

Calibrating plateau detectors in ADS-B data without ground truth

Gabriel Jarry ,* Ramon Dalmau , and Philippe Very

EUROCONTROL, Brétigny-sur-Orge, France

*Corresponding author: gabriel.jarry@eurocontrol.int

Abstract

Plateau detectors find the stretches of a flight signal that stay approximately constant and return each as one announced value, to provide inputs for trajectory models. Their tolerances are usually chosen by judgement, because the commanded values are unavailable for most surveillance channels and only partially observed for altitude. We propose a reference-free calibration procedure for five signals — altitude, vertical speed, flight-path angle, calibrated airspeed and Mach — deriving each one's error budget from its own noise floor: the broadcast quantisation step, the dispersion of the reanalysis reconstruction, or the propagation of both. The procedure then retains the setting committing the largest share of the trajectory whose every segment stays within that budget.

We evaluate four detector families on 5463 flights from 200 airframes observed on four dates over one year. Applied unchanged to the three dates excluded from calibration, the selected settings hold their coverage; the median segment error stays near a fifth of its budget. The worst segment is another matter: it leaves the budget on thirteen of fifteen held-out channel-dates. Across all four dates, the segments that exceed it number 31 in 139 964. An independent comparison with broadcast selected altitude finds agreement within 100 ft for 92.4 % of matched altitude plateaux.

The procedure makes the coverage-error trade-off explicit and reproducible without command labels. Its budgets constrain signal reconstruction; interpreting the detected values as commands requires independent evidence.

Keywords: Mode S; Comm-B; plateau detection; reference-free calibration; transfer cost

1. Introduction

Aircraft motion is naturally an ordinary differential equation: a state x evolves as $\dot{x} = f(x, u)$ under the command u the autoflight system is holding. A Neural ODE [1] learns f from recorded trajectories and integrates it forward to generate new ones; its controlled variant [2] does the same for a system driven by an external input, an approach we have used to model flight dynamics and to generate vertical profiles from open data [3, 4]. Such a model needs $u(t)$ alongside $x(t)$: to learn a response, it must see the command that caused it. This paper produces that input; the model consuming it is not evaluated here.

Those commands are the target values the autoflight system holds — an altitude, a vertical speed, a calibrated airspeed, a Mach number. Open surveillance data shows their consequence rather than the values themselves: stretches of a signal that stay quasi-constant. Detecting those stretches is how the values are recovered in practice, routinely used to manufacture the inputs trajectory models consume [4].

The word *selected* suggests a crew dialling a value. That is usually not what happens. In *selected* guidance the crew dials a value into the flight control unit. In *managed* guidance, the default in modern operations, the flight management system computes the target from the active flight plan and passes it to the autoflight system with no crew input at that moment. Identifying which mode

is active and estimating the target under it is its own line of work [5, 6, 7]. The distinction leaves the detection problem unchanged, since a held target leaves the same plateau whichever set it. It governs how a detected plateau may be described, and bears directly on how much of a trajectory holds a plateau at all. We write *commanded value* throughout and reserve any claim about who commanded it.

Every such detector has a setting that decides how flat a stretch must be to count; that setting has to come from somewhere. In a supervised problem it would be chosen against labels. Here the labels are partial, and the shape of that partiality is what makes the problem interesting. Mode S Enhanced Surveillance specifies a register carrying the selected altitude; over the cohort studied here it is populated on 30.4 % of samples. That is enough to compare a detector against on one channel and too little to calibrate one on. No equivalent field exists for vertical speed, flight-path angle, calibrated airspeed or Mach. For four channels of five the question of a reference does not arise: there is none. The setting is therefore chosen by judgement, a choice rarely reported and never costed. In a recent article in this journal the tolerance separating a level segment from a climb is described in its authors' own words as "chosen through trial and error" [8].

A reference is not the only thing a detector can be calibrated against. The signal carries its own error floor: a broadcast channel is quantised at a step the data reveals, while a channel reconstructed from a reanalysis carries the dispersion of that reconstruction. Reading a tolerance off those floors replaces a judgement with a measurement: the detector setting becomes reproducible without becoming arbitrary. What such a procedure owes its user is an account of what it costs, because a setting calibrated on one population is routinely applied to another.

This paper reports four quantities and establishes no hypothesis. First, the error budget each of five channels imposes on itself, derived from quantisation, from reanalysis reconstruction error and from propagation, and measured independently on four dates a season apart. Second, the operating point a declared rule retains on each channel out of 27 414 candidates over four detector families. Third, the price of freezing those points, measured across aircraft subsets within the calibration cohort and on three dates the selection never read. Fourth, on the one channel where a broadcast selected altitude exists, how the detector and that register relate as two observers of the same quantity.

The remainder of this paper is organised as follows. Section 2 reviews the lines of work this problem sits between. Section 3 describes the cohort, the registers behind each channel and the provenance that bounds the speed results. Section 4 states the selection rule and the protocol. Section 5 reports the four measurements above. Section 6 states what the procedure does and does not license.

2. Literature review

Four lines of work bear on this problem: level-segment detection on trajectory data, set-point inference from surveillance, change-point detection without labels, and the methodology of choosing a parameter when no labels exist. Each supplies one ingredient.

Detecting level segments in open surveillance data is established practice. The tolerance that decides one is consistently set by hand. Sun, Ellerbroek, and Hoekstra [9] identify flight phases from ADS-B with a fuzzy-logic scheme whose membership functions carry fixed vertical-rate and altitude thresholds; it is shipped in OpenAP [10] and reached through traffic [11], the de-facto standard for this task. Perrichon, Gendre, and Klein [8] compare a hidden Markov model against a naive level-detection baseline whose tolerance, in their words, "is chosen through trial and error". They state the reason plainly: ADS-B data provide no ground truth for a segmentation, so they escape to flight-data recordings to obtain labels at all. EUROCONTROL's Open Performance Data Initiative detects flight milestones over the same network with published thresholds. These methods emit

a phase *label*; the procedure reported here emits a committed *value* with a coverage and an error against it. There is no metric both sides report, so no benchmark against them is offered. On four channels of five no prior detector outputs a value at all.

On the aviation side, Mode-S Enhanced Surveillance mixes kinematic measurements with fields describing selected targets, and BDS 4,0 carries status-qualified MCP/FCU and FMS selected altitudes [12, 13]. Availability in the protocol does not ensure availability in a research product, particularly over a crowdsourced network [14, 15]. The closest prior art runs banks of interacting multiple-model filters to identify the active guidance mode [5, 6], most recently dropping the assumption that the set-point is known and estimating it jointly with the mode [7]. That line is model-based: it posits dynamic models and noises and estimates set-points by online Bayesian filtering. Its own calibration burden sits in process and measurement covariances that are tuned rather than measured. It answers which mode is active; the question here is how to fix a detector's tolerance when no labels are available to fix it against.

