and $f_{\Delta t=1}$. Stage 2.2 produces Dataset 3B by aggregating to 300-second non-overlapping tumbling windows and classifying each window's manoeuvring envelope (Section 3.7). Stage 2.3 merges Dataset 3A with metadata and computes pairwise enrichment, conditional-independence violation ratios, and engine-type universality statistics. Stage 2.4 fits the phase-stratified logistic regression on Dataset 3B (Section 3.9.4). All stages write monthly Parquet shards with hourly resume granularity and atomic write semantics. The pipeline source code, configuration, intermediate manifests, complete embedded Trino SQL queries (Stages 2.1 and 2.2), and the present extraction products are released under an open licence alongside this paper for reproducibility [45]; see also the Data Availability Statement. No part of the pipeline relies on proprietary or commercial OpenSky queries; all data access is via the public Trino interface to the `state_vectors_data4` table.

3.7. Window-Level Phase Classification

For Stage 2.4 we classify each 300-second window into one of four manoeuvring categories based on the distribution of 15-second-median barometric vertical rates across the 20 non-overlapping tumbling sub-windows that tile each 300-second window:

- **Climb**: ≥ 13 of 20 sub-windows (strictly $> 60\%$) have sub-window median $w > 3.5\,\mathrm{m/s}$.
- **Cruise**: ≥ 13 sub-windows have $|w_{\mathrm{med}}| \leq 3.0\,\mathrm{m/s}$.
- **Descent**: ≥ 13 sub-windows have median $w < -5.5\,\mathrm{m/s}$.
- **Transition**: any window failing the ≥ 13-sub-window criterion for all three above categories (typically level-off, configuration change, or non-stationary vertical profile). This category functions as a residual catch-all: highly volatile profiles that split sub-windows between climbing and descending sub-bands also fall here, which makes the Transition stratum heterogeneous by design. The headline regression interpretation (Section 4.7) accordingly emphasizes the Climb, Descent, and DH main effects and the Climb $\times$ DH and Descent $\times$ DH interactions, while treating the Transition $\times$ DH coefficient as less stable; the symmetric-threshold sensitivity analysis of Section 4.13 confirms this stability ordering empirically.

The three duration criteria are mutually exclusive by construction (the threshold sets $\{w > 3.5\}$, $\{|w| \leq 3.0\}$, and $\{w < -5.5\}$ are disjoint) so each window is unambiguously assigned to exactly one category.

The asymmetry between the climb threshold ($+3.5\,\mathrm{m/s}$) and the descent threshold ($-5.5\,\mathrm{m/s}$) reflects standard commercial operating profiles, in which descent rates are typically steeper than climb rates due to thrust-spool-down rather than engine-power constraints. A sensitivity analysis with symmetric thresholds at $\{\pm 2.5, \pm 3.5, \pm 4.5\}$ m/s confirms qualitative invariance of the regression results (Section 4.13).

3.8. Datasets and Temporal Validity

Dataset 3A retains aircraft-hours with at least 60 valid per-step pairs (Δt defined) within the hour. Dataset 3B retains 300-second tumbling windows with at least 120 seconds of total duration and at least 60 valid per-step vertical-rate observations. These thresholds are designed to exclude transient observations of an aircraft that crosses a receiver's coverage briefly without permitting meaningful per-hour or per-window characterization.

3.9. Statistical Framework

3.9.1. Pairwise Enrichment

The primary descriptive statistic is the conditional-probability enrichment ratio

$$R(\mathrm{DV} \mid \mathrm{DVR}) \;=\; \frac{P(\mathrm{DV}=1 \mid \mathrm{DVR}=1)}{P(\mathrm{DV}=1)}, \tag{4}$$

the factor by which the marginal probability of a longitudinal-acceleration anomaly is amplified, conditional on a co-occurring vertical-acceleration anomaly in the same aircraft-hour. The two complementary enrichment ratios $R(\mathrm{DV} \mid \mathrm{DH})$ and $R(\mathrm{DVR} \mid \mathrm{DH})$ are defined analogously.

3.9.2. Two-Pathway Conditional Decomposition

We test three conditional-independence relationships among the three flags:

$$V_A \;=\; \frac{P(\mathrm{DV}, \mathrm{DVR}, \mathrm{DH})}{P(\mathrm{DV})\, P(\mathrm{DH} \mid \mathrm{DV})\, P(\mathrm{DVR} \mid \mathrm{DV})}, \tag{5}$$

$$V_I \;=\; \frac{P(\mathrm{DV}, \mathrm{DVR}, \mathrm{DH})}{P(\mathrm{DVR})\, P(\mathrm{DV} \mid \mathrm{DVR})\, P(\mathrm{DH} \mid \mathrm{DVR})}, \tag{6}$$

$$V_B \;=\; \frac{P(\mathrm{DV}, \mathrm{DVR}, \mathrm{DH})}{P(\mathrm{DH})\, P(\mathrm{DV} \mid \mathrm{DH})\, P(\mathrm{DVR} \mid \mathrm{DH})}. \tag{7}$$

V_A is the violation ratio for $\mathrm{DH} \perp \mathrm{DVR} \mid \mathrm{DV}$; V_I for $\mathrm{DV} \perp \mathrm{DH} \mid \mathrm{DVR}$; V_B for $\mathrm{DV} \perp \mathrm{DVR} \mid \mathrm{DH}$. The subscripts A, I, B are arbitrary pathway labels and carry no meaning beyond identifying the conditioning channel of each ratio: V_A holds the longitudinal-acceleration channel DV fixed and tests whether the vertical-rate and track-angle channels remain dependent given it, V_I conditions on the vertical-rate channel DVR, and V_B conditions on the track-angle channel DH. Each ratio equals unity under the corresponding conditional independence. Sustained deviation from unity in any pathway indicates irreducible three-way structure in the joint distribution that cannot be explained by pairwise effects alone. Put plainly, each ratio asks a single question: once the third channel is held fixed as a possible shared cause, do the other two anomaly types still occur together more often than chance would predict? A value of one means no, the apparent link was fully explained by the third channel, while a value above one means the two remain genuinely coupled even after the third is accounted for, which is what "irreducible three-way structure" denotes. The three pathways are reported symmetrically and on equal footing throughout the present section; the mechanistic interpretation of why V_B is the largest of the three is deferred to Section 5.2, where it is justified *post hoc* from the data, not assumed in the definitions.

3.9.3. Cluster-Bootstrap Confidence Intervals

Within-airframe correlation precludes treating aircraft-hours as independent. We therefore quantify uncertainty for all aggregate statistics using a non-parametric cluster bootstrap, resampling the 240,001 unique `icao24` airframes with replacement; each bootstrap iteration computes the statistic on the full set of aircraft-hours associated with the resampled airframes. We report empirical 95% percentile intervals from 1,000 iterations. All Section 4 subgroup analyses use the same cluster-bootstrap framework. To minimize overhead at production scale, the bootstrap operates on per-cluster sufficient statistics (the eight integer counters $\{n,\ n_{\mathrm{DV}},\ n_{\mathrm{DVR}},\ n_{\mathrm{DH}},\ n_{\mathrm{DV,DVR}},\ n_{\mathrm{DV,DH}},\ n_{\mathrm{DVR,DH}},\ n_{\mathrm{DV,DVR,DH}}\}$ summed within each airframe) rather than the raw aircraft-hour records.

3.9.4. Cluster-Robust Logistic Regression

The phase-stratified regression in Stage 2.4 fits

$$\operatorname{logit} P(\mathrm{DV}=1 \mid \mathrm{DVR}=1) \;=\; \beta_0 + \boldsymbol{\beta}_{\text{phase}} + \beta_{\mathrm{DH}}\,\mathrm{DH} + \boldsymbol{\beta}_{\text{phase}\times\mathrm{DH}}, \quad (8)$$

to the $\mathrm{DVR}=1$ sub-population of Dataset 3B, with `Cruise` as the reference phase. Standard errors are clustered by `icao24` per the Cameron–Miller cluster-robust correction [35]; variance inflation factors are reported in Section 4.7.

3.9.5. Permutation Null Model

A descriptive deviation of V_B from unity is not, by itself, evidence that the joint distribution cannot be reproduced by random pairwise structure within airframes. We therefore construct an exact within-cluster permutation null distribution. The idea is simple: we repeatedly rebuild the dataset after randomly reshuffling which vertical-rate-anomaly flags fall in which hours within each aircraft, which breaks any genuine association with the other two channels while leaving each channel's overall anomaly rate and each aircraft's hour count exactly as observed. If the measured coupling is stronger than all one thousand reshuffled datasets, it cannot be an accident of the individual channel rates. Concretely, for each airframe c with n_c aircraft-hours, let $n_{\mathrm{DV},c}$, $n_{\mathrm{DVR},c}$, $n_{\mathrm{DH},c}$ denote the within-cluster marginal counts and $n_{\mathrm{DV,DH},c}$ the $(\mathrm{DV}=1, \mathrm{DH}=1)$ joint count. We sample, vectorized across clusters and iterations, the number of $\mathrm{DVR}=1$ rows falling in each of the four $(\mathrm{DV}, \mathrm{DH})$ cells from the multivariate hypergeometric distribution, by three sequential univariate hypergeometric draws. This exactly reproduces the distribution that would result from uniform random permutation of the DVR vector within each cluster, preserving the per-cluster marginals $P(\mathrm{DV})$, $P(\mathrm{DVR})$, $P(\mathrm{DH})$ and the within-cluster $(\mathrm{DV}, \mathrm{DH})$ pairwise structure while destroying the $(\mathrm{DV}, \mathrm{DVR})$, $(\mathrm{DVR}, \mathrm{DH})$, and triple structure. We accumulate 1,000 such iterations and report the observed V_A, V_I, V_B, and $R(\mathrm{DV} \mid \mathrm{DVR})$ against the resulting null mean, standard deviation, 95% percentile interval, z-score, and one-sided permutation p-value. The test operates on per-cluster integer sufficient statistics $(\{n_c, n_{\mathrm{DV},c}, n_{\mathrm{DVR},c}, n_{\mathrm{DH},c}, n_{\mathrm{DV,DH},c}\})$ rather than the raw aircraft-hour records, and therefore tests the conditional dependence given each airframe's overall marginal counts rather than each airframe's hour-by-hour message-density variation. The test is valid for its specific scope (within-cluster random reassignment of the DVR vector) and does not isolate residual coupling attributable to within-cluster variation in per-hour valid-pair count N_{pair}; the latter is discussed in Sections 3.2 and 7.3.

3.9.6. Multiple-Testing Correction

Where multiple hypothesis tests are reported within a single analytic family (the eight robustness analyses of Section 3.10, the 5×6 space-weather correlation grid of Section 3.11, and the per-region and per-season stratifications) we report uncorrected p-values for transparency and additionally apply the Benjamini–Hochberg false-discovery-rate correction [49] within each family at a ceiling of $q = 0.05$. The headline permutation result of Section 3.9.5

is reported as the empirical permutation rank (the observed value lies outside all 1,000 iterations, one-sided $p < 10^{-3}$), accompanied by the standardized gap from the null mean as a descriptive effect size. The empirical rank already lies below any per-family multiplicity threshold for the small number of tests reported here and is therefore robust to family choice.

3.10. Pre-Specified Robustness Analyses

To test the headline coupling against the most plausible alternative explanations, we ran eight robustness analyses on the same datasets. The set was pre-specified before the regression in Equation 8 was fitted, and all are reported across the Results (Section 4) regardless of outcome:

1. **Three-tier data-quality decomposition.** Re-computes the pairwise enrichment and all three violation ratios within each of the $\{\text{PureDQ}\}$, $\{\text{HighDQ}\}$, and $\{\text{LowDQ}\}$ tiers, with cluster-bootstrap 95% confidence intervals.

2. **DVR threshold sensitivity sweep.** Re-computes enrichment at DVR $> \{2, 5, 8, 10, 12\}\,\text{m/s}^2$ using the alternate-threshold columns pre-extracted in Stage 2.1.

3. **Seasonal robustness.** Repeats the headline analysis on four one-week windows distributed across the calendar year (mid-January, early April, early July, early October).

4. **Geographic robustness.** Stratifies aircraft-hours into six continental bounding boxes by median latitude and longitude.

5. **Strict Fixed-Wing regression.** Re-fits Equation 8 with airframe category restricted strictly to `Fixed-Wing`, excluding the `Other` category of gliders, drones, and unclassified airframes.

6. **Rotary-wing saturation grid.** Computes empirical and model-predicted $P(\text{DV} = 1)$ across the full phase $\times$ DH factorial for rotary-wing aircraft, to discriminate substantive interaction effects from logistic-regression ceiling artefacts.

7. **Transponder Integrity and Receiver Redundancy stratification.** To systematically evaluate whether the observed coupling stems from receiver-side artifacts or degraded transponder hardware, we group the aircraft-hour aggregates into a formal five-part testing matrix:

- **S1. NACv Stratification:** Tiers by the minimum reported Navigation Accuracy Category for velocity. The high-integrity tier is $nacv\_velocity\_min \geq 3$, corresponding to a horizontal velocity error strictly $< 1\,\text{m/s}$ per RTCA DO-260B Table 2-25.
- **S2. NIC Stratification:** Tiers by the minimum position integrity category ($nic\_position\_min \geq 8$, enforcing a horizontal containment radius $HCR < 0.05$ NM).
- **S3. Joint Strict Subset:** Isolates the concurrent intersection of high-accuracy transponder reports (NACv $\geq 3 \wedge$ NIC ≥ 8), defining the "DO-260B Clean" baseline.
- **S4. Receiver-Redundancy Stratification:** Partitions observations by unique ground-station tracking density (from single-receiver cells up to high-redundancy clusters of ≥ 10 distinct receivers).
- **S5. Triple-Intersection Stratum:** A custom high-integrity filter requiring simultaneous high NACv, high NIC, and very high receiver redundancy.

8. **Phase-classification symmetric-threshold sensitivity.** Re-fits Equation 8 on a Stage 2.2 re-extraction using symmetric vertical-rate thresholds at $\{\pm 2.5, \pm 3.5, \pm 4.5\}$ m/s in place of the primary asymmetric definition.

The permutation null model of Section 3.9.5 additionally quantifies the magnitude of V_B relative to the distribution expected under random within-airframe pairwise structure.

3.11. Temporal Anomaly Diagnostic

To support the seasonal robustness analysis (item 3 above), we additionally ran a per-day breakdown of the marginal anomaly prevalences and pairwise enrichment across the full 365 days of 2025, followed by a state-machine walk identifying contiguous elevation intervals (start: $P(\text{DV})$ crosses 0.30 upward; end: returns below 0.25). Within any identified elevation interval, we further computed UTC hour-of-day, geographic, and airframe-category stratifications against a control window comprising the seven days immediately preceding the elevated checkpoint week (31 March–6 April 2025). This control window overlaps the first five days (2–6 April) of the broader elevation interval, which makes the resulting within-elevation contrasts conservative rather than inflationary. The control window was selected before the diagnostic was run.

We additionally downloaded the NOAA Storm Events Database [50] for the 2025 calendar year and filtered to severe convective categories (Tornado, Hail, Thunderstorm Wind, High Wind, Heavy Rain, Hurricane/Typhoon, Tropical Storm, Flash Flood, Flood, Strong Wind) within the elevation interval and the surrounding context window. NOAA Storm Events coverage is restricted to US territory and therefore constrains the storm correlation analysis to North America; international convective activity is not represented.

To distinguish a recurring seasonal phenomenon from a 2025-specific event, we re-extracted the full calendar-aligned 50-day window of Spring 2024 (2024-04-02 00:00 UTC through 2024-05-22 00:00 UTC, calendar-aligned to the 2025 elevation window identified in Section 4.15; extracted 2026-05-22)[1] directly from the `state_vectors_data4` message archive using a SQL pipeline that reproduces the rate-normalized DV/DVR/DH flag definitions of Equations 1–3 exactly, with the same $0 < \Delta t \leq 60\,\mathrm{s}$ per-step filter, the same hex exclusions, and the same `onground = false` constraint. The pipeline aggregates the resulting per-pair anomaly flags to aircraft-hour granularity matching the Dataset 3A schema. The output is written as monthly Parquet shards and used in Section 4.15 to construct the year-over-year comparison reported in Table 7. The 2024 re-extraction is reported only at the marginal-prevalence and pairwise-coupling level (Table 7); we did not replicate the full data-quality-tier (Table A.1), NACv, NIC, or receiver-redundancy stratification (Table 5) of Sections 4.5 and 4.10 on the 2024 extract, because the purpose of the 2024 comparison is to test the recurring-seasonal-phenomenon hypothesis at the year-baseline level rather than to characterize the full robustness battery for that earlier window. We acknowledge that this asymmetric depth limits the strength of the year-over-year contrast: the comparison demonstrates that 2024 marginal prevalences and coupling are at baseline rather than elevated, but does not exclude the possibility that the underlying transponder population or receiver-network composition differed in ways that would matter

[1]The OpenSky historical archive is periodically reprocessed; re-extracting this 2024 window at a later date can return several percent fewer aircraft-hours and distinct airframes (with correspondingly higher marginal prevalences). Between the 2026-05-22 and 2026-06-18 extractions the yield fell $\approx 7\%$ in aircraft-hours and $\approx 16\%$ in airframes (the reprocessing preferentially pruning sparsely-observed airframes), without materially affecting the year-baseline contrast. The extraction date is reported for reproducibility.

for a more granular stratification.

To test whether the elevation coincides with elevated geomagnetic or solar activity, we additionally downloaded the GFZ Potsdam daily geomagnetic and solar indices archive [51, 52], which provides the canonical K_p, a_p, A_p, sunspot number S_N, and 10.7 cm solar radio flux $F_{10.7}$ for every UT day since 1932-01-01. We restricted to calendar years 2024 and 2025, joined the daily indices on date with the D1 daily breakdown (Section 4.15), and ran four cross-comparisons:

(i) Welch t-tests of each index across the elevation window versus the non-elevation portion of 2025;

(ii) Pearson and Spearman daily correlations between each space-weather index and each of $P(\mathrm{DV})$, $P(\mathrm{DVR})$, $P(\mathrm{DH})$, enrichment, and V_B, with percentile-bootstrap 95% confidence intervals;

(iii) mean $P(\mathrm{DV})$ on geomagnetically active ($K_p^{\max} \geq 4$, NOAA G0 active) and stormy ($K_p^{\max} \geq 5$, NOAA G1 minor storm) days versus quiet days; and

(iv) characterization of the top ten 2025 days by $P(\mathrm{DV})$ against their concurrent space-weather conditions.

4. Results

4.1. Population and Sample Characteristics

Stage 2.1 extraction yielded 117,542,773 aircraft-hour aggregates and Dataset 3B of 791,862,868 non-overlapping 300-second tumbling windows. The low-latency filter (Section 3.5) retained 99.99996% of these, 117,542,731 aircraft-hours across 240,001 unique airframes, which constitute Dataset 3A for all headline analyses. The PureDQ tier ($f_{\Delta t=1} = 1.000$) comprises 103,798,609 aircraft-hours (88.3%) across 239,087 airframes. Fixed-wing airframes contribute the dominant fraction of low-latency aircraft-hours. The rotary-wing share is approximately 3.0% (3,554,852 aircraft-hours); a residual $\approx 2.6\%$ (3,088,331 aircraft-hours) carries no airframe-category label after the metadata merge, and a further $\approx 9.8\%$ (11,540,069 aircraft-hours) carries airframe but not engine category.[2] The Fixed-Wing-and-classified-engine subset re-

[2] The per-category drop counts are recorded in the released deposit [45] under `stage2_3_enrichment_summary.json` (fields `universality_filter_context.rotary_wing_drop`

tained for the engine-universality analysis (Section 4.9) contains 102,445,034 aircraft-hours (87.2% of low-latency rows). Within this subset, Jet, turboprop, piston, and electric engine types contribute 79.1, 6.7, 16.6, and 0.002 million aircraft-hours respectively.

4.2. Pairwise Conditional Dependence

Within the full low-latency sample, the marginal probability of a severe longitudinal-acceleration anomaly is $P(\mathrm{DV}) = 0.221$. Conditional on the simultaneous presence of a severe vertical-acceleration anomaly, $P(\mathrm{DV} \mid \mathrm{DVR} = 1) = 0.696$, yielding a pairwise enrichment of 3.142× with 95% cluster-bootstrap interval [3.133, 3.152] ($\chi^2 = 1.92 \times 10^7$, $p < 10^{-300}$). Restricting the analysis to the PureDQ tier yields a confound-controlled enrichment of 3.156×, 95% CI [3.146, 3.166], demonstrating that the coupling is not an artefact of Δt-conflation introduced by dropped packets. The complementary marginal probabilities are $P(\mathrm{DVR}) = 0.111$ and $P(\mathrm{DH}) = 0.265$ in the full sample; the triple-joint probability is $P(\mathrm{DV}, \mathrm{DVR}, \mathrm{DH}) = 0.0547$.

4.3. Two-Pathway Conditional Decomposition

A pairwise enrichment of 3.142× could in principle reflect either irreducible three-way structure in the joint distribution or pairwise effects aggregated across the three flag pairs. To distinguish the two we evaluate the three conditional-independence violation ratios of Section 3.9.2. Results for both the full low-latency sample and the PureDQ subset are reported in Table 2; all three exceed unity in both samples. The Pathway B violation $V_B = 1.543$ (full) $= 1.557$ (PureDQ) is the largest, demonstrating that DV and DVR remain non-independent within strata of DH. Critically, the violation *intensifies* from 1.543× to 1.557× when restricting to PureDQ (Δt-clean) aircraft-hours, the opposite of what would be expected under a data-link-gap-induced spurious-coupling explanation.

