

Validating Storm Sensei's Lightning Detection Engine

Methodology, Results, and Supportable Claims

July 10, 2026 — revised October 2, 2026

Key Results

Using satellite lightning data only, which makes these figures a floor on production performance, the engine detected 72.2% of 295 independently verified storm events (95% CI 66.8–77.0%) and 74.0% of the events the satellite sensor could physically observe. We report these two figures together throughout this paper, because either one alone can misstate performance. Across 759 independently certified quiet periods, the engine produced no false positives, bounding the false-positive rate below 0.40% at 95% confidence, and no false alarms at the highest-severity tier. Average lead time on detected events was 36.5 minutes.

Detection varies by region, from 55.1% to 93.8% across five climatologically distinct US regions. We report this spread rather than averaging it away. In Monroe County, PA, home to our deployed venue at Buck Hill Falls, the engine detected 61.8% of verified events, and 65.4% of those the satellite could observe. These figures come from a historical replay against verified storm records. They are not a measure of the on-site sensor's uptime or live performance.

Executive Summary

Storm Sensei's lightning detection engine has been independently, rigorously validated against 1,054 real historical time windows — 295 independently verified storm events and 759 independently certified quiet-period controls, spanning five climatologically distinct US regions. The measurement was run against the exact code currently in production, with every result traceable to its source data and software version.

The result: 72.2% detection of verified storms, with an average 36.5-minute warning lead time, and a false-positive rate bounded below 0.40% at 95% confidence with zero false positives observed across all 759 certified quiet periods. This document presents those results, what they mean, and how they should and should not be used.

How the Detection Engine Ensures Reliability

Storm Sensei's alerting is built around a deterministic safety floor, not solely an AI judgment call. A plain, non-AI calculation from hard evidence — strike distance, time-to-arrival, active National Weather Service warnings, on-site sensor readings — establishes the minimum alert level the system is required to report at any moment. An AI model adds context and plain-language guidance on top of this floor, but its output can only raise the alert level to meet the floor, never lower it below what the hard evidence requires — on every code path, including any degraded or failed-connection scenario.

A single, shared detection module

The deterministic detection logic — the calculations that decide alert severity — lives in one code module used directly by both the live production system and the validation process described in this document. There is no separate reimplementation for testing purposes, which is what makes the results in this paper a direct measurement of the actual deployed system, not an approximation of it.

Until September 2026 this was true of the server-side system and the validation harness, while the in-app display necessarily carried a hand-maintained copy of the same logic, because a web browser cannot load a server module directly. That copy has since been removed: the app now loads the identical module over the network and refuses to

begin monitoring at all if it fails to load or arrives incomplete. As of September 23, 2026, every path that can raise an alert — the server, the background alerting pipeline, the in-app display, and the validation process behind this document — executes the same file.

Storm-motion and threshold logic

- Time-to-arrival calculations require two consistent, same-source data points before producing an estimate, preventing spurious short-time-to-arrival readings from comparing incompatible sensor snapshots.
- Arrival estimates are refused entirely when the two data points they would be computed from are closer together in time than a storm’s movement can meaningfully be measured over — a short interval makes ordinary flicker in the data look like extreme storm motion.
- A computed storm position is discarded when it is not corroborated by the nearest strike actually detected, and a computed storm speed is discarded when it exceeds what any real storm can sustain. Both produce no estimate rather than a doubtful one.
- The system distinguishes “no lightning detected” from “lightning data temporarily unavailable” — a degraded data feed never silently presents as confirmed-clear conditions.
- Storm-approach detection scales continuously with actual approach geometry, rather than using a fixed directional window.
- The highest-severity alert tier fires at 6 miles (distance rounded to 0.1 mile), consistent with standard public lightning-safety guidance (the “30-30 rule”: count 30 seconds from flash to thunder, seek shelter).
- An on-site lightning sensor, where installed at a venue, independently corroborates the assessment.