Within change-point detection, the CLaSP and CLaP family places change points without annotations by optimising a self-classification score [16, 17]. Count penalties are standard [18, 19]: each optimises a single scalar, and the penalty's influence is documented, weak penalties retaining noise-driven changes while strong ones suppress all but the largest. Optimising one objective while holding another below a fixed bound is the ε -constraint formulation of a multi-objective problem [20, 21]. What matters is where the bound comes from: a defensible procedure fixes it outside the sweep. Here it is fixed by the signal's own error floor rather than by judgement, which is the difference this paper reports.

Two constraints come from outside aviation. Hennig [22] states the one this work is built around: a low reconstruction error establishes that a constant segment summarises an observed signal; independent information is required before that segment may be called a command. Repeated samples from one aircraft describe it in more detail without creating independent population units [23], so every evaluation below groups by airframe.

The choice is not ours alone. A recent speed-profile model trained on 616 336 ADS-B flights builds its targets by smoothing the profile and cutting it where the gradient against altitude falls under an empirically identified 2 kt per 1000 ft, and names a finer changepoint labelling as future work [24].

The most directly comparable detector is our own. The vertical-profile model this work feeds obtained its speed and vertical-rate targets from the same family of detectors, with quasi-steady intervals identified when a smoothed first derivative stayed below a fixed tolerance for a minimum duration [4]. The family is the one swept here; what differs is everything around the tolerance. It was fixed rather than derived, so no coverage was reported and no error was measured against the value each segment announces. Those are defensible engineering choices, priced here.

3. Data

3.1 Cohort

The study uses surveillance data from the OpenSky Network [14], retrieved through its Trino interface from the history, extended and flightlist tables. Mode S Comm-B replies were decoded with rs1090, through traffic [11]. The register census of Section 5.1 was recomputed independently with pyModeS [25], which shares no code with that path: an independent decoder is what makes the count evidence about the decoding rather than about itself.

The cohort is fixed at 200 airframes drawn from ten International Civil Aviation Organization (ICAO) typecodes (A320, B738, A20N, B38M, A21N, A321, A333, A359, AT76 and E190), twenty airframes for

Table 1. The cohort, per date. The same 200 airframes were followed on four UTC dates; an aircraft absent from a date did not fly that day. The last two columns give the share of resampled samples carrying a broadcast selected altitude in each of the two BDS 4,0 fields, the mode control panel (MCP) and the flight management system (FMS). The calibration date is 1 November 2025.

Date	Flights	Airframes	Samples	MCP (%)	FMS (%)
17 Feb 2025	1 168	161	858 670	32.5	14.6
21 May 2025	1 354	176	993 980	31.2	12.5
29 Aug 2025	1 434	179	1 093 076	28.8	11.0
1 Nov 2025	1 507	200	1 108 921	29.6	10.6
Cohort	5 463	200	4 054 647	—	—

each type. These same 200 airframes were then observed on four UTC dates: 17 February, 21 May, 29 August and 1 November 2025. The last is the calibration date on which every operating point in this paper was chosen; the other three were drawn from windows spread across the year and entered no choice of setting, budget or grid, which is what makes Section 5.4 a measurement of transport rather than of fit.

The fleet is fixed across dates by design. Were the airframes redrawn per date, a difference between two days would confound seasonal meteorology with fleet composition. The selected dates span the year and include weekdays and a Saturday.

The resulting cohort holds 5463 flights and 4 054 647 samples. Aircraft availability is not uniform across dates: an airframe that did not fly on a given day is simply absent from it, so the four dates carry 161, 176, 179 and 200 airframes respectively. Table 1 reports the composition and the availability of the broadcast register alongside it.

3.2 Channels and their provenance

The five channels come from three different provenances. Barometric altitude and barometric vertical speed are broadcast directly and arrive quantised at the encoding step of their register. Calibrated airspeed and Mach number are reconstructed for about 70 % of the samples: where the aircraft does not broadcast them, they are derived from ground velocity combined with the ERA5 reanalysis wind field, interpolated onto the trajectory with fastmeteo [26]. The flight-path angle is derived from vertical rate and true airspeed through $\gamma = \arcsin(v_z/TAS)$, and therefore inherits the error of both.

That separation governs everything downstream. A quantised channel carries an error floor set by its encoding; a reconstructed one carries the error of the reanalysis it was rebuilt from; a derived one carries a propagated combination. Section 5 returns to each in turn.

All trajectories were resampled onto a uniform 4 s grid. Samples were retained only where the flight was airborne and the channel finite. Two further exclusions apply. Runs of at least fifteen samples at either end of a flight over which five reference fields — altitude, ground speed, track and both coordinates — all repeat are treated as frozen and removed, since a stalled feed reports a perfect plateau. Every coverage figure below is a share of what survives those exclusions.

3.3 The broadcast selected altitude

Comm-B register BDS 4,0 carries the selected altitude the crew has dialled, in two distinct fields: the mode control panel or flight control unit (MCP/FCU) selection, and the flight management system (FMS) selection. It accounts for 29.51 % of the decoded Comm-B replies in this cohort; Comm-B itself accounts for 70.28 % of the Mode S stream.

Availability is uneven along every axis that matters. Of the 200 airframes, 170 emit the register at all and 30 never do. Per-date coverage of the MCP field ranges from 28.8 % to 32.5 % of resampled samples; the FMS field reaches 12.0 %. Reception itself depends on receiver geometry and interrogation rate, so what is absent is not a random subset of what was flown; that structural property bounds every comparison in Section 5.5.

No register in the Mode S stream carries a selected speed. Whatever exists as a reference exists for altitude alone.

4. Method

4.1 What a plateau commits to

A detector applied to one channel returns a set of segments, each announcing a single value: the mean of the channel over the segment's own extent. Two quantities follow from that definition and are used throughout.

The *committed coverage* of a setting is the share of finite samples that fall inside a segment it returns. The *error* of a segment is the mean absolute deviation of its samples from the value the segment itself announces,

$$e(s) = \frac{1}{|s|} \sum_{t \in s} |v_t - \bar{v}_s|, \quad (1)$$

where $\bar{v}_s$ is that announced value. No external reference enters Equation 1, which is what makes it computable on four channels where nothing is broadcast to compare against. Its limitation follows from the same property: a segment whose announced value is wrong in a way its own samples do not reveal is invisible to it.

4.2 The selection rule

Each channel receives one operating point. A candidate is admissible when its worst committed segment respects the channel's error budget E_c : the constraint binds the maximum over every segment of every flight. Among the admissible candidates the rule retains the one committing the most coverage. Candidates within 0.05 coverage points of the best are treated as tied; the tie is broken towards the lower error and then the fewer segments per flight.

The budgets E_c are measured rather than chosen. Section 5.2 reports that measurement as a result. Each channel is swept on the samples its predecessors in the cascade have not claimed, in the order Section 4.4 sets out.