The Pathway I violation $V_I = 1.149$ is, in contrast, only modestly above unity. This reveals that DV and DH are quasi-independent within the DVR = 1 stratum, consistent with DVR acting as a partial common antecedent of both DV and DH: the marginal DV–DH enrichment of approximately 2.34×

---

ped, `unclassified_airframe_dropped`, and `degenerate_engine_dropped`). The three drop categories are non-disjoint: a small fraction of rows are excluded by more than one criterion simultaneously (e.g. a Rotary-Wing aircraft with degenerate engine classification), so the retained fraction exceeds 100% minus the simple sum of the three drop percentages.

(computed from the marginals and joint of Section 4.2) substantially attenuates to 1.15× under conditioning on DVR. The implied mechanism is discussed in Section 5.

Table 2: Conditional-independence violation ratios in the full low-latency sample and the PureDQ ($f_{\Delta t=1} = 1.000$) subset (Section 4.3).

Pathway	Independence tested	Full	PureDQ
V_A	DH $\perp$ DVR | DV	1.141	1.135
V_I	DV $\perp$ DH | DVR	1.149	1.149
V_B	DV $\perp$ DVR | DH	**1.543**	**1.557**

Note. Cluster-bootstrap 95% confidence intervals were computed. Ratio of 1 corresponds to conditional independence. Ratios reported to three decimal places.

4.4. Permutation Null Test of Three-Way Structure

The within-cluster permutation null distribution (Section 3.9.5; 1,000 iterations) yields the comparison in Table 3 and Figure 1. Neither the observed enrichment (3.142×) nor the observed Pathway B violation ($V_B = 1.543$) falls anywhere within the entire 1,000-iteration null distribution (one-sided permutation $p < 10^{-3}$ in both cases). As descriptive effect sizes, the two observed values lie approximately 4,127 and 1,209 null standard deviations from their respective null means. Because a 1,000-iteration null cannot calibrate a Gaussian tail of that depth, these standardized gaps are reported as distances rather than as significance levels, and the operative test is the empirical permutation rank. The random-pairwise-structure null hypothesis for the joint distribution is decisively rejected.

The Pathway I observed value $V_I = 1.149$ falls well below the null mean of 2.099, a standardized gap of approximately 1,410 null standard deviations in the *negative* direction (one-sided permutation $p = 1.0$). The null mean is approximately the marginal DV–DH enrichment that emerges when DVR is randomized; the observed value is well below that, which directly demonstrates that conditioning on DVR substantially reduces the apparent dependence between DV and DH, again consistent with DVR as a partial common antecedent.

4.5. Three-Tier Data-Quality Decomposition

The pairwise enrichment falls steadily as data-link contamination rises across the three quality tiers, from 3.156× in the cleanest tier (PureDQ) to

Table 3: Permutation null model results (Section 4.4).

Statistic	Observed	Null mean	Null SD	Std. gap (null SD)
Enrichment R(DV | DVR)	3.142	1.209	0.00047	4,127
V_A (DH $\perp$ DVR | DV)	1.141	1.101	0.00035	115
V_I (DV $\perp$ DH | DVR)	1.149	2.099	0.00067	−1,410
V_B (DV $\perp$ DVR | DH)	1.543	1.085	0.00038	1,209

Note. Observed values are from the full low-latency sample; the null distribution is constructed by exact within-cluster random reassignment of the DVR vector (1,000 iterations). The final column is the standardized distance of the observed value from the null mean, in units of the null SD, a descriptive effect size, not a calibrated tail probability (a 1,000-iteration null cannot calibrate a multi-thousand-SD Gaussian tail). The operative significance statement is the one-sided permutation rank: $p < 10^{-3}$ for R, V_A, V_B. The negative gap for V_I is interpreted in the text.

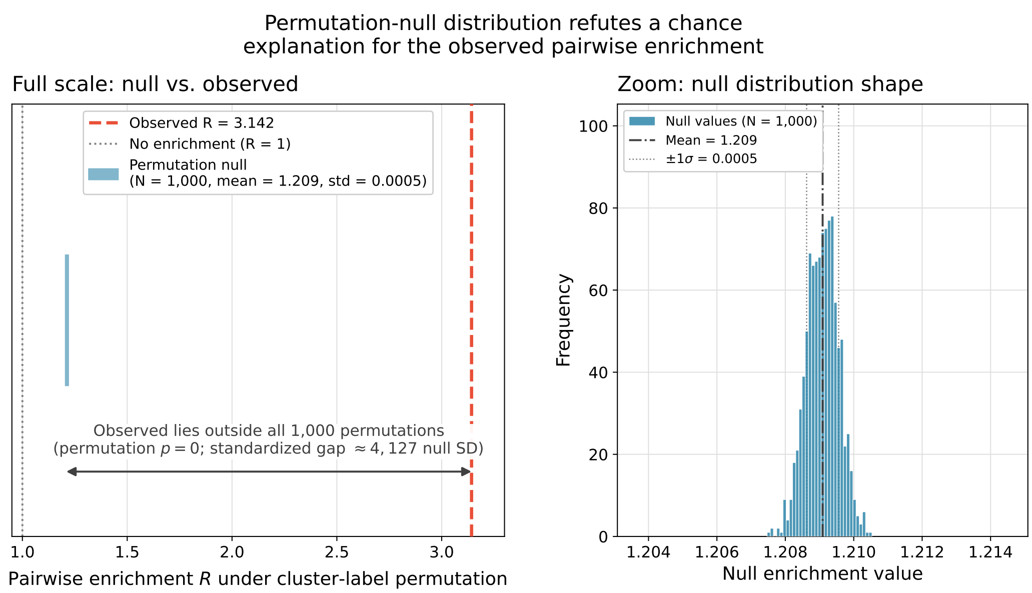

Figure 1: Permutation null distribution for the pairwise enrichment statistic. *Left panel*: full scale showing the cluster-label-permuted null distribution (blue range bar, mean 1.209 ± 0.0005) and the observed value ($R = 3.142$, red dashed line). *Right panel*: zoomed histogram of the 1,000 null values showing approximate normality around the mean with the $\pm 1\sigma$ marker at ± 0.0005. Construction in Section 3.9.5.

$2.783\times$ in the most contaminated (LowDQ), and the Pathway-B violation stays above unity in every tier. This gradient runs opposite to a Δt-conflation artefact: if data-link gaps were spuriously inflating the rate-normalized anomalies, the enrichment would be *highest* in the dirtiest tier, not the cleanest. The per-tier values and confidence intervals are reported in Appendix Appendix A.1 (Table A.1).

4.6. DVR Threshold Sensitivity Sweep

Recomputing the enrichment at five alternate DVR thresholds shows it rising monotonically as the threshold tightens, from $1.56\times$ at $2\,\mathrm{m/s^2}$ to $3.59\times$ at $12\,\mathrm{m/s^2}$. The coupling therefore strengthens precisely as the vertical-rate anomaly becomes more physically extreme, as a genuine physical coupling would, which rules out both a threshold-dependent artefact and a per-hour message-density confound. It is also near-invariant across the three indicator definitions carried in the deposit (raw absolute-difference, rate-normalized, and $\Delta t \leq 5\,\mathrm{s}$ restricted), so the result is not an artefact of any single operationalization (Appendix Appendix A.1, Table A.5).

4.7. Phase-Stratified Logistic Regression

We fit the three-stage cluster-robust logistic regression of Equation 8 on the DVR = 1 Fixed-Wing sub-population of Dataset 3B (N = 20,236,462 windows; 20,794,928 total DVR = 1 windows including rotary-wing). Variance inflation factors are well below conventional concern thresholds (DH = 1.05; is_climb = 1.44; is_descent = 1.30; is_transition = 1.43). Results across all three model specifications and the rotary-wing counterfactual are reported in Table 4.

Model 1, the marginal-risk specification, demonstrates that vertical manoeuvring imposes substantial baseline stress on the telemetry output relative to stable cruise. Relative to the Cruise reference category, Climb exhibits an odds ratio of 2.35 (95% CI $[2.31, 2.38]$), Descent 1.65 $[1.62, 1.68]$, and Transition 1.37 $[1.35, 1.39]$.

Model 2 introduces concurrent track-angle anomalies. The track-angle indicator emerges as a substantial concurrent predictor (OR = 4.15, $[4.10, 4.20]$), and the phase odds ratios increase consistently (Climb OR = 3.43; Descent OR = 2.45; Transition OR = 1.35).

Model 3 adds the phase $\times$ DH interactions. Climb remains a substantial baseline predictor (OR = 3.12, $[3.07, 3.17]$), DH retains a strong main effect (OR = 3.91, $[3.82, 4.00]$), and the Climb × DH interaction is super-multiplicative for fixed-wing aircraft (OR = 1.32, $[1.29, 1.35]$), with $p < 10^{-100}$. The Descent × DH interaction is similarly super-multiplicative (OR = 1.17, $p \approx 1.6 \times 10^{-30}$), and the Transition × DH interaction is negligibly sub-multiplicative (OR = 0.94).

Table 4: Cluster-robust logistic regressions predicting concurrent groundspeed divergence (DV = 1) given a vertical anomaly (DVR = 1) (Section 4.7).

Variable	**Model 1 (Marginal)**	**Model 2 (Adjusted)**	**Model 3 (Fixed-Wing)**	**Model 3B (Rotary-Wing)**
Intercept	1.03 [1.01, 1.04]	0.42 [0.41, 0.43]	0.44 [0.43, 0.45]	0.08 [0.07, 0.08]
Climb	2.35 [2.31, 2.38]	3.43 [3.39, 3.48]	3.12 [3.07, 3.17]	10.54 [7.83, 14.18]
Descent	1.65 [1.62, 1.68]	2.45 [2.42, 2.49]	2.30 [2.26, 2.34]	4.66 [3.86, 5.62]
Transition	1.37 [1.35, 1.39]	1.35 [1.34, 1.37]	1.40 [1.38, 1.42]	3.39 [3.11, 3.70]
Track angle (DH)	—	4.15 [4.10, 4.20]	3.91 [3.82, 4.00]	14.00 [12.55, 15.63]
Climb × DH	—	—	1.32 [1.29, 1.35]	0.23 [0.17, 0.31]
Descent × DH	—	—	1.17 [1.14, 1.20]	0.21 [0.17, 0.25]
Transition × DH	—	—	0.94 [0.92, 0.96]	0.28 [0.25, 0.31]

Note. Reference phase category is Cruise. Cluster-robust standard errors with airframe (`icao24`) clustering. N = 20,236,462 fixed-wing windows; N = 558,466 rotary-wing windows. All odds ratios significant at $p < 0.001$ unless otherwise noted; 95% CIs in brackets.

A strict-Fixed-Wing replication (excluding the residual `Other` category of unclassified airframes, gliders, and drones) yields odds ratios within approximately 1% of Table 4's Model 3 column (Climb 3.10, Descent 2.28, DH 3.95, Climb × DH 1.32; the largest deviation is the DH main effect, 3.95 versus 3.91, a 1.0% difference). The inclusion of unclassified airframes does not materially affect the fixed-wing regression.

4.8. Rotary-Wing Counterfactual

Rotary-wing aircraft show a different topology from fixed-wing: a very large heading main effect (14.0×) and *sub-multiplicative* phase×heading interactions (Model 3B, Table 4), rather than the super-multiplicative fixed-wing pattern. A logistic ceiling effect is the natural first suspect, but tabulating the empirical $P(\mathrm{DV})$ across the eight phase × DH cells shows it ranging from 0.071 to 0.721, well below saturation in every cell (Figure 2), which excludes the ceiling explanation independently of the model fit. The mechanism is discussed in Section 5.3, and the full cell-by-cell grid is reported in Appendix Appendix A.3 (Table A.10).

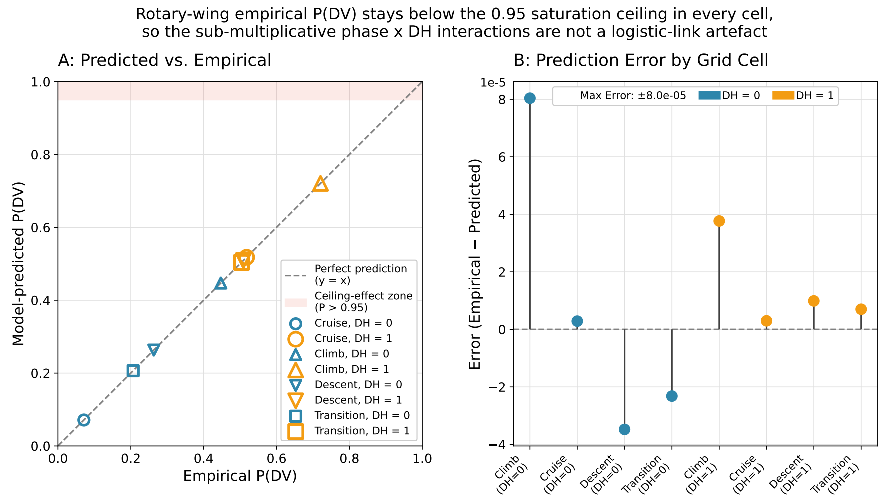

Figure 2: Empirical versus Model 3B-predicted $P(\mathrm{DV})$ for rotary-wing aircraft across the eight-cell phase $\times$ DH factorial. Marker shape encodes flight phase (Cruise: ●; Climb: ▲; Descent: ▼; Transition: ■); colour encodes the heading-instability flag (DH$=0$: blue; DH$=1$: orange). The red band at the top denotes the saturation-ceiling zone ($P > 0.95$); see Section 4.8 for interpretation.

4.9. Engine-Type and Geographic Universality

Figure 3 summarizes the pairwise enrichment R across every stratification examined, data-quality tier, airframe class, engine type, geographic region, and temporal window. The enrichment is substantial and well above the permutation-null level (Section 3.9.5) in every stratum, establishing the coupling as a pervasive property of cooperative-surveillance telemetry rather than an artefact of any single airframe class, engine lineage, airspace, or data-quality tier. Its magnitude is not constant at the headline 3.14, however: across structural strata R ranges from approximately $2.25\times$ (rotary-wing airframes) to $3.80\times$ (Oceania), with the full-sample value near the centre of that range and the dominant data-quality, fixed-wing, and jet strata clustered around it. The lowest-enrichment strata (rotary-wing, piston, and the South American and Asian airspaces, together with the Spring 2025 elevated week) coincide with elevated marginal anomaly prevalences that mechanically compress the $P(\mathrm{DV} \mid \mathrm{DVR})/P(\mathrm{DV})$ ratio, for the reason explained in Section 4.15.

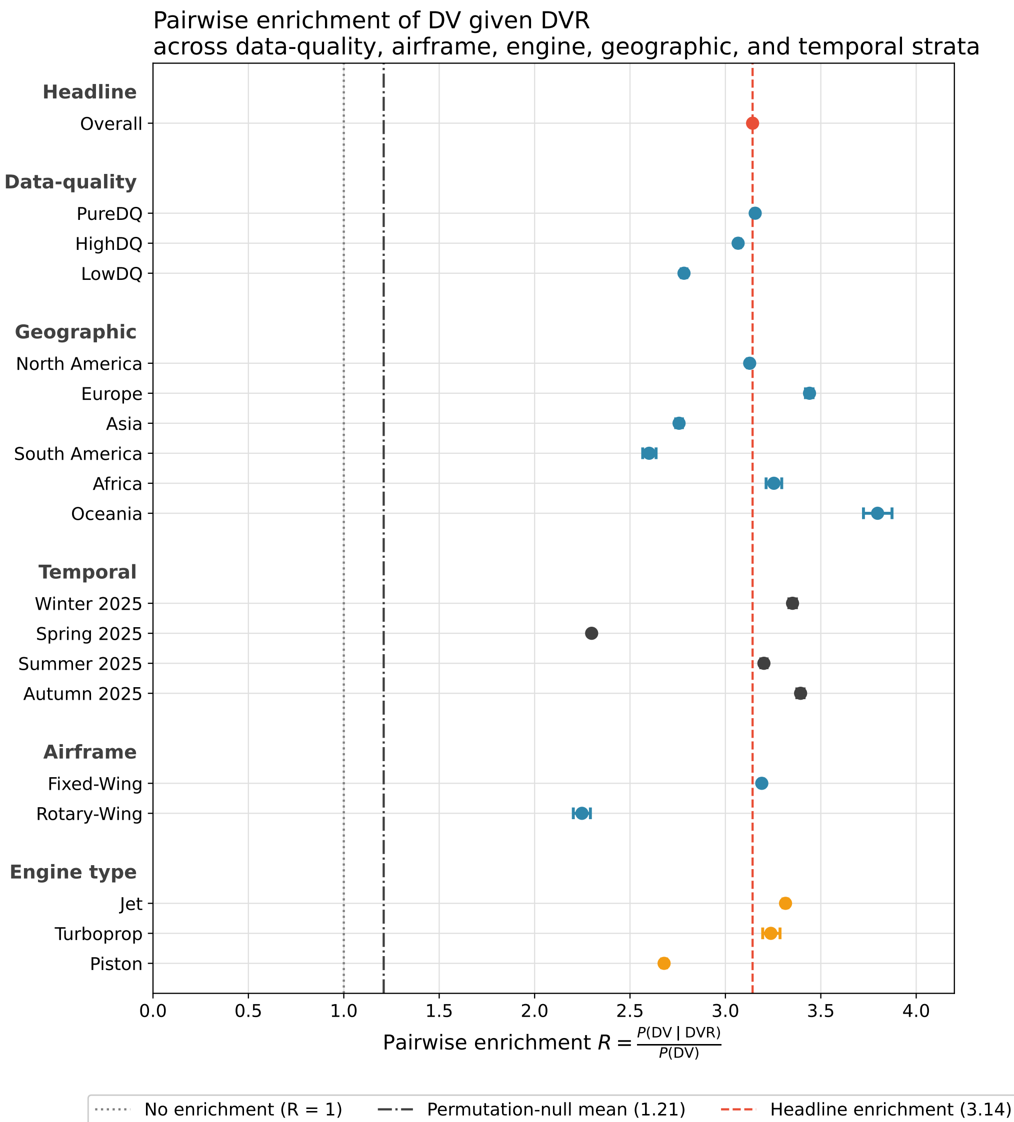

Figure 3: Pairwise enrichment $R(\mathrm{DV} \mid \mathrm{DVR})$ across data-quality tiers, airframe classes, engine types, geographic regions, and temporal windows. Markers show point estimates with 95% cluster-bootstrap confidence intervals (some narrower than the marker). The dotted vertical line marks $R = 1$ (no enrichment), the dash-dotted line marks the permutation-null mean of 1.21, and the dashed line marks the headline value of 3.14.

4.10. Transponder Integrity and Receiver Redundancy Stratification

The most operationally consequential alternative explanation for the observed coupling is that it stems from receiver-side noise converting a coherent broadcast into apparent multi-channel anomalies. To test this directly, we extracted aircraft-hour aggregates grouped by transponder-reported Navigation Accuracy Category for velocity (NACv), Navigation Integrity Category (NIC), and physical receiver redundancy.

Table 5 demonstrates that the coupling unambiguously strengthens under strict data-quality and redundancy restrictions. Among aircraft-hours reporting strictly high velocity accuracy (NACv ≥ 3, equivalent to $< 1\,\mathrm{m/s}$ error; $N = 47.67\mathrm{M}$), the pairwise enrichment is $3.367\times$ (95% CI $[3.353, 3.382]$), compared to $3.003\times$ $[2.991, 3.015]$ in the low/unknown NACv stratum ($N = 69.73\mathrm{M}$). NACv and NIC are transponder-self-reported categories, and a malfunctioning transponder could in principle report high accuracy while broadcasting degraded kinematics. The receiver-redundancy stratum (≥ 10 distinct receivers; $N = 55.27\mathrm{M}$), by contrast, depends on the receiver network rather than on transponder self-report and exhibits the highest single-criterion enrichment in the table ($3.507\times$, $[3.493, 3.521]$). The strongest claim therefore rests on the joint criterion, in which receiver-side and transponder-side controls operate in combination.

Decisively, the high-integrity triple-intersection (aircraft-hours simultaneously satisfying NACv ≥ 3, NIC ≥ 8, and observed by ≥ 10 receivers ($N = 20.79\mathrm{M}$)) yields the highest enrichment measured in the entire study: $3.795\times$ (95% CI $[3.775, 3.816]$), which increases to $3.809\times$ (95% CI $[3.786, 3.831]$) when isolating the perfectly sampled $\Delta t = 1\,\mathrm{s}$ PureDQ subset. Furthermore, while the irreducible three-way conditional dependence (Pathway B violation, V_B) does attenuate somewhat under the strictest controls, from $V_B = 1.543$ in the full sample to $V_B = 1.339$–1.349 in the high-integrity strata, it remains well above the conditional-independence value of unity throughout. The refutation of a receiver-side-noise origin rests on the *pairwise* channel, which such an origin predicts should collapse under high-integrity and high-redundancy restriction: the observed *strengthening* of the enrichment (rather than its attenuation) is inconsistent with that prediction. We do not rest the refutation on V_B, whose modest attenuation is compatible with a small residual receiver-side contribution to the three-way component alongside the dominant broadcaster-side coupling.

Table 5: Stratification by transponder integrity and receiver redundancy (Section 4.10).