Validation Methodology

Results are measured against a pre-registered methodology (METHODOLOGY.md, version 1.0), finalized before any data collection began, designed to produce a genuinely defensible measurement:

- Positive events are independently verified storm records from the National Centers for Environmental Information, not calendar samples — required to begin within 25 miles of a real venue coordinate, consolidated where multiple records describe the same storm.
- Negative (quiet-period) windows are only certified after independently passing four separate verification checks — no recorded storm event, no weather-service warning, no thunder reported at nearby weather stations, and no significant radar reflectivity — deliberately excluding the satellite sensor itself from certifying its own negative examples.
- Detection rate must always be reported two ways together (among all verified events, and among events the sensor could physically observe) and never interchanged — preventing a common way real detection results get overstated by silently reporting only the more favorable of the two.
- Five US regions with genuinely different storm climatology were used, with results required to be reported per-region as well as pooled, so no single region's pattern can hide inside an average.
- Every result is required to record the exact software version it was measured against, the random seed used for sampling, and full source data provenance — so any number can be independently reconstructed and verified.

Results

The methodology was executed in full: 295 independently verified positive events and 759 independently certified negative (quiet-period) windows were collected across the five regions. The current, fully-hardened detection logic was then replayed against every one of these 1,054 windows.

Pooled results

Metric	Result	Basis
Detection rate (D1 — all verified events)	72.2%	213 of 295, 95% CI 66.8–77.0%
Detection rate (D2 — sensor-visible events)	74.0%	213 of 288, 95% CI 68.6–78.7%
Reached at least a caution-level alert pre-event	96.3%	284 of 295
Average lead time (detected events)	36.5 min	median 45 min
False-positive rate	< 0.40%	0 observed / 759 certified negatives, 95% conf.
Critical (highest-severity) false positives	0	0 / 759
Caution-level alerts on confirmed-quiet windows	0.7%	5 / 759 — reviewed individually, all defensible

A related but distinct figure — not a competing detection rate: Detection rate (above) measures whether the system reached a specific tier for a genuinely dangerous event. A separate, much looser question is how often the system showed at least some elevated caution before an event, even for storms it didn't fully catch at that tier. Among all 295 verified events, 284 (96.3%) triggered at least a WATCH-tier signal before the event's verified onset. WATCH is deliberately the most sensitive tier — it fires on almost any measurable indication of risk — so this is a substantially lower bar than the detection rate above, and is not a substitute for it. It is included here as a complementary observation: even among events the system didn't fully catch, it rarely stayed completely silent.

Per-region results

Region	Detection (D1)	Detected	Median Lead	False Pos.
Citrus, FL	93.8%	15/16	50 min	0/53
Palm Beach, FL	86.4%	51/59	45 min	0/136
El Paso, CO	74.1%	86/116	50 min	0/176
Monroe, PA (Buck Hill Falls)	61.8%	34/55	35 min	0/204
Travis, TX	55.1%	27/49	25 min	0/190

This spread reflects real, known differences in how the satellite lightning sensor performs across storm types — it detects Florida sea-breeze convection most reliably and Texas dryline convection least reliably. This is disclosed deliberately rather than averaged away, consistent with the methodology's per-region reporting requirement.

Citrus, FL rests on the smallest sample of any region: its 93.8% detection rate (15/16) is genuine and reproducible, but should be read with that in mind rather than as evidence the system performs uniformly better in this region than the pooled figure suggests. The small sample is a reporting-density artifact, not a sign of infrequent storms in the area — NCEI report density tracks population and spotter/damage-assessment coverage, not actual meteorology, and Citrus produced the fewest consolidated reports of any region in this dataset (16 of

295 pooled). Citrus's reports also skew older than the other regions' (69% predate the MRMS archive's October 2020 start, versus ~38% pooled), compounding the effect for any rain-specific companion analysis drawing on the same regional sample.

Verification, not just trust

Before these results were accepted, ten individual event timelines were hand-inspected against the raw underlying data — three detected events, two missed events, and all five of the caution-level alerts that occurred during certified-quiet windows. Every classification held up under inspection. The two missed events were confirmed to be two genuinely different, honest failure modes: one where the storm's closest approach never actually justified a higher alert under the venue-proximity rules, and one case of complete satellite blindness for that specific window. Neither reflects a flaw in the scoring; both reflect real limits of the specific data source being measured here.