4.3 The candidate grid

Four detector families were swept on every channel, with family-specific segmentation and run validation but a common final raw-signal error criterion. The bilateral filter smooths while preserving edges, swept over its range parameter σ_r (6 to 9 values per channel), its spatial width σ_s (4 to 8) and the slope tolerance below which a sample counts as flat (24 to 26 values on a geometric ladder). Savitzky–Golay replaces that smoother by a polynomial low-pass over a window (10 widths) at the same slope tolerances. The variance threshold is the baseline with no smoothing at all: a sample is flat where its rolling standard deviation falls below a tolerance (12 values). PELT places change points by penalised segmentation [18], swept over the penalty (14 values), with candidate change points spaced 40 samples (160 s) apart. Each family was also swept over four low-pass pre-filter cut-offs, for 27 414 scored candidates in total; the grid itself is in the repository. Section 5.3 reports what each family reached.

The bilateral filter carries the three parameters Table 3 reports. Its range scale σ_r is expressed in the channel's own unit and sets how far two samples may differ in value while still smoothing one another; its spatial Gaussian standard deviation σ_s is expressed in samples, with kernel half-width $[3\sigma_s]$, on the 4 s grid; and the slope tolerance, in the channel's unit per sample, is the derivative below which the smoothed signal counts as holding.

Two further constants govern candidate runs in the main sweep. A run is rejected unless it spans a minimum number of samples, from 5 on calibrated airspeed to 15 on Mach. This gate is applied before clipping to finite observations; clipping can shorten the final segment. Bilateral and Savitzky–Golay runs also undergo a range check on the smoothed signal; PELT applies that check to its interpolated input. The variance baseline uses its rolling standard-deviation gate. Where applied, the range limit is twice a fixed per-channel tolerance. These constants are hand-set; Section 6.1 returns to their sensitivity.

4.4 Cascade ordering

The five channels run as two sequential cascades: altitude, then flight-path angle, then vertical speed; and Mach, then calibrated airspeed. Within a cascade each channel is scored on the residual its predecessors leave, and its retained point is fixed before the next is swept. Figure 1 shows both on one flight, at the settings Section 5.3 retains.

This ordering is declared here as part of the protocol. A plateau in flight-path angle over a stretch that altitude has already committed would be tuned on material the detector never meets at use time, which is what the residual removes. Scoring every combination of points jointly would search the product of a cascade's grids: 1.6×10^{11} triples on the vertical side alone, against the 17 142 candidates the ordering scores there. It was not attempted. Section 5.3 therefore reports the outcome of this ordering; a different ordering would retain different points.

5. Results

5.1 What the broadcast register can and cannot calibrate

A broadcast selected altitude is present on 30.4 % of the cohort's resampled samples and on 170 of its 200 airframes. That is enough material to compare a detector against on one channel and too little to calibrate one on: the share is intermittent within a flight, its absence follows receiver geometry rather than chance, and a register that decodes to a plausible altitude is not thereby a valid record of what the crew selected.

The binding shortage is elsewhere. No Mode S register carries a selected speed, so on Mach and calibrated airspeed there is nothing to score a detector against. The flight-path angle is derived rather than commanded. Four channels of five therefore admit no reference at all, whatever the fifth offers. A method that needs one is unavailable to them, which is the constraint the rest of this paper works under.

5.2 The budget each channel imposes on itself

The tolerance a plateau must respect is measured here rather than chosen. Table 2 reports the five budgets and the property each is read from.

The two broadcast channels are quantised, and the step is recovered from the signal: 99.5 % of consecutive altitude differences are multiples of 25 ft, while 70.7 % of vertical-rate differences are multiples of 64 ft/min. We use the full step as the calibration budget. The two speed channels are reconstructed from the ERA5 wind field on about 70 % of their samples; their budget is the robust standard deviation of the reconstruction error, measured on the 29.5 % of samples where a broadcast value and a

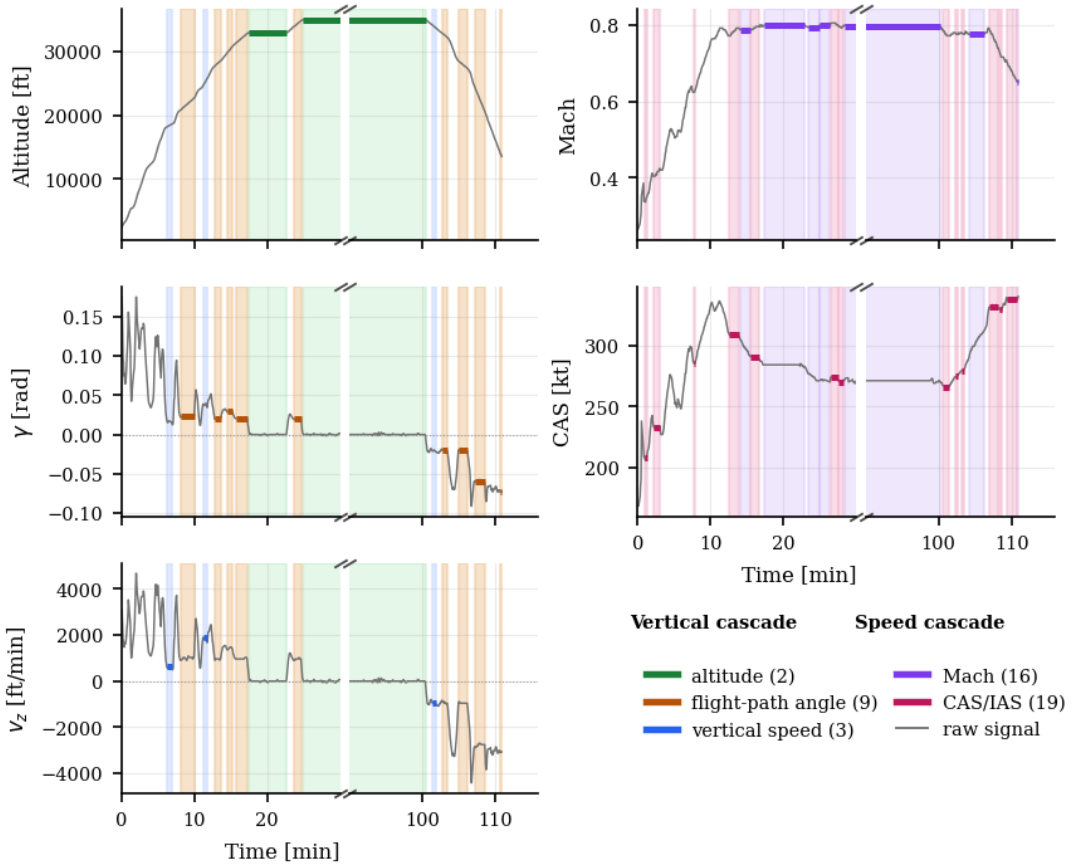


Figure 1. Both cascades on one flight. On the left, altitude commits the cruise and beneath it the flight-path angle and vertical speed fall silent: those stretches are the consequence of the altitude hold rather than independent commands, and the cascade stops all three claiming them. On the right, Mach holds through the cruise while the calibrated airspeed beneath it decays steadily; airspeed commands the climb and the descent where the Mach slopes. A sloping trace is therefore not a missed detection when the other channel is the one being held: each hold appears on both panels, as the held value on one and as its computed consequence on the other. Shading marks which channel owns each stretch of time, and the counts are runs of committed samples rather than commands. The axis is broken between 30 and 90 minutes: that hour holds one altitude and one Mach, and eliding it gives the climb and the descent, where both cascades arbitrate, the width they need. The counts are the flight's, so they exceed what the elided figure draws.

reconstruction coexist, though it is applied on the other 70.5 %. The flight-path angle is derived; its budget propagates the other two through $\gamma = \arcsin(v_z/TAS)$, the vertical-rate term dominating by a factor of ninety.