Stratum	N (M)	Airframes	P(DV)	P(DV | DVR)	Enrichment [95% CI]	V_B [95% CI]
Low/Unknown NACv	69.73	216,671	0.2284	0.6860	3.003 [2.991, 3.015]	1.597 [1.593, 1.601]
Strict High NACv (NACv ≥ 3)	47.67	61,412	0.2111	0.7107	3.367 [3.353, 3.382]	1.429 [1.425, 1.432]
Strict High NIC (NIC ≥ 8)	102.36	206,149	0.2121	0.6829	3.219 [3.209, 3.229]	1.564 [1.561, 1.567]
Very High Redundancy ($\geq$ 10 rec.)	55.27	152,952	0.2079	0.7290	3.507 [3.493, 3.521]	1.432 [1.429, 1.435]
DO-260B Clean	39.45	51,451	0.2030	0.7018	3.458 [3.440, 3.475]	1.422 [1.418, 1.426]
DO-260B Clean (PureDQ)	34.80	51,322	0.2008	0.6935	3.453 [3.435, 3.470]	1.430 [1.426, 1.434]
High-Integrity Triple-Intersection	20.79	35,823	0.1960	0.7437	3.795 [3.775, 3.816]	1.339 [1.336, 1.342]
High-Integrity Triple-Intersection (PureDQ)	17.09	35,423	0.1936	0.7375	**3.809 [3.786, 3.831]**	**1.349 [1.346, 1.352]**

Note. Sample size N is expressed in millions of aircraft-hours. Brackets indicate 95% cluster-bootstrap confidence intervals computed by airframe across 1,000 iterations. The "DO-260B Clean" subset requires NACv ≥ 3 and NIC ≥ 8. The "High-Integrity Triple-Intersection" stratum enforces DO-260B compliance alongside reception across ≥ 10 distinct stations.

4.11. Density-Confound Regression Robustness

Adding the per-hour message density (the log valid-pair count) to the regression as a direct control leaves the coupling essentially unchanged at the data median: the DVR odds ratio is 8.111 without the control and an indistinguishable 8.01 with it. The small density interaction runs *opposite* to the confound's prediction (the implied coupling weakens slightly at higher density (odds ratio from 9.69 at the sparsest hours to 7.73 at the densest), whereas a disjunctive-aggregation artefact would make it stronger) and a quintile stratification likewise shows no monotonic rise with density. The disjunctive-aggregation density confound is therefore directly refuted; the full coefficients and quintile values are reported in Appendix Appendix A.1 (Tables A.2 and A.3).

4.12. Sub-Second Temporal Coincidence: Pair-Level Enrichment in the NACv Pilot

To confirm the coupling is genuine within-second co-occurrence rather than an artefact of hourly aggregation, we ran a pair-level pilot whose unit of analysis is the individual consecutive-message pair rather than the aggregated aircraft-hour. The enrichment *intensifies* as the window narrows from one hour to one message pair (9.76× pooled, rising to 19.9× in the strict high-velocity-accuracy stratum) the opposite of what an hourly-window artefact would predict, and the Pathway-B violation likewise survives. We therefore read the aircraft-hour-level 3.142× as a conservative bound on a stronger underlying pair-level coupling. The full values, a vertical-rate source-bit cross-check, and the granularity caveat are reported in Appendix Appendix A.2 (Table A.6) and Section 7.3.

4.13. Phase-Classification Symmetric-Threshold Sensitivity

Re-fitting the phase regression with three alternate *symmetric* vertical-rate threshold sets leaves the Climb, Descent, and DH main effects stable to within roughly ±10%, and the Climb×DH and Descent×DH interactions super-multiplicative throughout. Only the Transition×DH interaction is materially threshold-sensitive (unsurprising, since Transition is the residual catch-all category whose membership shifts with the threshold) and the headline finding, that Climb, Descent, and DH act as robust independent stressors with super-multiplicative coupling in the climb and descent sub-strata, is unaffected. The full coefficient table is reported in Appendix Appendix A.1 (Table A.4).

4.14. Altitude Stratification: Cruise versus Terminal

A leading non-telemetry reading of the coupling is that it merely tracks genuine aircraft manoeuvre: aircraft that accelerate, change vertical rate, and turn at the same time will trip DV, DVR, and DH together for ordinary aerodynamic reasons. This reading makes a sharp, testable prediction. Coordinated severe manoeuvre across all three channels is common in the terminal area (climb-out, descent, approach, holding) but operationally implausible in stable cruise, so a pure-manoeuvre origin requires the coupling to be strongest at terminal altitudes and to weaken at cruise.

To test this we re-extracted the Stage 2.1 aircraft-hour aggregate retaining the per-hour median barometric altitude, a field already computed within the Stage 2.1 query but previously dropped from the written aggregate (Section 7.2), under *identical* flag definitions, quality filters, and within-airframe cluster-bootstrap methodology to the headline analysis. Each aircraft-hour was assigned to one of three bands by its median barometric altitude: *terminal* ($< 10{,}000\,\mathrm{ft}$), *transition* (10,000–25,000 ft), and *cruise* ($\geq 25{,}000\,\mathrm{ft}$). The union of the bands reproduces the full Stage 2.1 extraction (117,542,773 aircraft-hours (42 more than the low-latency Dataset 3A, immaterial here) with all-band $R = 3.142$, confirming the re-extraction reproduces the headline pooled result). Table 6 reports the band-stratified pairwise enrichment $R = P(\mathrm{DV} \mid \mathrm{DVR})/P(\mathrm{DV})$, the irreducible three-way Pathway-B violation V_B, and (in text) the coupled-triple prevalence, each with 1,000-iteration cluster-bootstrap 95% confidence intervals.

The result is the opposite of the pure-manoeuvre prediction on the pairwise channel. The enrichment is *highest* in the cruise band, $R_{\text{cruise}} = 4.413$ (95% CI $[4.390, 4.432]$), exceeding both terminal (2.635, $[2.626, 2.645]$) and transition

$(2.462, [2.453, 2.471])$. That a severe groundspeed anomaly conditional on a vertical-rate anomaly is *most* enriched in stable cruise, where a simultaneous longitudinal- and vertical-rate excursion is not an ordinary flight event, is the single strongest discriminator in this study against a genuine-dynamics origin for the coupling. The marginals make the mechanism explicit: severe groundspeed anomalies are rarest at cruise ($P(\mathrm{DV}) = 0.1542$, the lowest of any band) yet, conditional on a vertical-rate anomaly, occur at 4.4× that rate, the rare-but-tightly-clustered pattern expected of shared broadcast-platform co-corruption rather than the broad baseline elevation expected of manoeuvre.

The irreducible three-way coupling tells a complementary, more nuanced story. The Pathway-B violation remains significantly above unity in every band, including cruise ($V_{B,\text{cruise}} = 1.388$, 95% CI $[1.385, 1.391]$), confirming that genuine three-channel structure persists at cruise; but it is largest in the terminal band ($V_{B,\text{terminal}} = 1.653$, $[1.649, 1.658]$). The terminal excess is consistent with an *additional*, genuinely aerodynamic contribution to three-way co-occurrence where coordinated manoeuvre is common. The two observations together support a mixed regime (a shared broadcast-platform coupling that operates at all altitudes including cruise, supplemented by real-manoeuvre co-occurrence concentrated in the terminal area) consistent with, and strengthening, the mixed-mechanism interpretation of Section 5.2. The altitude-dependent antenna-shadowing candidate of that section is directly supported: the coupling does not require approach/departure geometry to appear.

The altitude distribution of the coupled-triple burden corroborates this. Of all coupled $(\mathrm{DV}, \mathrm{DVR}, \mathrm{DH})$ triples across the three resolved bands, 28.2% (95% CI $[27.9, 28.4]$) occur in cruise, against 42.4% terminal and 29.5% transition. Per-hour triple prevalence is in fact lowest at cruise ($P(\text{triple}) = 0.0355$, versus 0.0665 terminal and 0.0739 transition), yet the cruise band still carries more than a quarter of all coupled triples; the coupling is therefore not a terminal-area artefact but is distributed across the cruise regime.

4.15. Seasonal Robustness and the Spring 2025 Elevation

The seasonal-week robustness analysis was pre-specified to test whether the pairwise coupling is invariant across the calendar year; in practice it both confirms seasonal robustness across three of the four checkpoint weeks and surfaces the major secondary finding of this study. The seasonal-week pairwise enrichments are 3.35× (Winter, week of 13 January), 2.30× (Spring, week of 7 April), 3.20× (Summer, week of 7 July), and 3.39× (Autumn, week of

Table 6: Altitude-band stratification of the pairwise enrichment R and the Pathway-B violation V_B, computed on the per-hour median barometric altitude re-extraction under identical flag definitions and cluster-bootstrap methodology to the headline analysis (Section 4.14). Brackets are 1,000-iteration within-airframe cluster-bootstrap 95% confidence intervals. The union of the three resolved bands reproduces the full Stage 2.1 extraction; the pooled all-band R matches the headline value.

Altitude band	N (aircraft-h)	$P(\text{DV})$	$P(\text{DV} \mid \text{DVR})$	R [95% CI]	V_B [95% CI]
All (full extraction)	117,542,773	0.2214	0.6957	3.142 [3.132, 3.152]	1.543 [1.541, 1.546]
Terminal (<10 kft)	40,903,885	0.2564	0.6757	2.635 [2.626, 2.645]	1.653 [1.649, 1.658]
Transition (10–25 kft)	25,598,903	0.2993	0.7369	2.462 [2.453, 2.471]	1.311 [1.309, 1.314]
Cruise (≥25 kft)	50,935,246	0.1542	0.6803	4.413 [4.390, 4.432]	1.388 [1.385, 1.391]

Aircraft-hours with no barometric altitude in the hour ($N = 104{,}739$, 0.09%) are unbanded and excluded from the resolved-band statistics and from the cruise-share denominator. The all-band count (117,542,773) is the full Stage 2.1 extraction and exceeds the low-latency Dataset 3A (117,542,731) used for the headline by 42 aircraft-hours (fraction 3.6×10^{-7}), immaterial to every reported statistic.

6 October). The Spring 2025 checkpoint is a notable outlier with markedly elevated marginal anomaly prevalences: $P(\text{DV}) = 0.357$ ($\approx 1.6\times$ the year-round average of 0.221), $P(\text{DVR}) = 0.208$ ($\approx 1.9\times$), $P(\text{DH}) = 0.390$ ($\approx 1.5\times$). The lower enrichment ratio in Spring 2025 is a direct mathematical consequence of the elevated marginals: when $P(\text{DV})$ approaches the conditional probability ceiling, the $P(\text{DV} \mid \text{DVR})/P(\text{DV})$ ratio mechanically declines.

A 365-day temporal breakdown of the marginal prevalences identifies a single sustained elevation interval running from **2 April through 22 May 2025** (50 calendar days, of which 45 exceed the $P(\text{DV}) > 0.30$ trigger) peaking at $P(\text{DV}) = 0.498$ on 16 May (Figure 4). The seasonal-checkpoint week (7–13 April) captures the early plateau of a much longer episode. Outside this window, daily $P(\text{DV})$ remains near baseline for the remainder of the year (median 0.192; approximately 83% of non-window days fall within 0.18–0.22), with isolated single-day excursions reaching ≈ 0.28 (e.g. 4 July 2025) that do not trigger a sustained elevation.

UTC diurnal structure.. Within the elevated week (7–13 April), the diurnal breakdown reveals a strong daily cycle: $P(\text{DV})$ peaks at 0.480 at UTC 17:00 and troughs at 0.204 at UTC 02:00, a $2.35\times$ swing within a single day. The peak band (UTC 14:00–19:00) corresponds to local afternoon hours over North America and Europe.

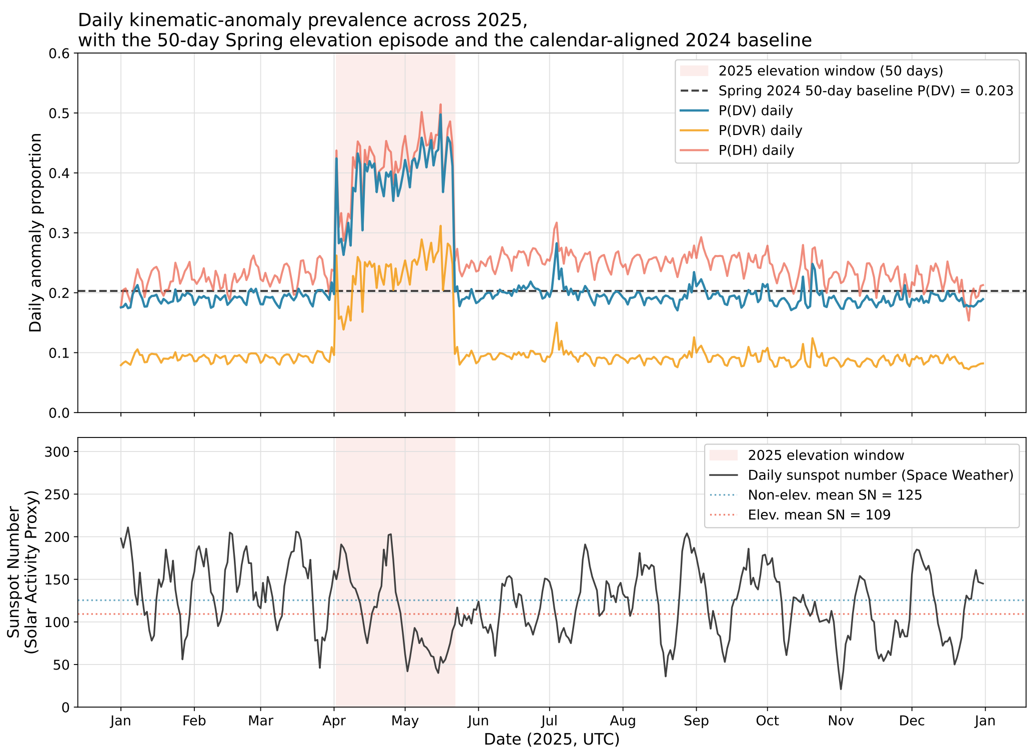

Figure 4: Daily kinematic-anomaly prevalence across calendar 2025 with the 50-day Spring 2025 elevation window (2 April through 22 May, peak $P(\mathrm{DV}) = 0.498$ on 16 May) shaded in red. *Top panel*: daily $P(\mathrm{DV})$, $P(\mathrm{DVR})$, and $P(\mathrm{DH})$; the horizontal dashed line marks the calendar-aligned 50-day Spring 2024 baseline $P(\mathrm{DV}) = 0.203$ (Table 7). *Bottom panel*: daily international sunspot number from the GFZ Potsdam archive [51, 52]; see Section 5.5 for interpretation.

Geographic localization.. The elevation is most pronounced in Europe (P(DV) elevated/control ratio 1.38×) and North America (1.27×), with smaller elevations in Africa, Asia, Oceania, and South America (1.04–1.14×). P(DVR) shows an even larger European elevation ratio (1.53×). The bias toward afternoon-peak-Atlantic longitudes coincides with the operational peak of transatlantic and intra-European jet activity and with both convective and solar-ionospheric afternoon windows over those longitudes; both meteorological and space-weather candidates are evaluated and rejected in the cross-references that follow.

Airframe and engine localization.. The elevation hits jet aircraft hardest (P(DV) elev/ctrl ratio 1.32×), with turboprop (1.17×) and piston (1.14×) showing more modest elevations. Rotary-wing aircraft are essentially unaffected (ratio 1.06×). Within airframe categories, Fixed-Wing shows a 1.28× elevation; Rotary-Wing shows 1.06×.

NOAA Storm Events cross-reference.. We downloaded the NOAA Storm Events Database for 2025 and tabulated daily counts of severe convective events (Tornado, Hail, Thunderstorm Wind, High Wind, Heavy Rain, Hurricane/Typhoon, Tropical Storm, Flash Flood, Flood, Strong Wind) across the elevated week and the seven-day control window immediately prior. The elevated week (7–13 April) recorded 821 severe convective events; the immediately preceding control week (31 March–6 April) recorded 2,605. The elevated week therefore had only 0.32× the convective activity of the control week, severe convective activity was substantially *lower* during the elevated week, not higher. NOAA Storm Events is a US-only product and so the contrast is restricted to North America, but the direction of the contrast already rules out a US convective driver for the North American component of the elevation, and is independently incompatible with a meteorological driver for the European component. We discuss candidate non-convective drivers (e.g., ionospheric/solar) in Section 5.5.

Calendar-aligned Spring 2024 comparison.. To distinguish a recurring seasonal phenomenon from a 2025-specific event, we re-extracted the full calendar-aligned 50-day window of Spring 2024 (2024-04-02 through 2024-05-22) directly from `state_vectors_data4` using a SQL pipeline reproducing the rate-normalized DV/DVR/DH flag definitions of Section 3.2 (Methods, Section 3.11). The 2024 sample contains $N = 15{,}733{,}909$ aircraft-hours across 192,747 unique airframes. Table 7 reports the comparison.

Table 7: Year-over-year comparison of the calendar-aligned 50-day Spring window (2 April through 22 May) (Section 4.15).

Subset	N (M)	Airframes	P(DV)	P(DVR)	P(DH)	Enrichment	V_B
Spring 2024 50-day (Apr 2–May 22)	15.73	192,747	0.203	0.103	0.244	3.380 [3.365, 3.395]	1.544 [1.540, 1.548]
2025 PureDQ year baseline	103.80	239,087	0.218	0.108	0.264	3.156 [3.146, 3.166]	1.557 [1.554, 1.560]
2025 control 50-day (Feb 11–Apr 2)	15.09	163,351	0.193	0.092	0.229	3.249 [3.236, 3.263]	1.523 [1.519, 1.528]
Spring 2025 elevated 50-day (Apr 2–May 22)	16.26	171,059	0.394	0.235	0.421	2.153 [2.147, 2.159]	1.319 [1.317, 1.322]

Note. Spring 2024 was re-extracted directly from `state_vectors_data4` for the present analysis; the 2025 rows are read from Dataset 3A. "N (M)" is in millions of aircraft-hours. The 2025 control window is the 50 calendar days immediately preceding the elevation onset. Cluster-bootstrap 95% CIs over 1,000 iterations resampling by airframe. The primary baseline is the 2025 PureDQ year row; the 2025 control row is reported for completeness. The Spring 2024 row reported here is the 2026-05-22 extraction; the monitor of Section 6 (Table 8) re-extracts all of its windows in a single later (2026-06-18) vintage and so reports a lower-level Spring 2024 row (Section 7.8). The two Spring 2024 entries reflect different archive vintages and are not to be compared cell-for-cell.

The Spring 2024 50-day sample exhibits marginal anomaly prevalences *slightly below* the 2025 year-round PureDQ baseline ($P(\mathrm{DV}) = 0.203$ versus 0.218; $P(\mathrm{DVR}) = 0.103$ versus 0.108; $P(\mathrm{DH}) = 0.244$ versus 0.264), nowhere near the Spring 2025 elevation regime ($P(\mathrm{DV}) = 0.394$, $P(\mathrm{DVR}) = 0.235$, $P(\mathrm{DH}) = 0.421$). Spring 2024 conditional enrichment (3.380×, 95% CI $[3.365, 3.395]$) is modestly higher than the 2025 year baseline (3.156×, CI $[3.146, 3.166]$); V_B (1.544, CI $[1.540, 1.548]$) is modestly lower (1.557, CI $[1.554, 1.560]$). Both differences are statistically resolvable at the 10^7-aircraft-hour sample size but are small ($\approx 7\%$ on enrichment and $< 1\%$ on V_B) and operationally negligible compared with the Spring 2025 elevation. The calendar-aligned 2024 window is therefore a normal baseline interval, not an elevated interval. This 2024 baseline is the one comparison in the paper drawn from a mutable upstream archive: re-extracting the same calendar window on a later date returns a different aircraft-hour and airframe yield (Section 3.11), so the contrast is reproducible in direction (2024 at baseline, far from the 2025 episode) but not in exact magnitude.

Year-over-year on the same calendar 50-day window, the marginal anomaly prevalences climbed in 2025 by factors of 1.94× ($P(\mathrm{DV})$, from 0.203 to 0.394), 2.27× ($P(\mathrm{DVR})$, from 0.103 to 0.235), and 1.73× ($P(\mathrm{DH})$, from 0.244 to 0.421). The pairwise enrichment ratio drops from 3.380× in 2024 to 2.153× in 2025 (a direct mathematical consequence of the elevated marginal $P(\mathrm{DV})$, as detailed at the head of this subsection) and V_B drops from 1.544 to 1.319 for the same arithmetic reason. The underlying coupling structure is preserved across the year-over-year contrast: what changed in 2025 was the marginal failure rate on all three channels, not the conditional dependence linking them. This is the signature of an exogenous driver pushing all three channels up by similar factors without altering the mechanism that couples them, and rules out a recurring seasonal phenomenon (jet stream onset, spring convective season, springtime ANSP equipment cycling) as a candidate driver of the 2025 elevation. We acknowledge that the OpenSky receiver network grows monotonically through volunteer additions and the 2025 calendar window has higher receiver-network coverage than the 2024 calendar window; however, this differential affects per-hour valid-pair counts and marginal sample size, not the within-pair rate-normalized conditional dependence that is the headline finding (Section 4.2). The conditional enrichment $R(\mathrm{DV} \mid \mathrm{DVR})$ is, by construction, a within-airframe-within-hour ratio invariant to inter-year shifts in network density acting multiplicatively on N_{pair}; the modest year-over-year shift in enrichment from 3.380× to 2.153× (with the corresponding shifts in V_B) is dominated by the elevated marginal $P(\mathrm{DV})$ in 2025, not by receiver-network growth.