A correction found by tightening the engine, not by re-running the numbers

This section describes the September 23 revision, measured on the prior 5-mile configuration. Its figures were slightly lower than those published on September 20, and the reason is worth stating plainly.

In September 2026 three additional plausibility checks were added to the part of the engine that estimates how fast a storm is moving and when it will arrive. Two of them matter here: a ceiling that discards any computed storm speed fast enough to be physically implausible, and a corroboration test that discards a computed storm position when it is not supported by the nearest lightning strike the system actually detected.

Re-running the identical backtest — same 1,054 windows, same satellite data, same scoring, nothing changed but the engine itself — against the tightened code moved the pooled detection rate from 70.5% to 69.8%. Two of the 295 verified events changed classification. Both turned out to be alerts the earlier code should never have issued.

The clearer of the two is an El Paso County storm from June 2025. The pre-September engine issued a highest-severity alert for it, twice, at five and ten minutes before the event — on an arrival estimate of three minutes. At that moment the nearest lightning strike the satellite had actually detected was 16.2 miles from the venue, and no strike at all was within the alert radius. The alert was arithmetic, not evidence: an averaged storm position that no detected strike supported. Under the corrected engine the same window is a caution-tier alert, which is what the evidence justified. It was previously counted as a correct detection, and it was inflating the published detection rate.

The second case, a Monroe County storm from July 2021, is the same failure in a milder form: a warning issued 35 minutes ahead, from an arrival estimate built on four scattered flashes averaged into a storm position 12.4 miles away, with nothing inside the alert radius. Notably, the same event is still detected at the wider 8-mile near-band tested elsewhere in this document — there, real strikes 6.2 miles out carry it on their own. It never needed the estimate.

The false-positive results did not move at all. Zero false positives across all 759 certified quiet windows, zero critical false positives, five caution-tier windows — identical, in every region, before and after. Every change was on the detection side, and every change was downward, which is what the mechanism requires: removing an arrival estimate can only lower an alert tier, never raise one.

A measurement that only ever improves is not measuring anything. This correction cost 0.7 percentage points of headline detection rate and removed a highest-severity alert that pointed at a storm sixteen miles away. We regard that as the measurement working. The September 23 figures were lower than the previous revision's and better supported by the evidence behind them, and we would rather publish the second kind of number than the first.

What These Numbers Actually Mean

This is a floor, not the full system

Important scope note: This measurement replays satellite lightning data only. The live production system also incorporates ground-network lightning data, National Weather Service warnings, forecast data, atmospheric instability data, and on-site sensor data where available — all of which can only raise an alert tier, never lower one. Because those signals can only raise a tier, the deployed system detects at least as many real storms, at least as early, as this document reports — so the detection-rate and lead-time figures are a floor. That reasoning does not extend to the false-positive figure, which the same mechanism works against; see “Near-band sensitivity” and “Polling-cadence sensitivity” below and “How to Read These Results” below. This document states only what has actually been measured, not what is inferred or assumed.

Detection rate is a conservative statement, if anything

Because a verified positive event can occur anywhere within 25 miles of a venue while the system's warning threshold is a tighter 6 miles, some fraction of the events counted as “not detected” in this measurement are storms that came close enough to register in the historical record but not close enough to the venue to warrant an alert under the product's own venue-focused design. This means the reported detection rate likely understates, rather than overstates, how the system performs against threats that genuinely reach a monitored venue.

Results for the Deployed Venue: Buck Hill Falls

Until Storm Sensei operates at multiple venues, the pooled 72.2% figure is a statement about the underlying technology's general capability, not a prediction of what a golfer at Buck Hill Falls specifically experiences. That number is 61.8% (65.4% of events the satellite could observe). For the current deployment, this is the more relevant figure; the pooled figure describes the product's general validated capability.