Two qualifications travel with those numbers. A floor is a lower bound on achievable error; treating it as a budget is a decision, made here and stated as one. The reconstruction also carries a systematic offset as well as a spread: ERA5 runs about 2.4 kt fast on calibrated airspeed, 92 % of that channel's budget, and only the dispersion enters the tolerance.

A natural intuition is that the reconstruction error tracks wind strength. It does not: the correlation between wind magnitude and reconstruction error is $r = -0.036$ on calibrated airspeed and $r = -0.031$ on Mach.

Deriving the budgets on a single day would leave them a property of that day. Measured indepen-

Table 2. The five error budgets and the signal property each is derived from. The right-hand column gives the ratio between the largest and smallest budget obtained when the derivation is repeated independently on each of the four dates.

Channel	Derived from	Budget E	Ratio across dates
Altitude	full quantisation step	25 ft (7.62 m)	1.000
Flight-path angle	propagation through $\arcsin(v_z/TAS)$	0.082 deg (0.0014288 rad)	1.029
Vertical speed	full quantisation step	64 ft/min	1.000
Mach	robust σ of the ERA5 reconstruction error	0.0067	1.271
CAS/IAS	robust σ of the ERA5 reconstruction error	2.57 kt	1.254

dently on all four dates, the two quantised budgets come back identical every time (ratio 1.000) and only the reconstructed ones move, by 1.254 on calibrated airspeed and 1.271 on Mach. A budget-bearing quantity that is stable does not thereby make the budget adequate; it establishes that the budget is not the part that varies.

5.3 What the rule retains

Applied to 27 414 candidates over four detector families, the rule retained one operating point per channel; the bilateral filter with no pre-filtering took all five. Table 3 gives the settings and their coverage; Table 4 gives what each family reached.

Committed coverage totals 73.9 % over the vertical cascade and 85.2 % over the speed cascade. Those totals span a wide range: on the vertical side, altitude commits 62.1 % of the channel, the flight-path angle 6.9 % and vertical speed 4.9 %, the latter two in segments of a median 60 and 40 s.

The spread is partly the ordering and partly the data. Sweeping the entire grid with the budget switched off, the highest coverage any candidate reaches is 99.9 % on altitude, 21.4 % on the flight-path angle and 16.9 % on vertical speed. The last two are swept on the residual altitude leaves, which bounds them before any tolerance applies; the budget then cuts 16.9 % to 4.9 %. These ceilings describe the coverage reached by the searched detectors on the cascade residual. Calibrated airspeed inverts the pattern: its ceiling of 43.0 % sits below the 68.7 % of the trajectory the cascade leaves it, so there the swept grid bounds the retained 32.4 %.

What vertical speed does commit, it commits within tolerance: a median segment error of 0.47 of its budget, with 8 segments in 14 809 exceeding it, the highest breach rate of the five channels on the smallest committed share.

The rule binds a plateau’s mean absolute deviation and leaves its shape free, so a committed segment may decline slowly rather than hold a value. At the retained points the per-segment drift is small in the typical case, with medians from 0.00 of budget on altitude to 0.50 on vertical speed. Constraining drift within twice the budget as well costs 6.8 coverage points on altitude and nothing on Mach; the two-constraint rule was not adopted. Both frontiers are reported in the repository.

Two properties qualify the retained points. The vertical points are a greedy joint choice, altitude selected first, the flight-path angle on what it leaves and vertical speed on what both leave. A point sitting at the edge of a swept range was decided by the range rather than by the data.

Four channels of five stand at 95.7 % or more of their budget; only Mach keeps real margin at 70.2 %. That is what an argmax against a binding constraint does, and it governs everything in Section 5.4.

5.4 What freezing the points costs

The retained points were evaluated unchanged across aircraft subsets within the calibration cohort and on dates excluded from calibration. Figure 3 shows both axes.

Table 3. The operating point retained on each channel, its committed coverage, and the share of its budget the worst committed segment uses. The range scale σ_r and the slope tolerance are in the channel's own unit, the latter per sample; σ_s is the spatial Gaussian standard deviation in samples. Only Mach is not pressed against its constraint.

Channel	Retained setting	Coverage (%)	Budget used (%)	Admissible candidates
Altitude	bilateral, no pre-filter, σ_r 5, σ_s 20, slope tol. 3.63	62.1	96.9	260
Flight-path angle	bilateral, no pre-filter, σ_r 0.002, σ_s 4, slope tol. 0.000144	6.9	95.7	347
Vertical speed	bilateral, no pre-filter, σ_r 50, σ_s 4, slope tol. 21.9	4.9	97.7	257
Mach	bilateral, no pre-filter, σ_r 0.0025, σ_s 8, slope tol. 0.000316	52.8	70.2	581
CAS/IAS	bilateral, no pre-filter, σ_r 1.5, σ_s 8, slope tol. 0.353	32.4	97.0	553

Table 4. Candidates scored and candidates inside budget, by detector family and by pre-filter cutoff. All four families produce admissible candidates the retained one had to beat, though PELT clears the budget on two channels of the five.

Detector family	Scored	Admissible	Channels	
			scored	admissible
Bilateral filter	19 872	1 169	5	5
Savitzky-Golay	4 880	430	5	5
Variance threshold	2 390	370	5	5
PELT (change point)	272	29	5	2
Total	27 414	1 998	5	5
Admissible candidates by pre-filter cutoff (s)				
none	6 846	1 673		
120	6 857	219		
180	6 857	101		
300	6 854	5		

On the airframe axis, the cohort was split into five groups of forty airframes. No channel breached its budget on any fold, and the arithmetic makes that unavoidable: a maximum over a partition cannot exceed the maximum over the whole, and the frozen point was selected to satisfy the budget on the full cohort. What the folds do measure is the deficit against each fold's own optimum; it separates the channels by their margin. Calibrated airspeed, pinned at 97.0 % of its budget, has nowhere to move: its median deficit is zero and the rule returns the frozen setting outright on three folds of five. Mach, at 70.2 % of budget, has room: re-selecting on a fold buys back a median 7.58 coverage points, which is the size of the 8.31 points its coverage already spans between folds, so the gain is not separable from ordinary between-fold variation. Altitude's 0.33-point deficit sits far below its own 9.38-point spread. Every fold's aircraft contributed to the frozen point, so this prices carrying a setting to airframes that did not dominate its selection.