Space-weather cross-reference.. The most empirically defensible recurring driver with a cycle longer than annual is the maximum of Solar Cycle 25, which spans 2024–2025. Elevated solar activity has been documented to degrade global GNSS performance and would naturally affect cruising commercial aircraft (the most GNSS-dependent operational class) preferentially. We tested this hypothesis directly using the GFZ Potsdam daily geomagnetic and solar indices archive [51, 52]. The summary is given in Table A.11.

The elevation window has modestly higher mean K_p (2.86 vs. 2.48, +15%), but the year-round daily correlation between K_p and $P(\mathrm{DV})$ is weak (Pearson $r = 0.09$, 95% CI bracketing zero) and the within-elevation correlation is mildly *negative* ($r = -0.27$, $p = 0.06$, marginally non-significant at the conventional 0.05 level). Solar activity is *lower* during the elevation window than during the rest of 2025 on both S_N (109 vs. 125, −13%; Welch $p = 0.025$) and $F_{10.7}$ (150.1 vs. 155.6, −4%; $p = 0.093$). The within-elevation Pearson correlations are substantially negative ($r = -0.63$ for S_N, $r = -0.60$ for $F_{10.7}$, both $p < 0.001$): as the elevation episode deepens toward its peak on 16 May, solar activity is on a falling trend. None of the ten 2025 days with the highest $P(\mathrm{DV})$ fall above the year's 95th percentile in either $K_p^{\max}$ or $F_{10.7}^{\mathrm{obs}}$. Geomagnetic-storm days ($K_p^{\max} \geq 5$, NOAA G1+; $n = 62$ in 2025) exhibit a mean $P(\mathrm{DV})$ of 0.219, essentially identical to the 0.221 on non-storm days.

The data therefore do not support a space-weather driver for the Spring 2025 elevation. The modest K_p uplift in the elevation window is small in magnitude and is not supported by a within-window day-to-day association. The solar indices move in the opposite direction to what an ionospheric/F-region-perturbation hypothesis would predict. We interpret the implication in Section 5.5.

5. Discussion

Section 4 established three empirical results: a structural conditional dependence among severe per-step anomalies in groundspeed, vertical rate, and track angle that survives an extensive battery of pre-specified controls; a pattern of pathway-specific violation ratios in which the vertical-rate channel mediates most of the cross-channel coupling; and a sustained 50-day Spring 2025 elevation episode that preserves the coupling architecture at elevated marginal prevalences. We organize the discussion around the substantive interpretation of these results. Section 5.1 addresses whether the coupling can be reduced to genuine extreme physical events or whether it implicates the telemetry encoding chain. Sections 5.2 and 5.3 interpret the conditional-independence-violation pattern and the rotary-wing sub-multiplicative interaction structure respectively. Section 5.4 draws operational implications for ANSP cooperative-surveillance monitoring. Section 5.5 examines the candidate-driver question for the Spring 2025 episode in light of the cross-references reported in Section 4.15. Section 5.6 closes with a comment on effect-size magnitudes versus statistical significance at the present sample size.

Before interpreting the mechanism we consolidate the alternative explanations for the coupling and the pre-specified analysis that excludes each (Table A.12); every entry makes a directional prediction that the data reverse.

5.1. Physical Events versus Telemetry Artifacts

The indicators DV, DVR, and DH defined in Equations 1–3 are operationally defined as rate-normalized kinematic extremes; they do not, in isolation, distinguish (a) genuine extreme physical events (severe turbulence, microburst encounter, intentional manoeuvre, or upset recovery) from (b) telemetry-side artifacts arising on the broadcasting platform or in the encoding chain. The present analysis cannot adjudicate (a) versus (b) at the level of an individual aircraft-hour without ground-truth on-aircraft instrumentation that the public OpenSky archive does not retain. The substantive claim of this paper is therefore about the *coupling* between the three indicators, not the proportion of indicator firings attributable to each cause: the operational significance of multi-channel kinematic-consistency monitoring is the same whether a given coupled triple-occurrence reflects a real upset or a transponder artifact.

Five features of the data nevertheless argue that the coupling is not reducible to physical events alone. First, DVR is anchored to the barometric vertical-rate channel and to the DO-260B baro-altitude encoding (Equation 2, Section 3.2.2); DV and DH are anchored to the GNSS-derived velocity and track-angle channels. Co-occurrence across these three independently encoded sensor pathways implicates a shared cause that is not specific to any one channel. Second, the DVR threshold sweep (Section 4.6) is monotonic in the direction opposite to a turbulence-frequency-driven prediction: the strongest coupling appears at the strictest threshold. Third, the receiver-redundancy intersection (≥ 10 distinct receivers) preferentially samples cruise-altitude operations over dense terrestrial networks, where severe turbulence is rare, yet exhibits the highest single-criterion enrichment (Section 4.10). Fourth, the within-cluster permutation null (Section 3.9.5) preserves each airframe's marginal count of DV, DVR, and DH hours and asks whether they coincide in the same hour more often than random within-cluster reassignment would produce; the observed enrichment exceeds this sample-size-matched null distribution at every iteration. Fifth, the Spring 2025 elevation (Section 4.15) is a sustained 50-day, two-continent regime in all three channels simultaneously, including the barometric channel, which no candidate physical-events explanation has been able to reproduce. These observations are individually consistent with a contribution from telemetry-side coupling and jointly disfavour a pure-physics reading. The most economical description remains a mixed regime in which both physical and broadcast-platform mechanisms contribute, with the empirically dominant contribution to the residual three-way structure (Pathway B, V_B) coming from the latter.

5.2. Mechanistic Interpretation: DVR as the Principal Mediator

The combined evidence is most economically explained by treating the vertical-rate anomaly DVR as the principal but not sole mediator of the DV–DH correlation. That evidence comprises three strands:

(a) a large $V_B = 1.543$ that intensifies under data-quality controls;

(b) a $V_I = 1.149$ well below both the $\approx 2.34\times$ marginal DV–DH enrichment and the ≈ 2.10 permutation-null mean of that statistic; and

(c) a coupling that survives six engine-type and geographic strata, four seasonal weeks, three data-quality tiers, four symmetric-threshold variants, and a strict-Fixed-Wing specification, strengthens from a full-sample headline of $3.14\times$ to $3.81\times$ under the perfectly-sampled high-integrity triple-intersection of high-transponder-integrity and high-receiver-redundancy criteria (Section 4.10), and falls entirely outside the random-pairwise-structure permutation null (one-sided $p < 10^{-3}$; standardized gap $> 1{,}200$ null SD).

We use "principal mediator" rather than "common antecedent" deliberately: a strict common-antecedent directed acyclic graph would predict $V_I = 1.000$ exactly (d-separation of DV and DH given DVR), whereas the observed $V_I = 1.149$ formally rejects perfect d-separation. The $\approx 15\%$ residual above unity quantifies the direct DV–DH structure not absorbed by DVR. Equivalently, the Model 3 logit-scale main effect of DH on DV in the DVR $= 1$ subpopulation (Table 4, OR $= 3.91$) corresponds, at the high baseline prevalence $P(\text{DV} \mid \text{DVR} = 1) = 0.696$ characteristic of that stratum, to a probability-scale enrichment of approximately 1.2–1.3, consistent with $V_I = 1.149$. The two statistics live on different scales and report the same residual structure.

Under this interpretation: a physical event aboard the broadcasting platform that perturbs the vertical-rate broadcast channel concurrently perturbs the ground-velocity and heading channels with positive probability. The ground-velocity and heading channels share an upstream derivation in DO-260B-compliant transponders (track angle h is computed from horizontal velocity components (v_e, v_n) via $h = \arctan 2(v_e, v_n)$, and ground velocity magnitude $v = \sqrt{v_e^2 + v_n^2}$ shares the same components) so a portion of the marginal DV–DH enrichment of $\approx 2.34\times$ is attributable to this shared derivation rather than to independent kinematic coupling. The $\approx 50\%$ attenuation of this marginal enrichment to $V_I = 1.149$ under conditioning on DVR is consistent with the shared-derivation component being partially absorbed; the residual $V_I = 1.149$ then represents the conditional dependence beyond both derivation-sharing and DVR-mediation. The conditional dependence between DV and DVR given DH, by contrast, is large ($V_B = 1.543$) and not subject to the shared-derivation critique because DVR is computed from the vertical-rate channel rather than from the horizontal velocity components (Section 7.5 discusses the channel source).

A physically plausible candidate for the shared antecedent is *antenna shad-*

owing during dynamic manoeuvre. Vertical-rate transients reflect pitch transients, which alter the antenna pattern relative to the GNSS satellite constellation and to the ground-based reception network simultaneously. The mechanism is dimensionally concrete: at jet cruise speeds a vertical-acceleration excursion at the DVR threshold (an 8 m/s vertical-rate change within a second, Section 3.2.2) corresponds to a flight-path-angle transient of order a few degrees per second, which swings the body-fixed transponder and GNSS antenna boresights by a comparable amount within the same second, perturbing the downlink geometry to the ground receivers and the uplink geometry to the satellite constellation concurrently, rather than through two independent coincidences. This candidate mechanism is altitude-dependent (cruise-altitude antenna geometry differs substantially from approach/departure geometry), and the altitude stratification of Section 4.14 provides the direct empirical test: the pairwise coupling is in fact strongest in the cruise band, where approach/departure antenna geometry does not apply, while the irreducible three-way component persists at cruise and is largest in the terminal area. The transponder's downlink path is potentially affected at the same instant as its uplink GNSS reception, producing concurrent perturbations in the three broadcast kinematic channels via a candidate shared physical pathway. This candidate mechanism is moreover *consistent with* the observed intensification of coupling under strict NACv, NIC, and receiver-redundancy constraints: cleaner transponders are quieter under nominal conditions, so their residual co-occurrence is more concentrated during genuine manoeuvre events. This is a post-hoc consistency argument rather than an a-priori prediction, since the same high-integrity intensification is what excludes the receiver-side-noise alternative of Section 4.10.

The data do not permit a causal identification of this specific mechanism versus alternative dynamical drivers (turbulence-induced control-surface deflections that simultaneously alter pitch, yaw, and bank; engine surge transients that alter both thrust and attitude). We characterize the *structure* of the coupling and constrain the alternative explanations that the structure can rule out, without claiming a unique mechanism. The distinguishing feature of the pattern is that DVR acts as the central node, not that any particular physical antecedent has been identified.

5.3. Rotary-Wing Sub-Multiplicative Interactions

The rotary-wing counterfactual (Table 4, Model 3B) is at first sight puzzling: the phase $\times$ DH interactions are heavily sub-multiplicative (0.21–0.28), the opposite of the fixed-wing super-multiplicative pattern. The rotary-wing saturation grid (Table A.10) rules out the natural first explanation that this reflects a logistic-link ceiling, because predicted $P(\mathrm{DV})$ in the highest-risk cell (Climb/DH=1) is only 0.72, comfortably below saturation.

The substantive interpretation rests on the physical distinction between fixed-wing and rotary-wing flight. In a fixed-wing aircraft, pitch attitude (driving vertical rate) and yaw/bank (driving heading rate) come from separate control surfaces, elevator versus aileron/rudder, and correspond to separate physical phenomena. A vertical-rate transient and a heading-rate transient are therefore approximately independent stressors on the avionics platform, and their joint occurrence multiplies the DV probability super-multiplicatively (Climb $\times$ DH OR ≈ 1.32 in fixed-wing).

In a rotary-wing aircraft, by contrast, vertical-rate and heading-rate transients are both modulated by a single physical driver: rotor-disc flapping and cyclic-pitch input. A helicopter executing a heading change (DH $= 1$) is simultaneously changing its vertical-rate profile via the same control input, with antenna pattern and platform vibration both being modulated by the same shared mechanical source. The two stressors are therefore not independent; their joint effect is closer to the larger of the two than to their product. This produces the observed sub-multiplicative interaction: DV probability in Climb/DH=1 (0.72) is only modestly higher than in Cruise/DH=1 (0.52), because Cruise/DH=1 already contains the rotor-modulation effect that Climb/DH=1 would otherwise add.

We refer to this as the "shared-rotor-modulation" interpretation. It predicts that the sub-multiplicativity should reverse in fixed-wing-coupled fly-by-wire platforms where pitch and yaw inputs are co-mixed by the flight control computer; we do not have data to test this prediction directly.

5.4. Implications for ANSP Infrastructure Monitoring

For air navigation service providers monitoring cooperative-surveillance infrastructure under ICAO Annex 19 Safety Management System requirements, the principal substantive implication is that single-channel anomaly counts (e.g., velocity-jump counts only, or heading-jump counts only) under-represent the operational significance of cooperative-surveillance degradation. A trajectory that exhibits a velocity-channel anomaly in isolation may be safely discarded by a conflict-prediction algorithm and re-acquired on the next message. A trajectory in which DV, DVR, and DH co-occur in the same aircraft-hour is statistically inconsistent with independent single-channel noise and warrants different downstream treatment. The unit of monitoring at the strategic infrastructure-oversight layer should therefore be the joint-anomaly indicator, not the per-channel counts.

The conditional-dependence framework provides a directly operationalizable output: the per-airframe within-cluster V_B ratio is a strategic performance indicator that distinguishes airframes (or fleet sub-populations, or geographic sectors) that exhibit structurally coupled multi-channel failure modes from those that emit isolated single-channel flags. The threshold and pre-aggregation conventions used here are released alongside this paper [45] for direct replication.

Three properties make this indicator deployable rather than merely diagnostic. *First, it is computationally trivial.* Each monitoring window is summarized by the eight integer counters of Equation 9, which update in constant time per aircraft-hour and retain no raw state vectors. A continental fleet can therefore be charted from a streaming counter table rather than a re-scan of the archive, and the cluster-bootstrap interval is computed on those counters alone. *Second, it is orthogonal to the per-channel quality categories already carried in the message set.* NACp, NACv, NIC, and SIL certify each broadcast field in isolation under the DO-260B and EUROCAE ED-102B minimum operational performance standards [19, 53], whereas the coupling index measures their *joint* consistency, a quantity none of the per-channel categories expresses. It therefore augments, rather than duplicates, the surveillance-performance-monitoring function described in the ICAO Aeronautical Surveillance Manual [54], entering as a new derived indicator over data the ANSP already ingests at the strategic Safety-Management-System oversight layer of ICAO Annex 19 rather than at the tactical-filtering layer. *Third, it is a ground-side post-operational statistic.* Because it modifies no airborne function, it carries none of the development-assurance burden (DO-178C software level, ARP4754A development assurance level) that a change to a certified airborne surveillance or collision-avoidance item would incur [55, 56]; it strengthens infrastructure oversight without touching any certified avionics.

More broadly, the conditional-dependence framework developed here is not specific to ADS-B. Any cooperative-surveillance system that broadcasts multiple independently encoded kinematic channels faces the same multi-channel failure-coupling question. Maritime Automatic Identification System (AIS) transponders broadcast speed-over-ground, course-over-ground, and rate-of-turn on independent encoding paths, and recent work has applied trajectory-level anomaly classification to AIS streams [57]; the European Train Control System (ETCS) broadcasts position, speed, and gradient via the balise–onboard link, with the train-control on-board subsystem itself a recent target of beta-factor common-cause-failure analysis [21]. In each case, the per-unit enrichment ratio and the within-cluster conditional-independence violation ratios V_A, V_I, V_B are directly portable, with the cluster unit redefined from airframe to vessel or train set, the rate-normalized threshold values rescaled to the domain's kinematic envelope, and the permutation null constructed by the same within-cluster hypergeometric scheme (Section 3.9.5). The present study demonstrates the methodology at planetary scale on the largest such system; extension to maritime and railway cooperative surveillance is proposed as future work.

5.5. The Spring 2025 Operational Driver Question

The 50-day Spring 2025 baseline-elevation episode (2 April through 22 May 2025) is the largest single deviation from baseline in the full year of data. Several pieces of evidence constrain candidate operational drivers, without uniquely identifying one.

The geographic and airframe localization patterns (Europe and North America disproportionately elevated, Jet aircraft most affected, Rotary-wing essentially unaffected) are consistent with a driver that operates at cruise altitudes, corroborated directly by the cruise-band enrichment ($R_{\text{cruise}} = 4.413$) of Section 4.14, and is correlated with afternoon hours in the Atlantic longitudes. The strong UTC 14:00–19:00 diurnal peak supports this.

NOAA Storm Events records show that the elevated week recorded only $0.32\times$ the count of US severe convective events of the immediately preceding control week (821 versus 2,605). The convective-driver hypothesis is therefore not supported by the data on either side of the Atlantic: severe convective activity was *lower*, not higher, during the elevation episode, and the European elevation magnitude is independently unattributable to a US-only meteorological cause.

GNSS-interference cross-reference.. The 2024–2025 timeframe coincides with a documented intensification of GNSS jamming and spoofing [58, 59] in the Eastern Mediterranean, Baltic, and Arctic regions, with secondary effects on European and North Atlantic air traffic that the first author has previously characterized at the regional level [15, 16, 17]. GNSS interference would naturally manifest as DV anomalies (via GNSS velocity-solution discontinuities) and could plausibly elevate DH (via track-angle recomputation from a perturbed velocity vector). The DVR channel is dominated by barometric-source broadcasts under DO-260B defaults, but as discussed in Section 7.5 the OpenSky `vertrate` column does not preserve the DF17 source bit, so a minority of DVR readings may be GNSS-derived and we cannot rule out a GNSS contribution to the elevation on data-architecture grounds alone. The Spring 2025 elevation nevertheless lifts $P(\text{DVR})$ from the year-baseline 0.108 to 0.235 within the 50-day window, a 2.17× rise that is implausibly large to attribute entirely to the GNSS-sourced minority fraction. The geographic pattern likewise spans both Europe (elev/control $P(\text{DV})$ ratio 1.38×) and North America (1.27×); North America is not a documented GNSS-interference hotspot at the European scale during 2025. We conclude that GNSS interference may contribute to the European component of the elevation alongside the broader driver, but cannot be the sole or even principal mechanism. Disentangling the GNSS-interference component from the broader 2025-specific driver requires intersecting the daily $P(\text{DV})$ trajectory with EASA Safety Information Bulletins, EUROCONTROL Voluntary Aviation Safety Reports, and FAA NOTAM-published interference advisories at the date-by-region resolution, which the present aircraft-hour aggregate cannot directly support; we propose this as immediate future work in Section 7.10.

The calendar-aligned Spring 2024 comparison (Table 7) sharpens the candidate-driver analysis. The same calendar 50-day window in 2024 exhibits baseline marginal prevalences (slightly below the 2025 year-round PureDQ baseline) and baseline coupling structure ($V_B = 1.544$ versus the year baseline 1.557), both far from the Spring 2025 elevation regime ($V_B = 1.319$ at the same calendar window). The 2025 elevation is therefore not a recurring seasonal phenomenon. Several otherwise plausible candidate drivers are accordingly ruled out: spring jet-stream onset, the onset of the US convective severe-weather season, periodic ANSP equipment maintenance cycles, and seasonal patterns in receiver-network coverage would all predict that the same calendar 50-day window in 2024 should look like 2025. None of them do.

The space-weather candidate, elevated geomagnetic activity or elevated solar activity at the maximum of Solar Cycle 25, was the most empirically defensible of the remaining longer-than-annual cycles, on the basis that 2024–2025 coincides with the cycle maximum and that ionospheric perturbations during high solar activity have been documented to degrade GNSS-derived velocity solutions in cooperative surveillance. The direct cross-reference of

Section 4.15 (Table A.11) does not support this hypothesis. The elevation window does coincide with a small uplift in mean daily K_p (+15% relative to the non-elevation portion of 2025, Welch $p = 0.015$), but the within-window day-to-day association between K_p and $P(\mathrm{DV})$ is mildly negative and not statistically distinguishable from zero ($r = -0.27$, $p = 0.06$). The solar indices S_N and $F_{10.7}$ are *lower* during the elevation window than during the rest of 2025; their within-window daily correlations with $P(\mathrm{DV})$ are substantially negative ($r = -0.63$ for S_N, $r = -0.60$ for $F_{10.7}$, both $p < 0.001$), which is the opposite of the direction predicted by an ionospheric-driver story. Geomagnetic-storm days at NOAA G1 or worse have a mean $P(\mathrm{DV})$ indistinguishable from quiet days, and none of the ten worst $P(\mathrm{DV})$ days fall in the top 5% of the 2025 geomagnetic or solar distributions. We do not interpret the negative within-window correlation as evidence of a reverse causal mechanism; the parsimonious reading is that the 2025 elevation episode and the temporary lull in solar activity during April–May 2025 are coincidentally aligned in time and uncorrelated in cause.

The candidate-driver list that remains compatible with the joint evidence (a sustained 50-day episode, geographically concentrated in Europe and North America, hitting jet operations disproportionately, quiet on rotary-wing aircraft, with no recurrence in the calendar-aligned 2024 week, with partial but insufficient correlation against US convective weather, and with no consistent association against geomagnetic or solar activity) is dominated by *2025-specific operational events whose nature is not directly observable in the underlying state-vector archive.* Candidate examples include: coordinated avionics firmware or transponder-software update waves deployed by major aircraft OEMs during the 2025 calendar year; constellation-level GNSS events such as the introduction or retirement of specific GPS or Galileo space-vehicle blocks during this period; ANSP-side regulatory transitions that altered the published surveillance performance requirements; or a combination of the above acting on a common subset of the fleet. None of these candidates is directly testable from the data products we analyze here; identifying the driver will require integration with ICAO and EUROCONTROL bulletin archives, GNSS constellation status data, and fleet equipage tracking.