What this number is, and what it is not. The Buck Hill Falls figure is the Monroe County, PA regional result from the historical backtest described in this document: the share of independently verified storm events near that venue for which the engine, replayed against archived satellite lightning data, reached warning tier before the event began. It is a measurement of the detection logic against the past. It is not a statistic about the deployed installation — not sensor uptime, not data-feed availability, not how often the venue has been correctly alerted in live operation. No figure in this document describes those things, including this one.

Near-band sensitivity, 5 → 8 miles (historical): does the 5-mile threshold drive the zero?

This study was run on the prior configuration, whose satellite near-band was 5 miles (the production band has since been widened to 6 miles; see “Near-band widening, October 2026” below); the in-app path uses 8 miles. To test whether that threshold difference alone changes the false-positive result, the same 759 certified quiet windows and the same 295 positive events were replayed on the same satellite data with the near-band widened from that prior 5 miles to 8 miles. One constant changed; no tier logic, window definition, or event certification was altered. The run was pre-registered before execution; the pre-registration, harness, and full results are retained in the codebase.

False positives at warning level or above were zero at both thresholds, pooled and in every region. Windows reaching the caution tier were five at both. Not one of the 759 quiet windows changed classification.

The reason matters as much as the result. Only 7 of the 759 quiet windows contain any satellite flash within 25 miles of the venue, and the closest flash anywhere in the quiet set is 11.6 miles out. The 5-to-8-mile band is empty across every quiet window. Widening the threshold changed nothing because there was nothing in that band to catch. The correct reading is therefore narrow: on this dataset, the false-positive count is not sensitive to the near-band threshold. This is not evidence that an 8-mile band is generally safe, and a quiet-period set containing more marginal activity could behave differently.

This test also does not validate the ground-network feed. It widened the satellite band; the app adds a different sensor at 8 miles, with its own detection characteristics. It measures threshold geometry, not the deployed merge.

Reported for completeness and not as headline figures: at the 8-mile band the same replay detected 76.9% of events (D1) and 78.8% of satellite-visible events (D2), with 38.8 minutes of average lead time. These describe a

configuration nothing in production runs and do not replace the Key Results figures. Of the 25 events whose classification changed relative to the prior 5-mile configuration, 21 gained pre-event detection; the other four escalated only after the event had begun.

Near-band widening, October 2026

The background alert path's effective proximity-DANGER radius was 5 miles while published material, including this paper's statement that the highest-severity tier fires at 6 miles, said 6. The cause: the background path runs on satellite data only, and the list of satellite flashes counted as "near" was filtered at 5 miles, while the highest-severity proximity test requires both a near flash and a closest flash within 6 miles. A single satellite flash 5.5 miles away therefore produced a WATCH, not a highest-severity alert, on that path. On the in-app path, a 5.5-mile satellite flash reached the highest-severity tier only if the ground network independently reported some other strike within 8 miles.

The correction was evaluated under a pre-registered protocol before any configuration was run. The plan was locked on commit db96703 and amended on commit 7d2d723 (execution mechanism only, before any run); results were recorded on commit fbeab10, which makes no adoption decision; the decision to adopt 6 miles was recorded separately on commit 48d3daa on October 2, 2026. The evaluation replayed the same 1,054 windows with the satellite near-band at 5 miles (the control, which reproduced the previously published figures exactly), 6, 7 and 8 miles.

Results at 6 miles: detection among all verified events (D1) 72.2% (213 of 295; 95% CI 66.8–77.0%); among sensor-visible events (D2) 74.0% (213 of 288; 95% CI 68.6–78.7%); average lead time 36.5 minutes, median 45. Zero false positives at warning level or above and zero critical false positives across all 759 certified quiet windows, and five caution-tier windows — identical to the 5-mile configuration.