On the day axis the picture changes. Carried to three dates the selection never read, the frozen points exceed their budget on 13 of the 15 held-out channel-dates, altitude reaching 253 % of its own. Coverage, meanwhile, barely moves: at most 2.74 points on any channel across those dates. The two halves of the same measurement behave differently. Re-calibrating repairs the breach and costs coverage: on twelve of the fifteen, the point re-selected on the day itself respects that day's budget and commits less than the frozen point, which reaches its own coverage by exceeding it. Two properties qualify the comparison: the four days carry 161, 176, 179 and 200 airframes, so a difference between days is a difference between days and fleets together; and the two speed channels are about 70 % reconstructed, so what transports on them is the detector and the reanalysis jointly.

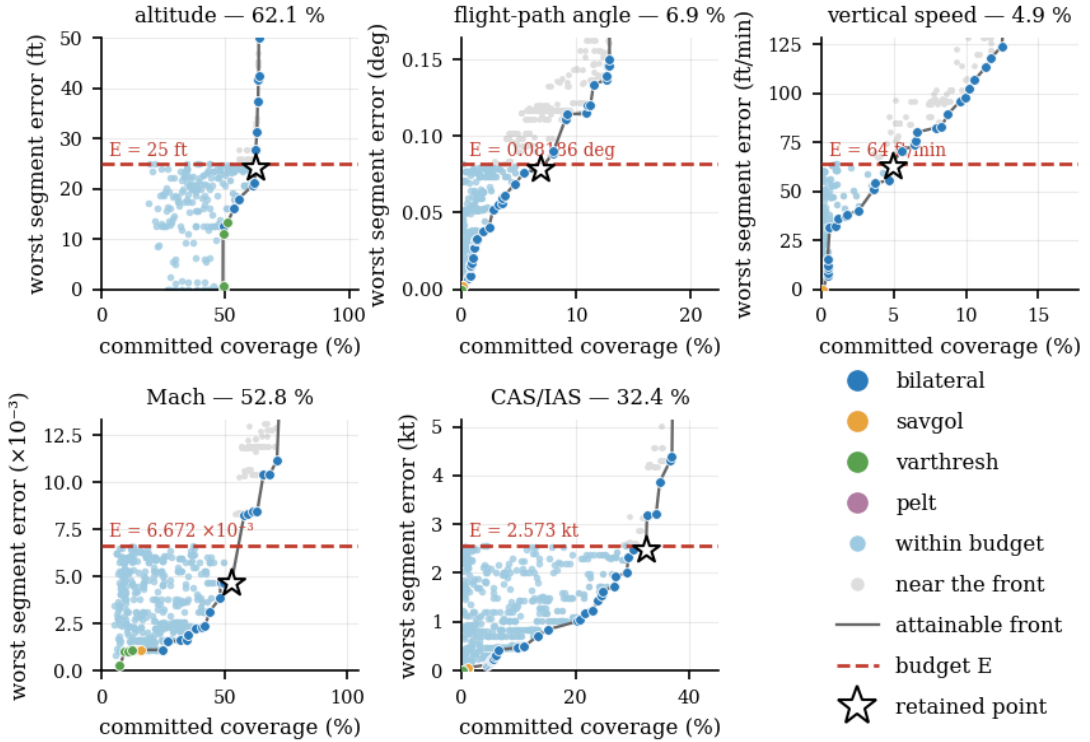


Figure 2. The attainable front on each channel, in that channel’s own unit. The line is the least error reached at or below each level of coverage, its points coloured by the family that reached them; pale points are the other candidates inside budget, and the dashed line is the budget itself. The star is the retained point. It sits on the front on all five channels; on four of the five it sits against the budget, because that is where the front still rises.

Those maxima are one statistic of a distribution; the pooled per-segment medians remain close. Across the four dates 31 of 139 964 committed segments exceed their budget, 0.022 %. The calibration date and the three held-out ones carry median errors of 0.214 and 0.216 budget units, pooled across the five channels (Figure 4). A maximum is the statistic most sensitive to sample size, and the held-out dates carry 102 354 segments against the calibration date’s 37 610; the frozen point was also selected to sit against its budget on that smaller sample. Both push the held-out maximum upward; neither is separated here from the change of date itself. Section 6.1 states what that does and does not license.

5.5 Agreement with the broadcast register

One channel admits an external check. Where the detector commits an altitude plateau and a broadcast selected altitude also exists over its extent, two observers of the same quantity can be compared. The comparison is legitimate only because the register entered no part of the selection: it appears in no budget, no grid and no retained point.

Of the 5290 altitude plateaux the frozen detector commits on the calibration date, 1665 have a broadcast value over their own extent; that share is counted over segments, where every other percentage in this paper is a share of samples. Across those plateaux, the two observers differ by a median 0 ft, with an interquartile range of 7.9 ft; 92.4 % agree within the chosen comparison threshold of 100 ft (Figure 5). BDS 4.0 encodes selected altitude in 16 ft steps [13]. Restricting to plateaux above 1000 ft, which removes aircraft on the ground with a level still dialled, moves that share to 93.4 %.

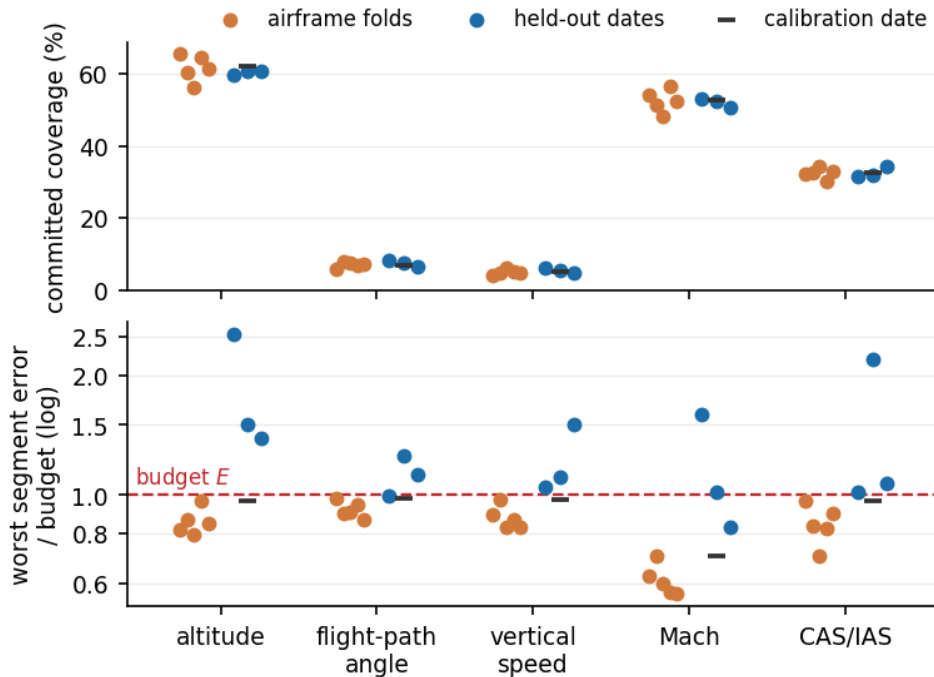


Figure 3. Committed coverage (top) and the worst committed segment's error as a share of budget (bottom, log scale), for five aircraft subsets of the calibration cohort and the three held-out dates, against the calibration date. Coverage transports on both axes; the worst segment leaves the budget on the day axis alone.