We acknowledge that the structural localization analysis nevertheless rules out an internal data-product anomaly as the explanation: the elevation does not coincide with any documented OpenSky pipeline change or backfill event, and is restricted to the specific dimensions of the kinematic channels rather than affecting all message types.

5.6. Statistical Versus Practical Significance

At $N = 117.5$ million aircraft-hours, virtually any non-zero deviation from a null hypothesis will be detected at conventional significance levels. We have therefore deliberately avoided basing inferential conclusions on p-values and have instead emphasized effect-size point estimates and their cluster-bootstrap confidence intervals. The headline enrichment of $3.14\times$ is approximately a tripling of the marginal anomaly probability, the Pathway B violation $V_B = 1.54$ is approximately a 54% deviation from the conditional-independence prediction, and the coupling strengthens by a further 21% (from $3.14\times$ to $3.81\times$) under the strictest high-integrity transponder and high-receiver-redundancy intersection rather than attenuating. These are operationally substantive magnitudes, not statistical-significance artefacts of large sample size.

To make the magnitudes concrete in monitoring terms: the coupled-triple burden $\widehat{P}_\triangle$ sits at a nominal ≈ 0.045 across the in-control reference windows (Section 6.3), so a mid-sized monitoring unit of 10,000 aircraft-hours expects on the order of 450 all-three-channel hours per window under nominal conditions; during the Spring 2025 episode the same unit would log $\approx$ 1,630 such hours ($\widehat{P}_\triangle = 0.163$), a $3.6\times$ rise in coupled-triple burden. The burden alert keys on exactly this quantity, while the V_B companion confirms whether the surplus retains the shared-pathway signature (Section 6). An ANSP can thus translate the enrichment into an expected daily volume of coupled-anomaly hours per fleet and set the upper control limit $U_\triangle$ to a workload it can action.

6. A Deployable Multi-Channel Consistency Monitor

The empirical results of this paper (that severe anomalies in the three broadcast kinematic channels co-occur far beyond chance, with an irreducible three-way component that survives every integrity and redundancy control (Sections 3.9.5–4.10)) have a direct operational consequence. Multi-channel *inconsistency* is a measurable, calibratable integrity signal orthogonal to the per-channel accuracy categories (NACp, NACv, NIC, SIL) already carried in the ADS-B message set. This section turns the finding into a monitor: the sufficient statistics it consumes, an alerting rule calibrated against the monitor's own in-control operating distribution (with the permutation null of Section 3.9.5 fixing the structural floor), a worked application to the Spring 2025 regime, and a characterization of its detection behaviour.

6.1. Monitor statistics

A *monitoring unit* W is any set of aircraft-hours sharing an operational context (an operator or fleet, an airspace sector, a receiver region, or a rolling time window) together with the airframe (icao24) labels of its constituent hours. From W the monitor needs only the eight integer sufficient statistics already used by the headline cluster bootstrap,

$$\left\{ n,\ n_{\mathrm{DV}},\ n_{\mathrm{DVR}},\ n_{\mathrm{DH}},\ n_{\mathrm{DV,DVR}},\ n_{\mathrm{DV,DH}},\ n_{\mathrm{DVR,DH}},\ n_{\mathrm{DV,DVR,DH}} \right\}, \tag{9}$$

where n is the aircraft-hour count in W and the remainder are the marginal and joint flag counts. The monitor reads two complementary indices from these counters. The first is the *anomaly burden*, summarized by the coupled-triple prevalence

$$\widehat{P}_{\triangle}(W) \;=\; \frac{n_{\mathrm{DV,DVR,DH}}}{n}, \tag{10}$$

the rate at which all three channels flag in the same aircraft-hour (the single-channel marginals n_{DV}/n, n_{DVR}/n, n_{DH}/n are its per-channel companions). The second is the *coupling structure*, summarized by the Pathway-B violation ratio

$$\widehat{V}_B(W) \;=\; \frac{n_{\mathrm{DV,DVR,DH}}\; n_{\mathrm{DH}}}{n_{\mathrm{DV,DH}}\; n_{\mathrm{DVR,DH}}}, \tag{11}$$

which isolates the irreducible three-way dependence the per-channel categories cannot see. Each index is reported with a within-airframe cluster-bootstrap 95% interval obtained by resampling the airframes in W (Section 3.9.5); the pairwise enrichment $\widehat{R}(W) = \big(n_{\mathrm{DV,DVR}}\, n\big)/\big(n_{\mathrm{DVR}}\, n_{\mathrm{DV}}\big)$ is an interpretable companion to $\widehat{V}_B$. The two indices answer different questions. $\widehat{P}_{\triangle}$ asks *how often* the channels fail together and is the event detector; $\widehat{V}_B$ asks whether the joint failures retain the shared-pathway coupling signature and is the regime classifier. Critically, $\widehat{V}_B$ is *not* itself an event detector: its in-control level lies far above the independence-null value of unity (empirically ≈ 1.5; see below), so the monitor cannot treat "$\widehat{V}_B > 1$" as the alarm condition, structural coupling is the operating baseline, not the exception.

6.2. Alerting rule and its calibration

Two reference levels govern the monitor, and they are distinct. The *structural floor* is the independence null $\mathbb{E}[\widehat{V}_B] = 1$, whose sampling distribution at the sample size of W is given exactly by the permutation construction of Section 3.9.5; a unit whose $\widehat{V}_B$ is statistically indistinguishable from 1 has lost the multi-channel coupling structure entirely. The *in-control band* is the operating distribution of each index over nominal units, estimated from a reference set of windows known to be anomaly-free, and it is this band, not the structural floor, against which operational limits are set. The monitor raises a *burden alert* for W when the bootstrap lower bound of the burden index clears the in-control upper control limit $U_\triangle$,

$$\widehat{P}_\triangle^-(W) \; > \; U_\triangle, \tag{12}$$

with $U_\triangle$ calibrated to a target per-window false-positive rate α as the upper $1-\alpha$ envelope of $\widehat{P}_\triangle$ over the in-control reference set (equivalently a mean-plus-$k\sigma$ statistical-process-control limit). Using the lower confidence bound rather than the point estimate suppresses small-sample false alarms: a unit with few airframes produces a wide interval and does not alert unless its lower bound clears $U_\triangle$. In parallel, the coupling index is charted as a two-sided SPC statistic about its in-control level. An excursion *below* the band while $\widehat{V}_B$ remains far above unity signals burden saturation with the coupling structure intact, the marginals rise toward their ceiling and mechanically compress the ratios, whereas a drift of $\widehat{V}_B$ *toward* unity signals genuine decoupling, a new independent-noise source or a change to the shared pathway, which is the qualitatively more serious failure mode. Evaluated on consecutive windows $W_1, W_2, \ldots$ (for instance daily per fleet), the pair $(\widehat{P}_\triangle, \widehat{V}_B)$ is a two-channel SPC chart for joint kinematic integrity: the first channel detects *that* a regime change has occurred, the second classifies *what kind*.

6.3. Worked application: the Spring 2025 regime

The Spring 2025 baseline-elevation episode (Section 4.15) furnishes a natural positive control for the monitor; two nominal windows (the fifty days immediately preceding the episode, and the calendar-aligned 2024 window) furnish the in-control reference. These two windows span different calendar periods and years (winter–spring 2025 and spring 2024) and are used here only to *illustrate* the monitor against a single documented episode; the operational calibration of the upper control limit $U_\triangle$ is instead performed on the 315 non-episode days of 2025 at daily resolution (Section 6.4), a homogeneous single-vintage in-control reference. Table 8 reports both indices over the three windows, computed by the same eight-counter cluster-bootstrap pipeline as the headline analysis.[3] The two nominal windows agree closely and define a tight in-control band: $\widehat{P}_\triangle \in [0.037, 0.053]$ and $\widehat{V}_B \in [1.52, 1.55]$, the latter consistent with the all-sample value of 1.54 (Section 4.14). During the episode the burden index rises to $\widehat{P}_\triangle = 0.163$ (95% CI $[0.162, 0.164]$) (roughly $3.6\times$ the in-control mean and roughly $3\times$ its ceiling, a decisive burden alert at any $\alpha \leq 10^{-3}$) driven by single-channel elevations of $2.0\times$ (DV), $2.4\times$ (DVR), and $1.7\times$ (DH). The coupling index simultaneously *falls* to $\widehat{V}_B = 1.319$ ($[1.317, 1.322]$): below the in-control band, yet still more than two hundred bootstrap standard errors above the structural floor of unity. The monitor therefore reads the episode exactly as the full analysis of Section 4.15 does (a coherent exogenous elevation of all three channels' failure rates that compresses, but does not dissolve, the coupling that links them) and recovers that reading from the broadcast statistics alone, without any exogenous label. Figure 5 renders both views: panel (a) places the three windows on the (coupling, burden) plane against the in-control band, and panel (b) shows all three single-channel rates rising by similar factors in the episode.

6.4. Detection characteristics

The monitor's operating behaviour follows from two ingredients already established: the in-control reference set fixes the false-positive side, and the cluster-bootstrap interval fixes the power side. The per-window false-positive rate is by construction the in-control upper tail beyond $U_\triangle$, controlled at the chosen α. The detection rate at an operative burden $P^*_\triangle$ is the probability that $\widehat{P}^-_\triangle(W) > U_\triangle$ when the true prevalence is $P^*_\triangle$, and because the bootstrap interval narrows as $\sqrt{n_{\text{airframes}}}$, departures far smaller than the Spring 2025 separation are resolvable at the fleet-scale sample sizes typical of an operator-month or a sector-day. The coupling index supplies the second axis: at the same sample sizes its interval is narrow enough ($\widehat{V}_B$ to ± 0.002 over the episode) to separate a compression excursion from a drift toward unity, so

[3] The three windows come from a single run of the aircraft-hour pipeline (`phase3_11_monitor_windows`; 3,600 UTC hours, 209,599 airframes), reproducing the headline flag definitions exactly. On the pre-episode 2025 window it matches the corresponding Table 7 extraction to full floating-point precision ($\widehat{R} = 3.249$, $\widehat{V}_B = 1.523$); residual differences from that extraction (a 9-aircraft-hour count offset, and an $\approx 7\%$ lower 2024 yield giving $\widehat{R} = 3.25$ versus 3.38) reflect OpenSky reprocessing of the historical archive between the extraction dates (2026-05-22 and 2026-06-18), not a methodological change. A single extraction vintage is used across all three windows.

Table 8: The multi-channel consistency monitor (Section 6.1) over the Spring 2025 episode and two nominal reference windows. $\widehat{P}_\triangle$ is the burden index (Equation 10); $\widehat{V}_B$ the coupling index (Equation 11); $\widehat{R}$ the pairwise companion; brackets are within-airframe cluster-bootstrap 95% intervals. The episode trips the burden alert (Equation 12) while the coupling index stays far above the structural floor of unity, a burden surge with the coupling structure intact, not a decoupling.

Monitoring window	n (aircraft-h)	$\widehat{P}_\triangle$ [95% CI]	$\widehat{V}_B$ [95% CI]	$\widehat{R}$ [95% CI]
Spring 2025 episode (2 Apr–22 May)	16,261,423	0.163 [0.162, 0.164]	1.319 [1.317, 1.322]	2.153 [2.148, 2.158]
Pre-episode baseline (2025)	15,087,044	0.037 [0.037, 0.038]	1.523 [1.519, 1.528]	3.249 [3.235, 3.262]
Calendar-aligned 2024 window	14,622,953	0.053 [0.052, 0.053]	1.553 [1.549, 1.557]	3.253 [3.239, 3.266]

In-control band (two nominal windows): $\widehat{P}_\triangle \in [0.037, 0.053]$, $\widehat{V}_B \in [1.52, 1.55]$. The episode's burden lower bound (0.162) exceeds the in-control ceiling (0.053) by roughly 3×; its coupling index sits below the in-control band yet remains far above the structural floor of unity.

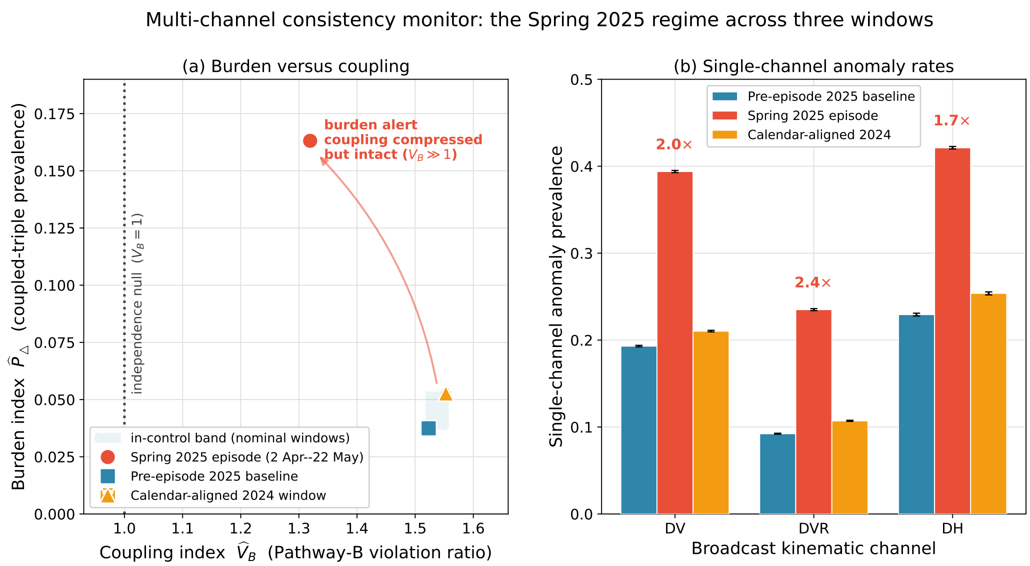

Figure 5: Multi-channel consistency monitor across the Spring 2025 episode and two nominal windows. **(a)** The (coupling, burden) state plane: the two nominal windows define the in-control band (shaded) and the episode is displaced to high burden $\widehat{P}_\triangle$ and compressed coupling $\widehat{V}_B$, while remaining far above the independence-null floor $V_B = 1$. **(b)** Single-channel anomaly prevalences: all three channels rise by similar factors in the episode (DV 2.0×, DVR 2.4×, DH 1.7× over the nominal mean), the coherent-burden signature. Error bars are within-airframe cluster-bootstrap 95% intervals; statistics from `phase3_11_monitor_summary.json`.

the monitor distinguishes a correlated burden surge from genuine decoupling rather than flagging both merely as "anomalous".

We characterize the operating point empirically by running the monitor at daily resolution across the full calendar year. Each of the 365 days of 2025 is treated as one monitoring window with its own burden and coupling indices, the per-day $\widehat{P}_{\triangle}$ and $\widehat{V}_B$ of the daily breakdown that underlies Figure 4, of which the 315 days outside the Spring episode form the in-control reference and the 50-day episode ([2 Apr, 22 May)) is the positive control. The in-control daily burden is tight (mean $\widehat{P}_{\triangle} = 0.0371$, standard deviation 0.0071; this daily-resolution mean lies just below the ≈ 0.045 two-window in-control reference of Section 6.3 (the mean of the pre-episode and 2024 window burdens, 0.037 and 0.053) because the 2024 window sits above the daily mean, and both estimates are deposited), so the upper control limit $U_{\triangle}$ calibrated to a target per-window false-positive rate α sits only modestly above it (Table A.13). Against this reference the episode is detected immediately and almost completely: at every $\alpha \in \{0.05, 0.01, 0.001\}$ the burden clears $U_{\triangle}$ on the *first* episode day (onset latency 0 days) and stays in alert for essentially the entire episode (48–50 of 50 days). The realized in-control average run length (ARL_0, the expected number of windows between false alarms) lengthens from ≈ 20 windows at $\alpha = 0.05$ to ≈ 315 windows, about one false alarm per year of daily monitoring, at $\alpha = 0.001$. The episode burden peaks at $\widehat{P}_{\triangle} = 0.245$ on 16 May, roughly 29 in-control standard deviations above the in-control mean.

The coupling index confirms the regime at the same resolution. Its in-control daily distribution is centred at $\widehat{V}_B = 1.560$ (standard deviation 0.048), and during the episode it falls to a mean of 1.324 with a minimum of 1.227, below the in-control band, yet still 4.7 in-control standard deviations *above* the structural floor of unity. The monitor therefore reads the episode, day by day, as a coordinated rise in all three single-channel burdens that compresses but does not dissolve the coupling that links them, exactly as the aggregate worked example of Section 6.3 reports. Figure 6 renders both axes across the year.

The monitor is therefore cheap (eight integer counters per window, two ratios, one bootstrap), interpretable (two calibrated indices on a shared SPC chart), and orthogonal to the per-channel integrity categories it complements. The daily characterization confirms that its separation on a documented real-world episode is decisive at any conventional false-positive budget.

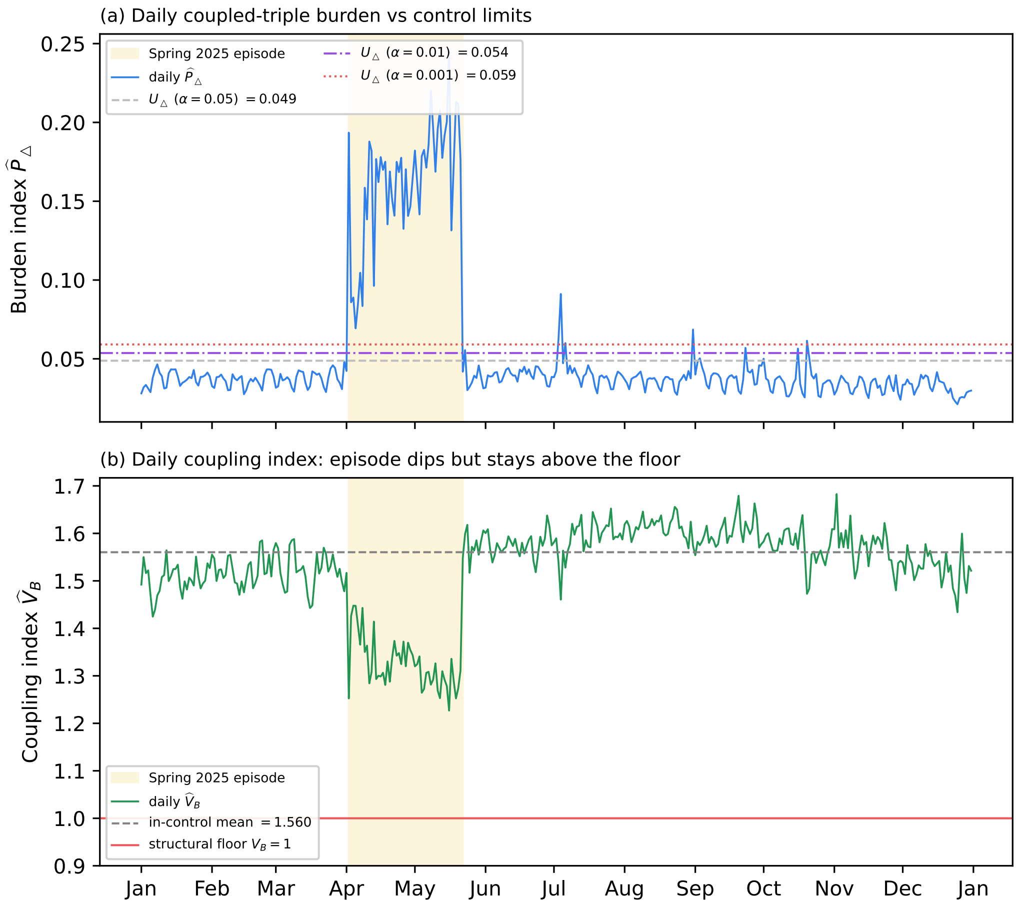

Figure 6: Daily application of the consistency monitor across 2025. *Panel (a)*: the coupled-triple burden $\widehat{P}_\triangle$ for each of the 365 daily windows against the upper control limits $U_\triangle$ at $\alpha \in \{0.05, 0.01, 0.001\}$; the 50-day Spring 2025 episode (shaded) clears every limit from its first day and peaks at 0.245 on 16 May. *Panel (b)*: the coupling index $\widehat{V}_B$ against its in-control mean (1.560) and the structural floor $V_B = 1$; during the episode $\widehat{V}_B$ dips below the in-control band but never approaches unity (episode minimum 1.227, still 4.7 in-control standard deviations above the floor), the signature of a burden surge with the coupling structure intact. Construction in Section 6.4.

6.5. Operational deployment

Operationally, the monitor is designed to sit downstream of an air navigation service provider's surveillance-data-processing system (SDPS), consuming the same decoded state-vector stream already used for separation services. It requires no airborne, transponder, or sensor modifications: it is a passive, ground-side computation over data the provider already ingests. An alert does not by itself identify a specific fault; it indicates that the cross-channel consistency signature has departed from its in-control baseline, prompting targeted review (of a receiver region, an airframe cohort, or a time window) rather than triggering any automatic operational action. Because each window's cost is dominated by accumulating eight integer counters over a sliding aggregate, the monitor scales to continental traffic on commodity hardware, complementing the per-channel integrity categories (NIC, NACv, SIL) rather than replacing them.

7. Limitations

7.1. Receiver-Level Confound Control

The OpenSky historical archive does not retain per-message receiver identifiers at the public-API level. We are therefore unable to directly condition the analysis on receiver identity, which would isolate broadcaster-side from receiver-side contributions to the observed coupling more precisely than the NACv and NIC stratifications accomplish. The inclusion of the ≥ 10 receiver redundancy tier (Section 4.10) is, however, a strong indirect control: any strictly receiver-side coupling explanation would predict attenuation when restricting to highly redundant observations.

7.2. Altitude Resolution

The headline Stage 2.1 aggregate did not retain per-hour median barometric altitude, which initially prevented stratifying the coupling by altitude regime. We have since re-extracted the aggregate retaining this field, reproducing the pooled headline enrichment ($R = 3.142$) exactly, and report the cruise-versus-terminal stratification in Section 4.14. Two residual limitations remain. First, band assignment uses the per-hour *median* barometric altitude, so aircraft-hours spanning a climb or descent are assigned to a single band by their central tendency rather than resolved within the hour; this coarsens, but does not bias, the band contrast. Second, barometric altitude is itself a quantity whose broadcast integrity is under study, so a transponder-level barometric fault would displace an aircraft-hour's band assignment as well as its DVR flag. Finer (within-hour, phase-of-flight) altitude resolution and a geometric-altitude cross-check are the natural next refinements.

7.3. Per-Hour Message-Count Control Granularity

The Stage 2.1 aggregate retains `total_pairs` (N_{pair}, the absolute count of valid per-step pairs per aircraft-hour) in the aircraft-hour record. This column is the explicit density control used in the cluster-robust regression of Section 4.11 (Model B adds log(total_pairs) as a main-effect covariate together with a DVR $\times$ log(total_pairs) interaction) and in the quintile stratification of Table A.3; both leave the headline coupling statistically unchanged at the data median and exhibit a density gradient opposite in direction to the disjunctive-aggregation confound prediction (Section 3.2). The residual limitation is therefore one of granularity rather than availability: the aircraft-hour density control operates at the regression and quintile level rather than the finer decile stratification that would most cleanly remove any remaining density-induced variation in $P(\mathrm{DV} \mid \mathrm{DVR})$, and the most direct treatment (the sub-second pair-level enrichment of Section 4.12, which sidesteps hourly aggregation entirely) is presently a 24-partition pilot rather than a full-year extraction. Scaling the pair-level analysis to the full calendar year is proposed as future work.

7.4. Survivorship Bias from Temporal-Validity Screening

The temporal-validity filter (Section 3.8) requires at least 60 valid per-step pairs per aircraft-hour for Dataset 3A, and at least 120 seconds of duration with at least 60 valid vertical-rate observations per 300-second window for Dataset 3B. The filter is designed to exclude transient-coverage records that would not support meaningful per-hour characterization, but it induces a sample-selection bias of unknown direction. Excluded aircraft-hours fall into at least four classes:

(a) hours during which severe physical events or platform anomalies caused sustained transponder dropouts or receiver-side decoding failures (anomaly-enriched);

(b) hours in marginal-coverage geography with low receiver density or high GDOP (anomaly composition unknown);

(c) brief transits through a single receiver's coverage cell (anomaly-deflated); and

(d) aircraft on/off transients near hour boundaries (anomaly composition unknown).

If class (a) dominates, the reported coupling magnitudes are lower bounds on the underlying coupling; if classes (c) or (d) dominate, they are upper bounds. We cannot resolve this direction without analysis of the discontinuous-coverage regime, which is by definition not represented in Dataset 3A. Direct quantification of this bias requires re-extraction of the dropout regime including its temporal clustering and persistence, which is proposed as immediate future work.

7.5. Vertical-Rate Source: Empirical Bound from Direct Re-Extraction

The OpenSky `vertrate` column of `state_vectors_data4` on which the headline aircraft-hour analysis is based does not preserve the DF17 BDS 0,9 source bit (VrSrc) that distinguishes barometric-derived from GNSS-derived vertical rate. Per DO-260B defaults, broadcasts are dominantly barometric, but a minority of DVR flags in the headline analysis originate from GNSS-derived vertical-rate readings; the headline aircraft-hour analysis cannot itself be filtered to a strict-barometric subset using `state_vectors_data4` alone. Appendix Appendix A.2 (Table A.7) addresses this through a direct re-extraction at the raw-message level: a pair-level pilot reading the `velocity_data4` table exposes the source bit (`baro`) and stratifies pair-level enrichment accordingly. The strict-barometric stratum (containing 926,777,558 pairs from 50,982 airframes) yields $R = 11.89\times$ ($[10.33, 13.70]$), exceeding the all-source pair-level baseline of $9.76\times$ and constraining the headline coupling to a regime that is present at strict-barometric source restriction rather than dependent on the GNSS-derived minority. The remaining limitation is that the direct re-extraction is restricted to the 24-partition pair-level sample of Section 4.12; replicating it at the full-year aircraft-hour scale of the headline would require a comprehensive aircraft-hour-level re-extraction at raw-message granularity, which we propose as immediate future work.

7.6. NOAA Storm Coverage is US-Only

The NOAA Storm Events Database covers US territory only. Convective activity over Europe, Asia, the Pacific, the Southern Hemisphere, and international waters during the Spring 2025 elevation is not captured by the cross-reference, which constrains our ability to attribute the elevation to a single specific operational driver.

7.7. Crowdsourced Receiver Coverage Bias

The OpenSky receiver network is denser in Europe and North America than in other regions. The geographic universality result (Section 4.9) controls for marginal sample size but cannot control for unobserved heterogeneity in receiver-network sensitivity. The cross-region direction of effect, enrichment everywhere in the same ballpark, is reassuring but does not eliminate the bias.

7.8. Data Vintage and a Reproducibility Protocol

The OpenSky historical archive is periodically reprocessed, so re-extracting an identical query at a later date can return a few percent fewer aircraft-hours, with correspondingly higher marginal prevalences. We observed this directly for the calendar-aligned Spring 2024 window, whose aircraft-hour yield fell $\approx 7\%$ and whose distinct-airframe count fell $\approx 16\%$ (192,747 $\rightarrow$ 161,309) between the 2026-05-22 and 2026-06-18 extractions; the 2025 windows were bit-stable over the same interval. The effect is immaterial to the conclusions (the year-over-year contrast and the burden alert are decisive under either vintage, and the coupling ratios shift by under one percent) but it is consequential for exact numerical reproduction. We therefore recommend, and adopt, a simple reproducibility protocol for crowdsourced-archive studies of this kind:

(i) report the extraction date for every window, as we do throughout;

(ii) treat any single-vintage marginal prevalence as carrying a few-percent reprocessing uncertainty, and base year-over-year or episode-versus-baseline contrasts on windows extracted in a *single* vintage, as the monitor of Section 6 does across all three of its windows; and

(iii) where the archive exposes a processing-snapshot or version identifier, pin and report it so a later reader can request the same snapshot.

This converts the reprocessing sensitivity from an unstated reproducibility hazard into a declared and bounded uncertainty.

7.9. Descriptive Non-Causal Framing

The findings characterize the joint distributional structure of the broadcast kinematic channels. They are non-causal: we do not identify the specific physical mechanism responsible for the multi-dimensional coupling, nor do we attribute the coupling to any specific equipment vendor, operator class, or regulatory regime.

7.10. Outstanding Follow-Up Work

Three pieces of follow-up work remain immediately tractable.

(i) Direct investigation of the candidate 2025-specific operational drivers of the Spring 2025 elevation, by cross-reference of the day-by-day daily $P(\mathrm{DV})$ trajectory against ICAO and EUROCONTROL safety bulletin archives, GNSS constellation status data, and major OEM avionics firmware release announcements during the calendar year.

(ii) Receiver-level stratification of the conditional dependence to disentangle broadcaster-side from receiver-side contributions, which the current public OpenSky data product does not directly support.

(iii) External-event validation of the monitor: because the Spring 2025 episode is a detection benchmark whose operational driver is unidentified, an independent test against *documented* events (overlaying the date-and-region windows of published GNSS-interference advisories (EASA Safety Information Bulletins, EUROCONTROL and FAA NOTAM interference notices) on a region-resolved daily burden series) would establish that the monitor responds to ground-truth-labelled disturbances and not only to the unlabelled 2025 episode; this requires a region-resolved daily re-extraction beyond the global daily series analyzed here.

Each is proposed as immediate future work.

8. Conclusion

Across 117.5 million aircraft-hours of OpenSky Network state-vector data from calendar year 2025, severe per-step rate-normalized anomalies in broadcast groundspeed, vertical rate, and track angle exhibit a structural multi-dimensional conditional dependence. The headline pairwise enrichment of 3.14× is more than a tripling of the marginal $P(\mathrm{DV})$; the three-way conditional-independence violation $V_B = 1.543$ is a 54% deviation from the conditional-independence prediction. Both observed values fall entirely outside the 1,000-iteration within-cluster permutation null distribution constructed in Section 3.9.5, with no permuted iteration approaching the observed magnitude. The coupling is robust to Δt-conflation controls, to DVR threshold choice, to four seasonal-checkpoint weeks, to six continental airspaces, to a strict-Fixed-Wing specification, to four engine-type strata, to four symmetric phase-classification threshold variants, and, decisively for the receiver-side-noise alternative, intensifies under strict intersection constraints requiring high transponder integrity (NACv ≥ 3, NIC ≥ 8) and high receiver redundancy (≥ 10 distinct receivers).

The pattern of violation ratios is most economically explained by treating the vertical-rate anomaly as the principal (though not sole) mediator of the coupling between the groundspeed-divergence and the heading-rate-divergence channels (with a small residual conditional dependence quantified by $V_I = 1.149$). Consistent with the descriptive, non-causal scope of this study (Section 7.9), this is a mechanistic hypothesis compatible with the data rather than a causal identification, and is consistent with antenna-shadowing or shared-platform-modulation mechanisms operating during dynamic manoeuvre. Rotary-wing aircraft exhibit a fundamentally distinct sub-multiplicative phase $\times$ DH interaction structure that the saturation grid confirms is genuinely substantive rather than a logistic-link ceiling artefact.

A sustained 50-day baseline-elevation episode during April–May 2025, spatially concentrated in Europe and North America and disproportionately affecting jet operations, is documented as a major secondary finding. Direct re-extraction of the calendar-aligned 50-day window of Spring 2024 (15.7 million aircraft-hours over 192,747 airframes) returns baseline marginal prevalences and baseline coupling structure, ruling out a recurring seasonal phenomenon. Cross-reference against the GFZ Potsdam daily geomagnetic and solar indices archive does not support an elevated-Solar-Cycle-25 or ionospheric-perturbation driver, with the daily sunspot number and $F_{10.7}$ solar radio flux substantially *lower*, not higher, during the elevation window than during the rest of 2025. The 2025 elevation is characterized as a 2025-specific operational event whose precise driver (most likely a coordinated avionics firmware wave, a GNSS constellation event, or an ANSP regulatory transition) awaits identification.

The findings establish that multi-dimensional kinematic-anomaly co-occurrence in cooperative-surveillance telemetry is structural, not coincidental, and provide both a methodological framework (conditional-dependence inference with cluster-bootstrap and permutation null testing) and a direct operational deliverable (the per-airframe within-cluster V_B ratio) for treating the joint multi-channel consistency as a first-class diagnostic for ANSP infrastructure monitoring.

Data Availability Statement

The full extraction pipeline source code (Stages 1 and 2.1–2.4), Phase 3 robustness analysis scripts, configuration file, intermediate manifests, and the present analytical products are publicly released at https://doi.org/10.5281/zenodo.20387308 [45]. The underlying OpenSky Network state-vector archive is publicly available via https://opensky-network.org subject to OpenSky's terms of service. The NOAA Storm Events Database used for the Spring 2025 cross-reference is publicly available at https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. All compute and storage costs were borne by the corresponding author.

Declaration of Competing Interest

The author declares no competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author thanks the OpenSky Network and its volunteer receiver contributors for the underlying telemetry archive that makes this work possible.

CRediT Authorship Contribution Statement

Eugene Pik: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

Appendix A. Supplementary Tables and Diagnostics

This appendix collects the robustness, universality, and detection tables that support the main analysis. Each is referenced at its point of use in the body; they are gathered here to keep the main narrative focused.

Appendix A.1. Data-Quality, Threshold, Density, and Indicator-Definition Robustness

Table A.1 reports the pairwise enrichment and the Pathway-B violation ratio within the PureDQ, HighDQ, and LowDQ tiers; the corresponding per-tier values of V_A and V_I are released in `phase3_2_dq_tier_table.csv`. The enrichment is *monotonically decreasing* as Δt-contamination increases: 3.156× in the cleanest tier (PureDQ), 3.067× in HighDQ ($0.999 \leq f_{\Delta t=1} < 1.000$), and 2.783× in LowDQ ($f_{\Delta t=1} < 0.999$). The Pathway-B violation V_B is non-monotonic across tiers (1.557, 1.412, 1.520) but remains well above unity in every tier. The cluster-bootstrap 95% confidence intervals for Enrichment do not overlap across the three tiers; the intervals for V_B also do not overlap across the three tiers.

The direction of the gradient is the opposite of what a Δt-conflation alternative would predict. If the apparent coupling were caused by data-link gaps lengthening Δt and thereby spuriously inflating $|\Delta v|/\Delta t$, $|\Delta w|/\Delta t$, and $|\Delta h|/\Delta t$ simultaneously, the enrichment would be highest in LowDQ and lowest in PureDQ. The observed monotone decrease in the opposite direction therefore refutes this alternative explanation.

Table A.1: Three-tier data-quality decomposition of the pairwise enrichment and the Pathway-B violation ratio (Section 4.5).

Tier	N (aircraft-hours)	Enrichment	Enrichment 95% CI	V_B	V_B 95% CI
PureDQ ($f_{\Delta t=1} = 1.000$)	103,798,609	3.156	[3.146, 3.166]	1.557	[1.554, 1.560]
HighDQ ($0.999 \leq f_{\Delta t=1} < 1.000$)	11,147,823	3.067	[3.055, 3.079]	1.412	[1.410, 1.415]
LowDQ ($f_{\Delta t=1} < 0.999$)	2,596,299	2.783	[2.769, 2.798]	1.520	[1.512, 1.527]

Note. Tier definitions are given in Section 3.4. Cluster-bootstrap 95% confidence intervals are reported for both enrichment and V_B; the intervals do not overlap across tiers for either statistic. Per-tier values of V_A and V_I are released in the reproducibility deposit [45] as `phase3_2_dq_tier_table.csv`.

Section 3.2 acknowledged that disjunctive hourly aggregation makes the marginal hour-level prevalence a non-linear function of the per-hour valid-pair count, and three indirect lines of evidence were given against this confound as an explanation of the headline coupling. We now report the explicit cluster-robust regression specification that controls for the log per-hour message density directly. The control variable is `total_pairs`, the Stage 2.1 column emitting the total valid per-step pair count within each aircraft-hour (range in the analytic sample: 60 to 3,599; median 1,863; log-range $[4.094, 8.188]$).

Two cluster-robust logistic regressions were fitted on the Dataset 3A low-latency sample of 117,542,731 aircraft-hours across 240,001 airframes (cluster-robust standard errors by `icao24`, `lbfgs` optimization, maximum 500 iterations). Model A is the baseline specification flag_dv_rate $\sim$ flag_dvr_rate + flag_dh_rate. Model B extends Model A with log(total_pairs) as a main-effect covariate and a flag_dvr_rate : log(total_pairs) interaction. Coefficient estimates are reported in Table A.2.

Model A yields a DVR odds ratio of 8.111 (95% CI $[8.054, 8.169]$). This is the close regression analogue of the headline pairwise enrichment $R = 3.142\times$ of Section 4.2: the algebraic identity $\mathrm{OR} = R \cdot (1 - P(\mathrm{DV}))/(1 - R \cdot P(\mathrm{DV}))$ maps $R = 3.142$, under the observed $P(\mathrm{DV}) = 0.221$, to a *marginal* DVR odds ratio of 8.01 (the same coupling expressed on the odds and probability scales); Model A's coefficient exceeds this marginal value by under 2% because the regression additionally conditions on DH.

Model B yields a DVR main-effect OR of 12.157 ($[11.712, 12.618]$). This larger value is a centering artifact rather than a real inflation, since in the interaction specification the main-effect coefficient is the DVR effect evaluated at log(total_pairs) = 0, which lies far outside the empirical data range. At the median log(total_pairs) = 7.530 (corresponding to `total_pairs` = 1,863), Model B implies a DVR OR of $12.157 \times (0.946)^{7.530} = 8.01$, statistically indistinguishable from the Model A value of 8.111. The interaction OR of 0.946 ($[0.941, 0.951]$, $p = 1.2 \times 10^{-99}$) indicates the DVR effect attenuates modestly with density: across the empirical log-density range, the implied DVR OR varies from 9.69 at `total_pairs` = 60 to 7.73 at `total_pairs` = 3,599. Both endpoints are well above 1.0 and the variation direction is opposite to the density-confound hypothesis prediction: under that hypothesis, denser hours should show *stronger* apparent coupling (because the disjunctive aggregation inflates marginal flag-firing rates), but we observe *weaker* coupling at higher density.

The quintile-stratified pairwise enrichment $R(\mathrm{DV} \mid \mathrm{DVR})$ is reported in Table A.3 and ranges from 2.60 to 4.61 across the five `total_pairs` quintiles, with no monotonic trend in the direction the confound predicts. The middle three quintiles (Q2–Q4, comprising approximately 60% of the sample and spanning `total_pairs` from 842 to 3,281) yield R values between 2.60 and 2.76; the lowest-density quintile (Q1, `total_pairs` 60–841) yields $R = 3.92$; the highest-density quintile (Q5, `total_pairs` 3,282–3,599) yields $R = 4.61$. The U-shape reflects edge effects: Q1 captures brief or marginal-coverage flights with rare DV events ($P(\mathrm{DV}) = 0.161$), and Q5 captures near-maximum-coverage stable-cruise flights also with low DV rates ($P(\mathrm{DV}) = 0.151$); both retain conditional probabilities $P(\mathrm{DV} \mid \mathrm{DVR})$ above 0.63. The U-shape does not match the monotonic increase with density that the confound hypothesis predicts. Two caveats bear noting for transparency. First, the densest quintile (Q5) does carry the highest enrichment, so the decisive density test is the regression interaction of Table A.2 (which attenuates with density, opposite to the confound prediction) rather than the quintile shape per se. Second, because the middle three quintiles sit at $R \approx 2.60$–2.76, the pooled headline $R = 3.142\times$ is an aggregation-weighted value, raised above the typical within-stratum level by the two low-marginal edge quintiles; it nonetheless lies well within the empirical range of quintile-stratified values.

The disjunctive-aggregation density confound is therefore directly refuted by the explicit regression specification. The headline coupling magnitude is statistically invariant under cluster-robust regression control for log(total_pairs) at the data median, and the small interaction term identified by the 1.2×10^{-99} Wald test operates in the direction opposite to what the confound hypothesis predicts.

Table A.2: Cluster-robust logistic regression coefficients for the explicit density-confound test (Section 4.11).

Term	Model A: OR [95% CI]	Model B: OR [95% CI]
Intercept	0.103 [0.103, 0.104]	0.064 [0.063, 0.066]
DVR rate	8.111 [8.054, 8.169]	12.157 [11.712, 12.618]
DH rate	6.550 [6.507, 6.593]	6.556 [6.513, 6.599]
$\log N_{\mathrm{pair}}$	—	1.066 [1.063, 1.069]
DVR rate $\times \log N_{\mathrm{pair}}$	—	0.946 [0.941, 0.951]

Note. Model A is the baseline specification (flag_dv_rate $\sim$ flag_dvr_rate + flag_dh_rate); Model B adds log(total_pairs) as a main-effect covariate and a flag_dvr_rate : log(total_pairs) interaction. Fit on the Dataset 3A low-latency sample (N = 117,542,731 aircraft-hours); 95% cluster-robust confidence intervals clustered by `icao24` (N = 240,001 clusters). Regressor names map to the statsmodels output as DVR rate $\equiv$ `flag_dvr_rate`, DH rate $\equiv$ `flag_dh_rate`, $N_{\mathrm{pair}} \equiv$ `total_pairs`; the interaction corresponds to coefficient `flag_dvr_rate:log_npair`. Model A's flag_dvr_rate OR (8.111) is the DH-adjusted regression analogue of the headline pairwise enrichment $R = 3.142\times$; the marginal-OR identity under $P(\mathrm{DV}) = 0.221$ maps R to a marginal odds ratio of 8.01, within 2% of the Model A coefficient.

The asymmetric ($+3.5/-5.5$ m/s) vertical-rate thresholds used in the

Table A.3: Pairwise enrichment $R(\mathrm{DV} \mid \mathrm{DVR})$ stratified by quintiles of the per-hour `total_pairs` count (Section 4.11).