That agreement holds on one channel of five, on the 31.5 % of plateaux the register reaches, across 127 airframes of 200, and on a single date. The register also holds a *dialled* value rather than a held one, which a crew may set minutes before the aircraft reaches it or leave in place after levelling elsewhere; a difference between the two observers can therefore arise from the detector, from the register's semantics or from the timing between them. This comparison separates none of the three. The plateaux that disagree by more than one flight level are short and late: a median 100 s against 268 s, three quarters of the way through the flight against half. The broadcast value changes underneath them 46 % of the time against 32 %.

The register's partial presence is what makes this check possible at all. On the four channels where nothing is broadcast, a committed plateau can be scored only against its own samples; here one channel can be scored against an independent observer.

6. Discussion and limitations

6.1 What the rule does not decide

The rule constrains one statistic of a plateau and leaves the rest free. Three consequences follow from that choice.

It binds the mean absolute deviation and leaves the shape free: a committed segment may decline steadily and still satisfy its budget. Constraining the drift as well was priced on the frontier, costing coverage on the channels already pressed against their budget and nothing on the one that is not; the two-constraint rule is not adopted here. A consumer needing segments that hold their value

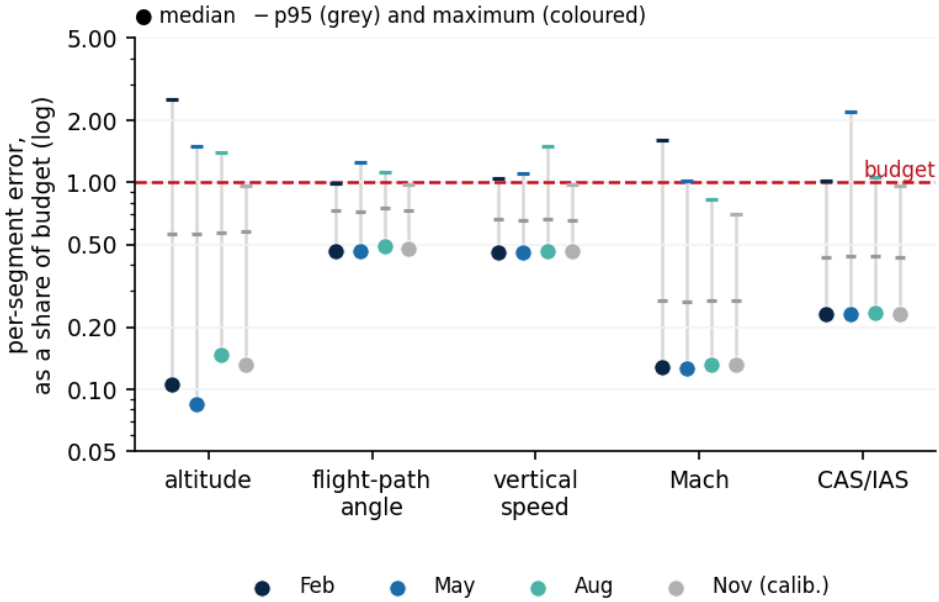


Figure 4. Per-segment error as a share of its channel’s budget, per channel and per date: median, 95th percentile and maximum, on a log axis. The dashed line marks the budget. The medians are flat across the four dates; the maxima are not.

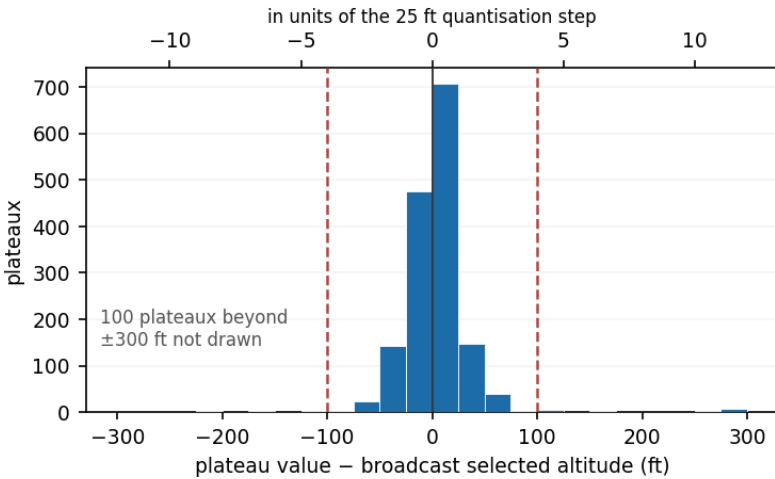


Figure 5. Difference between the value a committed altitude plateau announces and the selected altitude broadcast over its extent, in feet and in units of the 25 ft quantisation step. Dashed lines mark one flight level.

should apply it and accept that cost.

332

It constrains a maximum, and a maximum over many segments is a different object from the distribution beneath it. This is why the retained points sit against their budgets. It also bears on why a handful of segments exceed them on dates the selection never saw: an extreme over a larger sample tends to exceed an extreme over a smaller one, and the held-out dates carry nearly three times the segments. Rare failures make a criterion whose failures are rare, and no more than that. Whether a

333
334
335
336
337

rule constraining a percentile would retain better operating points is not measured here; Section 6.4 says what measuring it would take.

It also disfavours a non-zero pre-filter cutoff structurally: a candidate detects on the filtered view and is scored against the raw signal, so the more it smooths the further its committed constant sits from what was transmitted. That all five channels retain no pre-filtering is the criterion at work as much as a property of the data; the admissible count per cutoff in Table 4 makes it visible.

The rule is a choice about what a good operating point is. An objective favouring longer segments, fewer of them, or better transport would retain different points on the same data, which is why the frontier is published in full.

A sensitivity check on two settings fixed per channel by judgement — how long a run must be, and how tightly it must hold — leaves the retained point unmoved on altitude and CAS/IAS, and shifts Mach and vertical speed by under a point of coverage. The flight-path angle moves more: its tolerance sits at 1.4 times the derived budget, and the channel keeps 6.9 % of coverage under it.