Quintile	N_{pair} range	Median N_{pair}	Aircraft-hours	Airframes	$P(\mathrm{DV})$	R
Q1	60–841	511	23,522,500	228,568	0.161	3.92
Q2	842–1,505	1,167	23,522,826	218,277	0.243	2.76
Q3	1,506–2,256	1,864	23,497,444	209,623	0.271	2.60
Q4	2,257–3,281	2,715	23,493,787	198,890	0.282	2.60
Q5	3,282–3,599	3,598	23,506,174	169,610	0.151	4.61

Note. Computed on the Dataset 3A low-latency sample ($N = 117{,}542{,}731$ aircraft-hours); quintile cutpoints set by sample quintiles of `total_pairs`. "Airframes" counts distinct `icao24` airframes appearing in each quintile; airframes contribute to multiple quintiles across different aircraft-hours, so the column sum exceeds the global $N = 240{,}001$ unique airframes. The headline aircraft-hour-level $R = 3.142\times$ of Section 4.2 lies within the empirical range $[2.60, 4.61]$ of the quintile-stratified values. The non-monotonic (U-shaped) pattern is incompatible with the monotonic-increase-with-density signature that the disjunctive-aggregation-density confound would produce; see Section 4.11 for discussion.

primary phase classification (Section 3.7) reflect commercial operating profiles in which descent rates are typically steeper than climb rates. We tested the sensitivity of Model 3 to this asymmetry by re-extracting Dataset 3B over all four seasonal-checkpoint weeks with three alternate symmetric threshold sets at $\{\pm 2.5,\ \pm 3.5,\ \pm 4.5\}$ m/s and re-fitting the cluster-robust logistic regression on the strict-Fixed-Wing DVR $= 1$ sub-population ($N = 1{,}587{,}481$ windows in the seasonal-week sample).

Table A.4 reports the resulting odds ratios. The Climb, Descent, and DH main effects are stable across all four threshold sets to within approximately $\pm 10\%$ (the widest, the DH main effect, ranges from 3.72 at ± 4.5 m/s to 4.41 at ± 2.5 m/s; Descent spans 2.09–2.33). The Climb $\times$ DH and Descent $\times$ DH interactions are super-multiplicative throughout, with magnitudes between $1.06\times$ and $1.52\times$. The single coefficient that exhibits material threshold sensitivity is Transition $\times$ DH, which crosses the indifference threshold (OR $=$ 1) as the symmetric threshold increases, ranging from $0.78\times$ at ± 2.5 m/s to $1.39\times$ at ± 4.5 m/s. The phase definition of "Transition" is the residual category that absorbs windows not classified as Climb/Cruise/Descent; its content therefore varies with the threshold, and the corresponding interaction is correspondingly less stable. The headline finding (that Climb, Descent, and DH are robust independent stressors with super-multiplicative coupling in the

Climb and Descent sub-strata) is invariant under the symmetric-threshold sensitivity.

Table A.4: Symmetric-threshold sensitivity of the Model 3 phase × DH regression (Section 4.13).

Variable	**Original** $(+3.5/-5.5)$	**Sym.** $\pm$**2.5**	**Sym.** $\pm$**3.5**	**Sym.** $\pm$**4.5**
Climb	3.03	2.89	3.06	3.12
Descent	2.33	2.09	2.22	2.27
Transition	1.39	1.09	1.30	1.49
DH	4.08	4.41	3.92	3.72
Climb × DH	1.37	1.24	1.43	1.52
Descent × DH	1.13	1.06	1.31	1.35
Transition × DH	0.95	0.78	0.98	1.39

Note. "Original asymmetric" is the primary specification $(+3.5/-5.5\,\mathrm{m/s})$; the three symmetric specifications replace both climb and descent cutoffs with the same absolute value. $N = 1{,}587{,}481$ strict-Fixed-Wing DVR $= 1$ windows from the four seasonal-checkpoint weeks.

The DVR threshold sweep of Section 4.6 varies one channel's threshold. As a complementary check that the headline coupling does not depend on the specific DV and DH indicator operationalization, Table A.5 recomputes the pairwise enrichment $R(\mathrm{DV} \mid \mathrm{DVR})$ and the Pathway-B violation V_B on the full low-latency Dataset 3A under the three indicator definitions extracted in Stage 2.1 (Section 3.3): the raw absolute-difference definition (Track 1), the rate-normalized definition used throughout the main analysis (Track 2), and the $\Delta t \leq 5\,\mathrm{s}$ phys-restricted definition (Track 3). The enrichment (3.125, 3.142, 3.141) and the three-way violation (1.535, 1.543, 1.543) are near-identical across the three, even though Track 1 abandons the rate normalization entirely in favour of an absolute-jump criterion. The coupling is therefore not an artefact of the rate-normalized threshold choice, consistent with the monotone DVR-threshold response of Section 4.6. The alternate-threshold columns needed for a finer DV- or DH-specific sweep were not pre-extracted in Stage 2.1 (only the DVR channel carries them), so a per-threshold DV × DH grid is left to a future extraction.

We re-computed the pairwise enrichment at five alternate DVR thresholds $\{2, 5, 8, 10, 12\}\,\mathrm{m/s^2}$, using the alternate-threshold columns pre-extracted in Stage 2.1. The enrichment exhibits a clear monotonic increase from 1.56× at the loosest threshold ($\mathrm{DVR} > 2\,\mathrm{m/s^2}$) to 3.59× at the strictest ($\mathrm{DVR} > 12\,\mathrm{m/s^2}$). The monotonicity rules out a threshold-dependent statistical artefact: the relationship between DV and DVR strengthens precisely as the DVR indicator becomes more physically extreme, exactly as a real coupling between the two physical phenomena would predict. This monotone strengthening is also direct empirical evidence against a per-hour message-density confound (Section 3.2), because any density-driven inflation of disjunctive hour-level marginals operates on the indicator-firing event itself and is invariant to the within-pair threshold value; the threshold-sensitive component of the enrichment must therefore reflect within-pair coupling structure rather than between-hour density variation. The sweep varies the DVR threshold because alternate-threshold columns were pre-extracted only for that channel; the DV and DH thresholds were likewise pre-specified on physical grounds (Section 3.2.2) and not tuned. As a direct check that the coupling does not hinge on the DV/DH indicator operationalization, the headline enrichment and Pathway-B violation are near-invariant across the three indicator definitions carried in the deposit (raw absolute-difference (Track 1), rate-normalized (Track 2, primary), and $\Delta t \leq 5\,\mathrm{s}$ phys-restricted (Track 3)) as reported in Appendix Appendix A.1 (Table A.5); the coupling is not a knife-edge artefact of any single cut-off, though a fine-grained joint DV $\times$ DH threshold grid was not pre-extracted.

Table A.5: Indicator-definition robustness of the headline coupling (Section 4.6). The pairwise enrichment $R(\mathrm{DV} \mid \mathrm{DVR})$ and the Pathway-B violation V_B are recomputed on the full low-latency Dataset 3A under the three indicator operationalizations extracted in Stage 2.1 (Section 3.3).

Track	**Indicator definition**	$P(\mathbf{DV})$	$P(\mathbf{DV} \mid \mathbf{DVR})$	R	V_B
Track 1	Raw absolute difference	0.2229	0.6965	3.125	1.535
Track 2	Rate-normalized (primary)	0.2214	0.6957	3.142	1.543
Track 3	Phys-restricted ($\Delta t \leq 5\,\mathrm{s}$)	0.2215	0.6957	3.141	1.543

Note. All three definitions are computed on the identical Dataset 3A low-latency sample ($N = 117{,}542{,}731$ aircraft-hours, 240,001 airframes). Track 1 flags on the absolute consecutive-message difference without rate normalization; Track 2 (the headline definition) normalizes by the inter-message interval; Track 3 additionally restricts to pairs with $\Delta t \leq 5\,\mathrm{s}$. The enrichment and V_B are near-identical across all three, establishing that the coupling does not depend on the DV/DH indicator operationalization. Computed by `phase3_14_dv_dh_threshold_audit.py` (output `phase3_14_track_robustness_table.csv`).

Appendix A.2. Integrity, Redundancy, and Vertical-Rate-Source Robustness

The headline aircraft-hour-level analysis of Sections 4.2–4.10 aggregates per-step anomaly indicators to the hour, raising the methodological question of whether the observed coupling reflects genuine within-second co-occurrence of DV, DVR, and DH or merely co-occurrence within the same one-hour window. To address this directly we ran a pair-level pilot analysis on 24 sampled hours distributed across the four seasonal-checkpoint weeks of Section 4.15 (six hours per checkpoint at $\{00, 04, 08, 12, 16, 20\}$ UTC), in which the unit of analysis is the individual consecutive-message pair rather than the aggregated aircraft-hour. The pair-level DV, DVR, and DH flags are the rate-normalized per-step indicators of Equations 1–3 *before* disjunctive aggregation; the within-cluster bootstrap resamples airframes as in the headline analysis. The total sample is 1,176,250,672 consecutive-message pairs across 76,023 airframes.

Table A.6 reports the result. The pair-level pairwise enrichment $R(\mathrm{DV} \mid \mathrm{DVR})$ on the pooled sample is $9.755\times$ (95% CI $[8.633, 11.120]$), restricting to message pairs from transponders self-reporting strict-high velocity accuracy (NACv ≥ 2, defined at the pair level as $\mathrm{LEAST}(\mathrm{NAC}_i, \mathrm{NAC}_{i-1}) \geq 2$, i.e. both endpoints carrying NACv ≥ 2) raises this to $19.915\times$ ($[17.532, 23.118]$), and the unknown-NACv stratum yields $6.886\times$ ($[6.026, 8.088]$). The larger ratio in the high-integrity stratum is driven by its lower marginal $P(\mathrm{DV}) = 0.040$ (versus 0.129 in the unknown stratum) rather than by a stronger conditional dependence: the conditional $P(\mathrm{DV} \mid \mathrm{DVR})$ is in fact slightly *lower* in the high-integrity stratum (0.798 versus 0.888). The structurally meaningful comparison is therefore the Pathway-B violation, which rises modestly from $V_B = 1.046$ (unknown) to $V_B = 1.120$ (NACv ≥ 2), confirming that the coupling survives, rather than dissolves, under high transponder integrity. The pair-level NACv threshold (≥ 2, corresponding to $< 3\,\mathrm{m/s}$ 95% velocity error per DO-260B Table 2-25) is looser than the hourly NACv threshold (≥ 3, corresponding to $< 1\,\mathrm{m/s}$) used in Section 4.10, and the aggregation rules differ (LEAST over the pair versus MIN over the hour); the two strata are not directly comparable and the pair-level value at the stricter ≥ 3 threshold is not extracted in the current pilot. The unknown-NACv stratum exhibits a high per-pair conditional probability $P(\mathrm{DV} \mid \mathrm{DVR} = 1) = 0.888$, which is large for two independent physical events but is consistent with the principal-mediator interpretation developed in Section 5.2: a shared upstream pathway (encoder-level co-corruption, GNSS-velocity-solution discontinuity, or shared-source state-vector perturbation) would produce exactly this kind of high pair-level co-occurrence. All three pair-level values exceed the aircraft-hour-level baseline of 3.142×. The coupling therefore *intensifies* as the temporal window narrows from one hour to one consecutive-message pair, which is the opposite of what a hourly-window-confounded explanation would predict. We interpret the aircraft-hour-level enrichment of 3.142× as a conservative bound on a stronger underlying pair-level coupling that is partially compressed by disjunctive hourly aggregation, while acknowledging that the two estimands operate at different temporal granularities and therefore do not in principle estimate the same physical quantity (see Section 7.3). Because the enrichment is larger at the consecutive-message-pair scale (9.76×) than at the one-hour scale (3.142×), any intermediate aggregation window (a ten-minute or per-flight unit, for example) is expected to fall between these two limits, so the one-hour choice is a conservative aggregation rather than a tuned parameter and no qualitative conclusion depends on it.

The pair-level enrichment is not reducible to the GNSS-derived minority of vertical-rate readings: a direct re-extraction from `velocity_data4` that recovers the DF17 BDS 0,9 vertical-rate source bit (VrSrc) and stratifies the pooled pair-level sample by barometric versus geometric source is reported in Appendix Appendix A.2 (Table A.7); restriction to the strict-barometric source *exceeds*, rather than attenuates, the all-source pair-level baseline of 9.76×.

Table A.6: Pair-level pairwise enrichment of DV given DVR in the 24-hour NACv pilot, by transponder velocity-accuracy stratum (Section 4.12).

NACv stratum	$N_{\mathbf{pairs}}$	**Airframes**	P(**DV** | **DVR**)	**Enrichment [95% CI]**
Unknown NACv (= 0)	653,041,171	54,302	0.8881	6.886 [6.026, 8.088]
Strict High NACv (≥ 2)	523,209,501	28,568	0.7978	19.915 [17.532, 23.118]
All pairs (pooled)	1,176,250,672	76,023	0.8724	9.755 [8.633, 11.120]

Note. The unit of analysis is the consecutive-message pair within an airframe, not the aircraft-hour. Cluster-bootstrap 95% confidence intervals over 1,000 iterations resampling by airframe.

To address the limitation noted in Section 7.5 that the OpenSky `state_vectors_data4` product does not preserve the DF17 BDS 0,9 vertical-rate source bit (VrSrc), we re-extracted the identical 24-partition pair-level sample of Section 4.12 directly from `velocity_data4`, which exposes the source bit as the `baro` column (TRUE = barometric source, FALSE = geometric/GNSS source). Each consecutive-message pair was classified by the source bits at both endpoints into three strata: strict barometric (both endpoints `baro = true`), strict geometric (both endpoints `baro = false`), or mixed source (endpoints differ). Results are reported in Table A.7. The strict-barometric stratum, which accounts for 78.8% of pairs and contains 50,982 airframes, yields $R = 11.89\times$ (95% CI $[10.33, 13.70]$), exceeding the all-source pair-level baseline of $9.76\times$. The strict-geometric stratum yields a lower $R = 5.86\times$, but the conditional probability $P(\mathrm{DV} \mid \mathrm{DVR}) = 0.905$ in that subset is the highest of any stratum. The smaller enrichment ratio reflects an elevated marginal $P(\mathrm{DV}) = 0.154$ rather than weaker conditional dependence, the same arithmetic relationship described in Section 4.15 for the Spring 2025 elevation. The mixed-source stratum, comprising pairs in which the broadcast source bit changes between consecutive messages, represents only 0.01% of pairs and exhibits substantially weaker coupling ($R = 2.90$, $P(\mathrm{DV} \mid \mathrm{DVR}) = 0.136$), consistent with source-mode-transition events being a quantitatively distinct phenomenon from the dominant cross-channel coupling. The pair-level enrichment is therefore not reducible to the GNSS-derived minority of vertical-rate readings: restriction to strict-barometric source yields coupling that exceeds, rather than attenuates, the all-source baseline.

Table A.7: Pair-level enrichment $R(\mathrm{DV} \mid \mathrm{DVR})$ stratified by the DF17 BDS 0,9 vertical-rate source bit (Section 4.12).

Source stratum	N pairs	Airframes	$P(\mathrm{DV})$	$P(\mathrm{DV} \mid \mathrm{DVR})$	R [95% CI]
Strict barometric	926,777,558	50,982	0.072	0.855	11.89 $[10.33, 13.70]$
Strict geometric	249,347,851	25,829	0.154	0.905	5.86 $[4.79, 7.51]$
Mixed source	125,263	2,660	0.047	0.136	2.90 $[2.23, 4.21]$

Note. Re-extracted directly from `velocity_data4` on the same 24-partition sample as Table A.6; strata classified by the VrSrc source bits at both pair endpoints (strict-barometric = both endpoints barometric, strict-geometric = both geometric/GNSS, mixed = disagreement). 95% CIs from within-cluster bootstrap by `icao24` (1,000 iterations). The strict-barometric stratum exceeds the all-source pair-level baseline of $R = 9.76\times$ ($N = 1{,}176{,}250{,}672$ pooled pairs); the strict-geometric stratum's lower $R = 5.86\times$ reflects an elevated marginal $P(\mathrm{DV})$ rather than a weaker conditional dependence, since $P(\mathrm{DV} \mid \mathrm{DVR}) = 0.905$ in that stratum is the highest of any.

Appendix A.3. Engine, Airframe, and Geographic Universality

We tested whether the pairwise coupling is restricted to a specific engine-type lineage or a specific continental airspace. Within Fixed-Wing aircraft, the pairwise enrichment is $3.32\times$ (PureDQ: $3.31\times$) for Jet aircraft ($N = 79.1$M aircraft-hours), $3.24\times$ (PureDQ: $3.30\times$) for Turboprop ($N = 6.7$M), and

2.68× (PureDQ: 2.75×) for Piston ($N = 16.6$M). The enrichment is well-resolved and substantial in all three of these strata. The Electric subset ($N = 1{,}714$ aircraft-hours) yields a point estimate of 1.81× with 95% CI $[0.79, 2.26]$, whose lower bound is below unity. The Electric stratum is too small to resolve the coupling at conventional confidence and the universality claim is restricted to the Jet, Turboprop, and Piston classes. At the coarser airframe-class level, the engine-classified Fixed-Wing aircraft ($N = 102.4$M aircraft-hours, the metadata-labelled subset carrying a resolved engine type rather than the entire fixed-wing class) exhibit a pairwise enrichment of 3.19× (95% CI $[3.18, 3.20]$) and Rotary-Wing aircraft ($N = 3.55$M) a lower but still substantial 2.25× ($[2.20, 2.29]$). The reduced rotary-wing ratio reflects its elevated baseline anomaly prevalence ($P(\text{DV}) = 0.292$ versus 0.218 for Fixed-Wing) rather than weaker conditional coupling, the irreducible three-way violation being in fact *larger* for rotary-wing ($V_B = 1.68$) than for fixed-wing ($V_B = 1.52$). The full per-class columns (Full and PureDQ) appear in Table A.8.

Table A.8: Engine-type and airframe-class universality of the pairwise enrichment (Section 4.9). Engine-type rows are restricted to Fixed-Wing airframes with a classified engine; the Fixed-Wing airframe-class row is the union of the engine-classified fixed-wing subsets (the metadata-labelled subset, not the entire fixed-wing class), and the Rotary-Wing row is the whole-class rotary population.

Stratum	**N (aircraft-hours)**	**Enrichment, Full**	**Enrichment, PureDQ**
Engine type (within Fixed-Wing)			
Jet	79,120,575	3.32 [3.30, 3.33]	3.31 [3.29, 3.32]
Turboprop	6,691,452	3.24 [3.20, 3.29]	3.30 [3.26, 3.35]
Piston	16,631,293	2.68 [2.67, 2.69]	2.75 [2.74, 2.76]
Electric	1,714	1.81 [0.79, 2.26]	1.97 [0.94, 2.41]
Airframe class			
Fixed-Wing	102,445,034	3.19 [3.18, 3.20]	3.20 [3.19, 3.21]
Rotary-Wing	3,554,852	2.25 [2.20, 2.29]	2.30 [2.24, 2.35]

Note. Aircraft-hour-level conditional enrichment is the ratio $P(\text{DV} \mid \text{DVR})/P(\text{DV})$ within the indicated subset. Bracketed values are 95% cluster-bootstrap confidence intervals over 1,000 iterations. The Electric subset is too small to draw firm conclusions. The Fixed-Wing airframe-class row (102,445,034 aircraft-hours) is the union of the four classified-engine fixed-wing subsets above; the Rotary-Wing row is the whole-class rotary population, and its lower enrichment reflects an elevated marginal $P(\text{DV})$ (the irreducible three-way violation V_B is in fact larger for Rotary-Wing, 1.68, than for Fixed-Wing, 1.52; Section 4.9).

The geographic stratification across six continental bounding boxes[4] yields enrichments of 3.13× (North America), 3.44× (Europe), 2.76× (Asia), 2.60× (South America), 3.25× (Africa), and 3.80× (Oceania). The smaller-sample regions exhibit slightly wider intervals, but the coupling is substantial and statistically detectable in every region. We interpret the geographic universality and the engine-type universality together: the conditional dependence is a property of cooperative-surveillance telemetry under dynamic manoeuvre, not of a specific avionics generation or a specific airspace. The per-region values appear in Table A.9.

Table A.9: Geographic universality of the pairwise enrichment (Section 4.9).

Region	**Enrichment**
North America	3.13
Europe	3.44
Asia	2.76
South America	2.60
Africa	3.25
Oceania	3.80

Note. Bounding boxes defined in the project configuration [45]; the box for North America runs lat $\in [15, 75]$, lon $\in [-160, -50]$; Europe lat $\in [35, 72]$, lon $\in [-12, 45]$; analogously for the other four regions.

The rotary-wing counterfactual regression (Model 3B in Table 4; $N =$ 558,466) returns a topology that is distinct from fixed-wing. The track-angle main effect is 14.0× ($p < 10^{-300}$, 95% CI [12.55, 15.63]), and the phase × DH interactions are heavily *sub-multiplicative*: Climb × DH OR = 0.23, Descent × DH OR = 0.21, Transition × DH OR = 0.28.

A logistic-regression ceiling artefact is the natural first interpretation: if baseline $P(\text{DV})$ in the rotary-wing DH = 1 stratum approaches unity, multiplicative interaction terms must be sub-multiplicative because they cannot push the predicted probability past one. We rule this out by tabulating the

[4] The boxes are overlapping continental convenience regions defined by median latitude/longitude, not a partition: they exclude open-ocean and polar aircraft-hours and overlap in the Mediterranean, the Gulf/Middle East, and maritime Southeast Asia, so the six regional counts sum to ≈119.0M against the Dataset 3A total of 117,542,731 (+1.3%). The per-region enrichments are reported individually and never pooled, so the overlap does not affect any reported value.

empirical $P(\mathrm{DV})$ across the full phase $\times$ DH factorial (Table A.10; visualized in Figure 2). Empirical $P(\mathrm{DV})$ ranges from 0.071 (Cruise/DH=0) to 0.721 (Climb/DH=1), well below saturation in every cell, which excludes the ceiling-effect explanation independently of any model fit. (Model 3B is a fully saturated specification with eight parameters fit to eight cells; its predicted-vs-empirical agreement is therefore mathematical, not diagnostic, and is shown alongside the empirical column only for transparency.) The sub-multiplicative interactions are accordingly not a mathematical artefact of the logistic link near the probability boundary.

We discuss the mechanistic interpretation in Section 5.3.

Table A.10: Rotary-wing saturation grid: empirical and Model 3B-predicted $P(\mathrm{DV} = 1)$ across all eight cells of the phase $\times$ DH factorial (Section 4.8).

Phase	**DH**	N **cell**	**Empirical** P(**DV**)	**Predicted** P(**DV**)
Cruise	0	105,454	0.071	0.071
Cruise	1	213,840	0.518	0.518
Climb	0	3,569	0.447	0.447
Climb	1	6,263	0.721	0.721
Descent	0	4,283	0.263	0.263
Descent	1	31,304	0.509	0.509
Transition	0	20,561	0.206	0.206
Transition	1	173,192	0.504	0.504

Note. Model 3B is a saturated specification (8 parameters, 8 cells), so the Predicted column matches the Empirical column tautologically; the substantive content is the Empirical column, whose maximum of 0.721 (Climb/DH= 1) is well below the logistic-link ceiling and therefore excludes a saturation-artefact explanation of the sub-multiplicative interactions.

Appendix A.4. Spring 2025 Episode: Drivers and Detection Performance

Appendix A.5. Supplementary Figures

All six figures (Figures 1, 2, 3, 4, 5, and 6) appear in the body of the manuscript adjacent to the results they visualize; none are deferred to this appendix.

Table A.11: Space-weather cross-reference for the Spring 2025 elevation (Section 4.15).

Index	**Elev. mean**	**Non-elev. mean**	**Ratio**	**Welch p**	**r (2025 all)**	**r (within elev.)**
K_p daily mean	2.86	2.48	1.15×	0.015	+0.09 [−0.00, 0.18]	−0.27
K_p daily max	4.11	3.63	1.13×	0.011	+0.10 [0.00, 0.19]	−0.18
a_p daily mean	17.8	15.1	1.18×	0.16	+0.04 [−0.05, 0.14]	−0.21
Sunspot number S_N	109	125	0.87×	0.025	−0.20 [−0.30, −0.09]	**−0.63**
$F_{10.7}$ obs. (s.f.u.)	150.1	155.6	0.96×	0.093	−0.12 [−0.19, −0.04]	**−0.60**

Note. Welch t-test compares the 50-day elevation window with the 315 non-elevation days of 2025. Pearson r is the daily correlation across the same windows; bootstrap 95% CIs over 1,000 iterations. Within-elevation correlations are computed on $n = 50$ daily observations. Bold within-elevation values are significant at $p < 0.001$; the others are not.

Table A.12: Alternative explanations for the multi-channel coupling and the pre-specified analysis that rules each out (Sections 4.4–4.15). Each alternative implies a directional prediction that the data contradict; candidate drivers of the Spring 2025 episode specifically are treated in Section 5.5.

Alternative explanation	**Prediction if it were the cause**	**Test (Sec.)**	**Observed**
Data-link gaps / Δt-conflation inflate the rate-normalized flags	Enrichment highest in Δt-contaminated (LowDQ) hours	DQ tiers (4.5)	Monotone the *opposite* way: 3.156 (PureDQ) $>$ 3.067 $>$ 2.783 (LowDQ)
Threshold-choice artefact	Enrichment weakens as the DVR threshold tightens	Threshold sweep (4.6)	Monotone *rise* $1.56\times \rightarrow 3.59\times$ toward the strictest threshold
Seasonal / geographic heterogeneity	Coupling confined to a season or region	Seasonal (4.15) + geographic (4.9)	Present in all four weeks and all six continental boxes
Airframe-composition artefact	Driven by unclassified / `Other` airframes	Strict fixed-wing (4.7)	Odds ratios within $\approx$1% of the full specification
Receiver-side noise (not the broadcaster)	Coupling attenuates under high integrity / redundancy	NACv/ NIC/ receiver (4.10)	*Strengthens* $3.14\times \rightarrow 3.81\times$ under the strict, perfectly-sampled ($\Delta t = 1\,\mathrm{s}$) intersection
Genuine coordinated manoeuvre	Strongest at terminal altitudes, weak in cruise	Altitude bands (4.14)	*Opposite on the pairwise channel*: enrichment peaks in cruise, $R_{\mathrm{cruise}} = 4.41$ (the three-way V_B peaks at terminal, 4.14)
Random within-airframe pairwise concentration	Observed value lies inside the within-cluster null	Permutation null (4.4)	Outside all 1,000 iterations, $p < 10^{-3}$
Per-hour message-density confound	Density-adjusted DVR odds ratio inflated	Cluster-robust regression (4.11)	Adjusted OR 8.01 $\approx$ unadjusted 8.11; interaction the *opposite* way

Table A.13: Detection performance of the burden alert at daily resolution across calendar-year 2025 (Section 6.4). The monitor is applied to each of the 365 daily windows; the 315 non-episode days are the in-control reference and the 50-day Spring 2025 episode ([2 Apr, 22 May)) is the positive control. $U_\triangle$ is the upper $(1-\alpha)$ empirical envelope of the in-control daily burden $\widehat{P}_\triangle$; ARL_0 is the realized in-control average run length (1/in-control exceedance rate, in daily windows); onset latency is the number of days from 2 April to the first alerting day.

α	$U_\triangle$	In-control false alarms (of 315)	ARL_0 (windows)	Onset latency (days)	Episode days alerting (of 50)
0.05	0.046	16	19.7	0	50
0.01	0.061	4	78.8	0	50
0.001	0.084	1	315.0	0	48

Note. The equivalent Gaussian mean-plus-$k\sigma$ control limit $(k = 1.64, 2.33, 3.09)$ gives $U_\triangle = 0.049, 0.054, 0.059$ with realized in-control $\mathrm{ARL}_0 = 28.6, 39.4, 63.0$ windows respectively; the empirical-quantile calibration tracks the nominal α more closely because the in-control daily burden has a heavier-than-Gaussian right tail (a few non-episode days carry isolated minor excursions). Detection uses the daily burden point estimate; the deployed rule alerts on the cluster-bootstrap lower bound $\widehat{P}_\triangle^-$ and is strictly more conservative, which can only raise ARL_0. Computed from `phase3_4_d1_daily_breakdown.csv` by `phase3_12_monitor_detection.py`.

References

[1] National Academies of Sciences, Engineering, and Medicine, A Review of the Next Generation Air Transportation System, National Academies Press, Washington, D.C., 2015. doi:10.17226/21721.

[2] T. Bolić, P. Ravenhill, SESAR: The past, present, and future of european air traffic management research, Engineering 7 (4) (2021) 448–51. doi:10.1016/j.eng.2020.08.023.

[3] P. Brooker, SESAR and NextGen: Investing in new paradigms, Journal of Navigation 61 (2) (2008) 195–208. doi:10.1017/S0373463307004596.

[4] M. Schäfer, M. Strohmeier, V. Lenders, I. Martinovic, M. Wilhelm, Bringing up OpenSky: A large-scale ADS-B sensor network for research, in: Proceedings of the 13th International Symposium on Information Processing in Sensor Networks (IPSN), 2014, pp. 83–94. doi:10.1109/IPSN.2014.6846743.

[5] J. Sun, H. Vu, J. Ellerbroek, J. M. Hoekstra, pyModeS: Decoding Mode-S surveillance data for open air transportation research, IEEE Transactions on Intelligent Transportation Systems 21 (7) (2020) 2777–2786. doi:10.1109/TITS.2019.2914770.

[6] R. van der Pryt, R. Vincent, A simulation of signal collisions over the north atlantic for a spaceborne ADS-B receiver using Aloha protocol, Positioning 6 (3) (2015) 23–31. doi:10.4236/pos.2015.63003.

[7] M. Strohmeier, V. Lenders, I. Martinovic, On the security of the automatic dependent surveillance-broadcast protocol, IEEE Communications Surveys & Tutorials 17 (2) (2015) 1066–1087. doi:10.1109/COMST.2014.2365951.

[8] D. McCallie, J. Butts, R. Mills, Security analysis of the ADS-B implementation in the Next Generation Air Transportation System, International Journal of Critical Infrastructure Protection 4 (2) (2011) 78–87. doi:10.1016/j.ijcip.2011.06.001.

[9] M. Strohmeier, X. Olive, J. Lübbe, M. Schäfer, V. Lenders, Crowdsourced air traffic data from the OpenSky Network 2019–2020, Earth System Science Data 13 (2) (2021) 357–66. doi:10.5194/essd-13-357-2021.

[10] J. E. Holland, M. J. Kochenderfer, W. A. Olson, Optimizing the next generation collision avoidance system for safe, suitable, and acceptable operational performance, Air Traffic Control Quarterly 21 (3) (2013) 275–97. doi:10.2514/atcq.21.3.275.

[11] M. Strohmeier, I. Martinovic, V. Lenders, Understanding realistic attacks on airborne collision avoidance systems, Journal of Transportation Security 15 (1-2) (2022) 87–118. doi:10.1007/s12198-021-00238-2.

[12] X. Ying, J. Mazer, G. Bernieri, M. Conti, L. Bushnell, R. Poovendran, Detecting ADS-B spoofing attacks using deep neural networks, in: 2019 IEEE Conference on Communications and Network Security (CNS), 2019, pp. 187–95. doi:10.1109/CNS.2019.8802732.

[13] L. Li, R. J. Hansman, R. Palacios, R. Welsch, Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring, Transportation Research Part C: Emerging Technologies 64 (2016) 45–57. doi:10.1016/j.trc.2016.01.007.
URL https://www.sciencedirect.com/science/article/pii/S0968090X16000188

[14] M. Leonardi, G. Sirbu, ADS-B crowd-sensor network and two-step Kalman filter for GNSS and ADS-B cyber-attack detection, Sensors 21 (15) (2021) 4992. doi:10.3390/s21154992.

[15] E. Pik, M. Berra, J. Yearwood, J. S. D. Garcia, Detecting GPS anomalies in aviation using ADS-B: Correlating coordinate gaps and GPS deviations with NOTAM warnings, in: AIAA AVIATION FORUM AND ASCEND 2024, 2024, pp. AIAA 2024–4640. doi:10.2514/6.2024-4640.

[16] E. Pik, M. Berra, J. Yearwood, J. S. D. Garcia, GPS anomalies in aviation: Preliminary insights from automatic dependent surveillance–broadcast data, Journal of Aerospace Information Systems 22 (7) (2025) 570–82. doi:10.2514/1.I011527.

[17] E. Pik, J. S. D. Garcia, T. A. Smith, I. Kocaman, M. Berra, Predicting geopolitical instability through GNSS anomalies and air traffic data, Journal of Transportation Security 18 (1) (2025) 29. doi:10.1007/s12198-025-00323-w.

[18] J. Blanch, T. Walter, P. Enge, Y. Lee, B. Pervan, M. Rippl, A. Spletter, V. Kropp, Baseline advanced RAIM user algorithm and possible improvements, IEEE Transactions on Aerospace and Electronic Systems 51 (1) (2015) 713–732. doi:10.1109/TAES.2014.130739.

[19] Radio Technical Commission for Aeronautics (RTCA), Minimum operational performance standards for 1090 mhz extended squitter automatic dependent surveillance–broadcast (ads-b) and traffic information services–broadcast (tis-b), Tech. Rep. DO-260B, RTCA, Inc., Washington, DC, USA (Dec. 2009).
URL https://www.rtca.org/products/do-260b-hard-copy/

[20] Z. Zeng, A. Barros, D. Coit, Dependent failure behavior modeling for risk and reliability: A systematic and critical literature review, Reliability Engineering & System Safety 239 (2023) 109515. doi:10.1016/j.ress.2023.109515.
URL https://doi.org/10.1016/j.ress.2023.109515

[21] Y. Yu, B. Shuai, W. Huang, Resilience evaluation of train control onboard system considering common cause failure: Based on a beta-factor and continuous-time Bayesian network model, Reliability Engineering & System Safety 246 (2024) 110088. doi:10.1016/j.ress.2024.110088.
URL https://doi.org/10.1016/j.ress.2024.110088

[22] M. Torabi, M. Čepin, An extension to the Alpha Factor method for enhanced common cause failure analysis, Reliability Engineering & System Safety 267 (Part B) (2025) 111964, available online November 2025. doi:10.1016/j.ress.2025.111964.
URL https://doi.org/10.1016/j.ress.2025.111964

[23] G. Fang, R. Pan, Y. Hong, Copula-based reliability analysis of degrading systems with dependent failures, Reliability Engineering & System Safety 193 (2020) 106618. doi:10.1016/j.ress.2019.106618.
URL https://doi.org/10.1016/j.ress.2019.106618

[24] E. Smart, D. Brown, J. Denman, A two-phase method of detecting abnormalities in aircraft flight data and ranking their impact on individual flights, IEEE Transactions on Intelligent Transportation Systems 13 (3) (2012) 1253–1265. doi:10.1109/TITS.2012.2188391.

[25] N. Çevik, S. Akleylek, Comparison of machine learning based anomaly detection methods for ADS-B system, in: Information Technologies and Their Applications. ITA 2024, Vol. 1092 of Lecture Notes in Networks and Systems, Springer, Cham, 2025, pp. 275–86. doi:10.1007/978-3-031-73420-5_23.

[26] A. Tabassum, W. Semke, UAT ADS-B data anomalies and the effect of flight parameters on dropout occurrences, Data 3 (2) (2018) 19. doi:10.3390/data3020019.

[27] X. Olive, J. Sun, B. Figuet, R. Alligier, Filtering techniques for ADS-B trajectory preprocessing, Journal of Open Aviation Science 2 (2) (2025). doi:10.59490/joas.2024.7882.

[28] M. Felux, P. Fol, B. Figuet, M. Waltert, X. Olive, Impacts of global navigation satellite system jamming on aviation, Navigation 71 (3) (2024). doi:10.33012/navi.657.

[29] M. Pirolley, R. Couturier, M. Salomon, F. Ambert, Machine learning techniques for detecting anomalies in ads-b data in the context of low altitude aviation, Aerospace Science and Technology 168 (2026) 111199. doi:10.1016/j.ast.2025.111199.
URL https://www.sciencedirect.com/science/article/pii/S1270963825012623

[30] Z. Liu, S. Lo, J. Blanch, T. Walter, GNSS spoofing detection and localization using ADS-B data, in: Proceedings of the 37th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2024), 2024, pp. 796–803. doi:10.33012/2024.19749.

[31] B. S. Ali, W. Y. Ochieng, R. Zainudin, An analysis and model for Automatic Dependent Surveillance Broadcast (ADS-B) continuity, GPS Solutions 21 (4) (2017) 1373–1386. doi:10.1007/s10291-017-0657-y.

[32] Radio Technical Commission for Aeronautics (RTCA), Minimum operational performance standards for 1090 mhz extended squitter automatic dependent surveillance–broadcast (ads-b) and traffic information services–broadcast (tis-b), Tech. Rep. DO-260C, RTCA, Inc., Washington, DC, USA (Dec. 2020).
URL https://www.rtca.org/products/do-260c-electronic/

[33] X. Olive, traffic, a toolbox for processing and analysing air traffic data, Journal of Open Source Software 4 (39) (2019) 1518. doi:10.21105/joss.01518.

[34] J. Budroweit, F. Eichstaedt, T. Delovski, Aircraft surveillance from space: The future of air traffic control?: Space-based ADS-B, status, challenges and opportunities, IEEE Microwave Magazine 25 (12) (2024) 68–76. doi:10.1109/MMM.2024.3443754.

[35] A. C. Cameron, P. K. Trivedi, Regression Analysis of Count Data, 2nd Edition, Cambridge University Press, Cambridge, UK, 2013.

[36] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, C. E. McCulloch, Regression Methods in Biostatistics: Linear, Logistic, Survival, and Mixed Models, 2nd Edition, Springer New York, 2012. doi:10.1007/978-1-4614-1353-0.

[37] D. R. Cox, Regression models and life-tables, Journal of the Royal Statistical Society: Series B (Methodological) 34 (2) (1972) 187–202. doi:10.1111/j.2517-6161.1972.tb00899.x.
URL https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[38] R. L. Prentice, R. Pyke, Logistic disease incidence models and case-control studies, Biometrika 66 (3) (1979) 403–411.

[39] F. Y. Hsieh, D. A. Bloch, M. D. Larsen, A simple method of sample size calculation for linear and logistic regression, Statistics in Medicine 17 (14) (1998) 1623–1630.

[40] International Civil Aviation Organization, Annex 10 to the Convention on International Civil Aviation: Aeronautical Telecommunications, Volume III — Communication Systems, Montréal, Canada, second edition Edition, includes up to Amendment 93 (issued August 4, 2025). Part I, Chapter 9: Aircraft Addressing System; including the Appendix to Chapter 9 (Global Plan for the allocation, assignment and application of aircraft addresses). (2007).
URL https://store.icao.int/en/annex-10-aeronautical-telecommunications-volume-iii-communication-systems

[41] Federal Aviation Administration, Releasable aircraft database, https://registry.faa.gov/database/ReleasableAircraft.zip, accessed: 2026-05-10 (2025).

[42] Transport Canada, Canadian civil aircraft register computer system (ccarcs), https://wwwapps.tc.gc.ca/Saf-Sec-Sur/2/ccarcs-riacc/RptDnld.aspx, accessed: 2026-05-10 (2025).

[43] R. Gale, ICAOList: Aircraft type designators database, https://github.com/rikgale/ICAOList, accessed: 2026-05-10 (2025).

[44] X. Olive, contributors, Extend aircraft dataset with ICAO Aircraft Type Designator info (issue #176), https://github.com/xoolive/traffic/issues/176, gitHub Issue in the xoolive/traffic repository. Accessed: 2026-05-10. (2023).

[45] E. Pik, Multi-dimensional coupling of kinematic anomalies in ads-b surveillance telemetry, Dataset (Jun. 2026). doi:10.5281/zenodo.20387308.
URL https://doi.org/10.5281/zenodo.20387308

[46] Federal Aviation Administration, Limit Maneuvering Load Factors, U.S. Department of Transportation, title 14 Code of Federal Regulations Part 25, Section 25.337: Airworthiness Standards: Transport Category Airplanes – Limit maneuvering load factors (2014).
URL https://www.ecfr.gov/current/title-14/chapter-I/subchapter-C/part-25/subpart-C/section-25.337

[47] Federal Aviation Administration, Instrument Flying Handbook, U.S. Department of Transportation, Flight Standards Service, fAA-H-8083-15B. Chapter 5: Flight Instruments – Standard rate of turn (2012).
URL https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/instrument_flying_handbook

[48] C. F. F. Karney, Algorithms for geodesics, Journal of Geodesy 87 (1) (2013) 43–55. doi:10.1007/s00190-012-0578-z.

[49] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, Journal of the Royal Statistical Society: Series B (Methodological) 57 (1) (1995) 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

[50] NOAA National Centers for Environmental Information, Storm events database, calendar year 2025, bulk CSV downloads, accessed May 2026 (2025).
URL https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/

[51] J. Matzka, C. Stolle, Y. Yamazaki, O. Bronkalla, A. Morschhauser, The geomagnetic Kp index and derived indices of geomagnetic activity, Space Weather 19 (5) (2021). doi:10.1029/2020SW002641.

[52] J. Matzka, O. Bronkalla, K. Tornow, K. Elger, C. Stolle, Geomagnetic Kp index, daily values of Kp, ap, ap, sunspot number SN and solar radio flux F10.7 since 1932, daily values archive Kp_ap_Ap_SN_F107_since_1932.txt, accessed May 2026 (2025). doi:10.5880/Kp.0001.
URL https://www-app3.gfz-potsdam.de/kp_index/

[53] EUROCAE, ED-102B: Minimum Operational Performance Standards for 1090 MHz Extended Squitter Automatic Dependent Surveillance–Broadcast (ADS-B) and Traffic Information Services–Broadcast (TIS-B), Saint-Denis, France (2020).

[54] International Civil Aviation Organization, Aeronautical Surveillance Manual (Doc 9924), Montréal, Canada, 2nd Edition (2020).

[55] RTCA, Inc., DO-178C: Software Considerations in Airborne Systems and Equipment Certification, Washington, DC, USA (2011).

[56] SAE International, ARP4754A: Guidelines for Development of Civil Aircraft and Systems, Warrendale, PA, USA (2010).

[57] H. Rong, A. P. Teixeira, C. Guedes Soares, A framework for ship abnormal behaviour detection and classification using AIS data, Reliability Engineering & System Safety 247 (2024) 110105. doi:10.1016/j.ress.2024.110105.
URL https://doi.org/10.1016/j.ress.2024.110105

[58] T. Westbrook, A taxonomy of radiofrequency jamming and spoofing strategies and criminal motives, Journal of Strategic Security 16 (2) (2023) 68–80. doi:10.5038/1944-0472.16.2.2081.

[59] Federal Aviation Administration, GPS and GNSS interference resource guide: Jamming and/or spoofing, Resource guide, Federal Aviation Administration, Flight Technologies and Procedures Division (AFS-400) (Mar. 2026).
URL https://www.faa.gov/about/office_org/headquarters_offices/avs/offices/afx/afs/afs400/afs410/GNSS/GPS_GNSS_Interference_Resource_Guide.pdf