6.2 What the cohort bounds

Every measurement here rests on one crowdsourced receiver network. Reception over it is uneven in ways this design does not control. Mode S Enhanced Surveillance arrives on ground interrogation, so what is observed depends on receiver density and on local interrogation schemes as much as on what was flown. That bounds the register comparison in particular: the plateaux that can be checked against a broadcast value are not a random sample of the plateaux the detector produces.

Four dates of one year say whether a setting survives a change of season. They do not characterise the distribution of what a day can do; four points cannot. The fleet is also not identical across them: an airframe that did not fly on a date is absent from it, so a difference between two days is a difference between days and fleets together. Holding the airframes fixed by design bounds that confound without removing it. The airframe axis measured separately is the closest available estimate of its size. Broadening the cohort in season and in fleet is the remedy, and it requires only data.

The two speed channels are reconstructed from a reanalysis on about 70 % of their samples, so results on them describe a reconstructed signal rather than a broadcast observation. Their budget is the dispersion of that reconstruction, measured where a broadcast value and a reanalysis value coexist and applied where they do not. The reconstruction also carries a systematic offset that the budget does not absorb. Reanalysis error is regime-dependent, larger in shear and convection than in smooth synoptic flow, and a budget derived on one day inherits that day's regime. A cohort with substantially better Comm-B speed coverage would separate the detector from the reanalysis; we could not assemble one here.

6.3 What is not established

This section states what the procedure retains and how far that reaches. The settings here are calibrated against the signal itself; a record of what the crews actually selected would be needed to call them correct. The one external check available is the comparison of Section 5.5, on the single channel where a broadcast value exists.

That comparison is between two observers, not against a truth. The register holds a dialled value, which a crew may set minutes before the aircraft reaches it or leave in place after levelling elsewhere, and its encoding resolution differs from that of the observed altitude. Agreement between the two is therefore evidence that two imperfect observers of the same quantity concur; disagreement is not evidence that the detector is wrong. An external record of selected values, from flight-data

monitoring, would turn this calibration question into a validation question; that is the remedy, and it is outside what open data offers.

Finally, the budgets bound how well a committed constant summarises the signal it was measured on, which is a weaker statement than correctness. Independent information is required before that constant may be called a commanded value [22]. On four channels of five no such information exists in open surveillance, and none is claimed.

6.4 Future work

Three directions follow, each lifting one limitation above.

The budgets bound reconstruction rather than correctness. Only an external record of selected values can close that gap. Quick-access recordings carry the dialled set-points on all five channels, including the four where open surveillance offers nothing, so a cohort with both would turn the present calibration question into a validation question and would measure directly what the register comparison here can only bound on one channel.

The rule constrains a maximum, and whether a percentile constraint retains better operating points is unmeasured. Re-running the sweep while recording per-candidate error quantiles rather than the maximum alone would answer it on the same data, at the cost of the sweep, and would say whether the points retained here are the ones a better statistic would have chosen.

The motivation for detecting these values at all is that a trajectory model consumes them. Whether a setting calibrated this way changes what such a model learns, compared with a tolerance fixed by judgement, is the question the procedure was built to serve and the one it does not answer: it requires training the model on both inputs and comparing the trajectories it generates, which is a study in its own right.

7. Conclusion

A detector that recovers commanded values from open surveillance data cannot be validated against a reference, because on four channels of five none exists and on the fifth the available one is partial, intermittent and semantically distinct from what the detector claims. This paper has shown that such a detector can nonetheless be calibrated without one, by reading each channel's error budget off the signal itself and retaining the most covering setting that respects it.

The price of that calibration has been measured rather than assumed. Evaluation across aircraft subsets describes within-cohort variability. On the three dates excluded from calibration, coverage remains close and pooled median segment errors are similar, while the worst segment leaves the budget on most channel-dates.

A committed plateau is a stretch whose scatter about a constant is bounded by a budget the signal imposes, over the share of trajectory the tables report; it is not a transcript of what was commanded, and on the reconstructed channels it describes a reanalysis-derived signal. Published without its coverage and its budget, a setting says what was chosen, not what it will do.

Use of AI

OpenAI Codex and Anthropic Claude Code assisted with code development and manuscript drafting and editing. The authors take full responsibility for the methods, results and conclusions.

Author contributions

Conceptualization (G.J.); Methodology (G.J., R.D.); Software (G.J.); Validation (G.J., R.D.); Formal analysis (G.J.); Investigation (G.J.); Data curation (G.J.); Writing – original draft (G.J.); Writing – review and editing (G.J., R.D., P.V.); Visualization (G.J.); Supervision (R.D.).

Funding statement

The authors received no specific funding for this work.

Ethical basis

No ethical approval was required. All flight data are publicly available through the OpenSky Network under its terms of use, and no personal data are processed.

Open data statement

The cohort derives from the OpenSky Network [14] and was queried through its Trino interface, then decoded with pyModeS [25]. The underlying Mode-S and ADS-B records are available to researchers through the same interface under the OpenSky Network terms of use, and the retrieval, decoding and preprocessing steps that produce the cohort are released with the code. The raw observations are not redistributed here. The committed flight plan fixes the four dates, the ten typecodes and the 200 airframes rather than redrawing them per run; retrieval requires provider access. Meteorological enrichment uses the ERA5 reanalysis [27], interpolated with fastmeteo [26].

Reproducibility statement

Code, experiment results and figure-generation scripts accompany this paper at <https://github.com/JarryGabriel/papers-reproducibility>, in opensky-symposium/2026/plateau-calibration, under the EUPL v1.2. Access to the underlying data is subject to the providers' terms.

References

- [1] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. “Neural Ordinary Differential Equations”. In: *Proceedings of the 32nd Conference on Neural Information Processing Systems*. Montréal, Canada, 2018. arXiv: 1806.07366 [cs.LG].
- [2] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. “Neural controlled differential equations for irregular time series”. In: *Advances in neural information processing systems* 33 (2020).
- [3] Gabriel Jarry, Ramon Dalmau, Xavier Olive, and Philippe Very. *A Neural ODE Approach to Aircraft Flight Dynamics Modelling*. 2025. DOI: 10.48550/ARXIV.2509.23307. URL: <https://arxiv.org/abs/2509.23307>.
- [4] Gabriel Jarry and Xavier Olive. “Generation of Vertical Profiles with Neural Ordinary Differential Equations Trained on Open Trajectory Data”. In: *Journal of Open Aviation Science* 4.2 (Mar. 2026). ISSN: 2773-1626. DOI: 10.59490/joas.2026.8457. URL: <http://dx.doi.org/10.59490/joas.2026.8457>.
- [5] Ramon Dalmau, Marc Pérez-Batlle, and Xavier Prats. “Real-time Identification of Guidance Modes in Aircraft Descents Using Surveillance Data: An Interacting Multiple-Model (IMM) Filter Application”. In: *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. IEEE, 2018. DOI: 10.1109/DASC.2018.8569811.

- [6] Homeyra Khaledian, Raúl Sáez, Jordi Vilà-Valls, and Xavier Prats. “Interacting Multiple Model Filtering for Aircraft Guidance Modes Identification from Surveillance Data”. In: *Journal of Guidance, Control, and Dynamics* 46.8 (Aug. 2023), pp. 1580–1595. ISSN: 1533-3884. DOI: 10.2514/1.g007139. URL: <http://dx.doi.org/10.2514/1.G007139>. 460
- [7] Homeyra Khaledian, Henk A. P. Blom, Jordi Vilà-Valls, and Xavier Prats. “Interacting Multiple Model Joint Aircraft Guidance Mode and Control Set-Point Estimation”. In: *Journal of Guidance, Control, and Dynamics* 49.7 (July 2026), pp. 2127–2143. ISSN: 1533-3884. DOI: 10.2514/1.g009123. URL: <http://dx.doi.org/10.2514/1.G009123>. 461
- [8] Rémi Perrichon, Xavier Gendre, and Thierry Klein. “Hidden Markov Models and Flight Phase Identification”. In: *Journal of Open Aviation Science* 1.1 (2024). DOI: 10.59490/joas.2024.7269. 462
- [9] Junzi Sun, Joost Ellerbroek, and Jacco Hoekstra. “Flight Extraction and Phase Identification for Large Automatic Dependent Surveillance–Broadcast Datasets”. In: *Journal of Aerospace Information Systems* 14.10 (2017), pp. 566–572. DOI: 10.2514/1.I010520. 463
- [10] Junzi Sun, Jacco M Hoekstra, and Joost Ellerbroek. “OpenAP: An open-source aircraft performance model for air transportation studies and simulations”. In: *Aerospace* 7.8 (2020). DOI: 10.3390/aerospace7080104. 464
- [11] Xavier Olive. “traffic, a toolbox for processing and analysing air traffic data”. In: *Journal of Open Source Software* 4.39 (2019), p. 1518. DOI: 10.21105/joss.01518. 465
- [12] European Aviation Safety Agency. *Certification of Mode S Transponder Systems for Enhanced Surveillance*. AMC 20-13; ED Decision 2006/12/R, Annex I. European Aviation Safety Agency. 2006. URL: <https://www.easa.europa.eu/en/downloads/1497/en>. 466
- [13] Junzi Sun. *The 1090 Megahertz Riddle: A Guide to Decoding Mode S and ADS-B Signals*. 2nd ed. TU Delft OPEN Publishing, 2021. DOI: 10.34641/mg.11. 467
- [14] Matthias Schäfer, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, and Matthias 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*. 2014. DOI: 10.1109/ipsn.2014.6846743. 468
- [15] Martin Strohmeier, Xavier Olive, Jannis Lübke, Matthias Schäfer, and Vincent Lenders. “Crowd-sourced Air Traffic Data from the OpenSky Network 2019–2020”. In: *Earth System Science Data* 13 (2021), pp. 357–366. DOI: 10.5194/essd-13-357-2021. 469
- [16] Arik Ermshaus, Patrick Schäfer, and Ulf Leser. “ClaSP – Parameter-free Time Series Segmentation”. In: *Data Mining and Knowledge Discovery*. 2022. DOI: 10.1007/s10618-023-00923-x. arXiv: 2207.13987 [cs.LG]. URL: <https://arxiv.org/abs/2207.13987>. 470
- [17] Arik Ermshaus, Patrick Schäfer, and Ulf Leser. “CLaP – State Detection from Time Series”. In: *Proceedings of the VLDB Endowment*. 2025. arXiv: 2504.01783 [cs.LG]. URL: <https://arxiv.org/abs/2504.01783>. 471
- [18] R. Killick, P. Fearnhead, and I. A. Eckley. “Optimal Detection of Changepoints With a Linear Computational Cost”. In: *Journal of the American Statistical Association* 107.500 (Oct. 2012), pp. 1590–1598. ISSN: 1537-274X. DOI: 10.1080/01621459.2012.737745. URL: <http://dx.doi.org/10.1080/01621459.2012.737745>. 472
- [19] Charles Truong, Laurent Oudre, and Nicolas Vayatis. “Selective Review of Offline Change Point Detection Methods”. In: *Signal Processing* 167 (2020), p. 107299. DOI: 10.1016/j.sigpro.2019.107299. 473
- [20] Yacov Y. Haimes, Leon S. Lasdon, and David A. Wismer. “On a bicriterion formulation of the problems of integrated system identification and system optimization”. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1.3 (1971), pp. 296–297. DOI: 10.1109/TSMC.1971.4308298. 474
- [21] George Mavrotas. “Effective implementation of the ϵ -constraint method in multi-objective mathematical programming problems”. In: *Applied Mathematics and Computation* 213.2 (2009), pp. 455–465. DOI: 10.1016/j.amc.2009.03.037. 475

- [22] Christian Hennig. “What Are the True Clusters?” In: *Pattern Recognition Letters* 64 (2015), pp. 53–62. DOI: [10.1016/j.patrec.2015.04.009](https://doi.org/10.1016/j.patrec.2015.04.009). 510
- [23] Stanley E. Lazic. “The Problem of Pseudoreplication in Neuroscientific Studies: Is It Affecting Your Analysis?” In: *BMC Neuroscience* 11 (2010), p. 5. DOI: [10.1186/1471-2202-11-5](https://doi.org/10.1186/1471-2202-11-5). 511
- [24] Angelika Swatowska, Henrich Glaser-Opitz, Angela Nuic, and Vincent Mouillet. “Data-Driven Aircraft Speed Profile Prediction for Enhanced Trajectory Realism”. In: *Air Transportation Research & Development Symposium (ATRDS)*. Delft, Netherlands: EUROCONTROL/FAA, 2026. DOI: [10.5281/zenodo.21826781](https://doi.org/10.5281/zenodo.21826781). 512
- [25] Junzi Sun, Huy Vũ, Joost Ellerbroek, and Jacco Hoekstra. “pyModeS: Decoding Mode S Surveillance Data for Open Air Transportation Research”. In: *IEEE Transactions on Intelligent Transportation Systems* (2019). DOI: [10.1109/TITS.2019.2914770](https://doi.org/10.1109/TITS.2019.2914770). 513
- [26] Junzi Sun and Esther Roosenbrand. “Fast contrail estimation with OpenSky data”. In: *Journal of Open Aviation Science* 1.2 (Nov. 2023). DOI: [10.59490/joas.2023.7264](https://doi.org/10.59490/joas.2023.7264). 514
- [27] Hans Hersbach et al. “The ERA5 global reanalysis”. In: *Quarterly Journal of the Royal Meteorological Society* 146.730 (2020), pp. 1999–2049. DOI: [10.1002/qj.3803](https://doi.org/10.1002/qj.3803). 515