Nine events changed classification relative to the 5-mile configuration: seven gained pre-event detection, and two escalated to the highest-severity tier only after the event had begun. Every one of the nine was hand-inspected against the raw flash records. Counts: Genuine 8, Location-marginal 1, Artifact 0. The pre-registration defined these categories but not numeric thresholds; the thresholds were fixed before the evidence for each event was read. Genuine = qualifying flash corroborated by ≥ 3 other flashes ≤ 10 mi within ± 5 min, or ≥ 100 other flashes ≤ 25 mi within ± 10 min (dense verified storm), or an unmistakable multi-flash approach sequence consistent with the NCEI storm; Location-marginal = within the geolocation uncertainty of the 5-mi line (true for all 9) and no such corroboration; Artifact = no corroborating flash in space or time (no other flash ≤ 25 mi within ± 10 min). The one Location-marginal event is a Palm Beach County event from October 2020, an after-onset escalation. All seven gains rest on a flash 0.4–1.0 mi from the former 5-mile line, within the 5 km location-accuracy requirement NOAA states for the satellite sensor; that is true by construction for any flash in the 5–6 mile ring and is reported, not used as a criterion.

Distances are rounded to 0.1 mile before they are compared with the band, so the effective cutoff is approximately 6.05 miles. Two of the seven gains rest on a flash recorded at exactly 6.0 miles.

Zero false positives at 6 miles is uninformative about false-alarm risk in the widened ring. The closest satellite flash in any certified quiet window is 11.6 miles from the venue, and only 7 of the 759 quiet windows contain any flash within 25 miles, so no quiet window has anything in the 5-to-6-mile ring: the result shows only that the quiet set cannot test the widened band. Live monitoring of that ring applies — background-path alerts whose closest flash lies between 5 and 6 miles are logged for review.

Seven and eight miles were evaluated in the same run and were not adopted. At 7 miles: D1 73.9% (218 of 295), D2 75.7% (218 of 288). At 8 miles: D1 76.9% (227 of 295), D2 78.8% (227 of 288). Because the highest-severity proximity test requires a flash within 6 miles, widening beyond 6 can only add warning-level alerts, not highest-severity ones.

Polling-cadence sensitivity: does the five-minute measurement interval drive the zero?

The backtest samples every window at a fixed five-minute interval, matching the app’s automatic polling. A fault that only appears at some other interval could not have shown up in it. In September 2026 exactly that happened in a live session: a user tapping the manual refresh button shortly after an automatic poll produced two readings roughly a minute apart, and the resulting arrival estimate was physically impossible. The engine was fixed; this study asks whether the measurement holds at intervals the original backtest never exercised.

The same 1,054 windows, the same satellite data, the same scoring, the same engine. Only the sampling interval changed: two, three, five and seven minutes, plus a deliberately irregular schedule averaging five minutes. The study was pre-registered before execution, including a control requirement that the five-minute run reproduce the main result exactly before any other interval was attempted. It did.

Zero false positives at warning level or above, and zero critical false positives, at every interval tested, in every one of the five regions. Caution-tier windows were five of 759 at every fixed interval and four under the irregular schedule.

Two of those intervals are not hypothetical. The app polls automatically every five minutes, but the manual refresh button inserts an extra reading at whatever moment it is pressed. A tap two or three minutes after an automatic poll produces exactly the two- and three-minute intervals tested here. These are not laboratory conditions; they are ordinary use. The three-minute case matters most: it is the shortest interval a user can produce at which the engine will still compute an arrival estimate at all. It produced zero false positives across all 759 certified quiet windows.

The cost is real and is stated here as a cost. Below three minutes the engine refuses to compute storm motion at all, on the grounds that two readings that close together cannot distinguish a moving storm from ordinary flicker in the data. That refusal has a price, and this study measures it: at a two-minute interval the detection rate falls from 69.8% to 61.7% — 8.1 percentage points, 24 of the 295 verified events. Every one of those 24 had been detected solely on an arrival estimate, with no strike inside the alert radius at any point before the event; none lost a detection that rested on actually-detected lightning. Twenty-three of the 24 still reach caution tier before the event. One does not, and has no pre-event signal at all at that interval.

This is not a description of what the deployed system gives up. Automatic polling is fixed at five minutes and a shortened interval affects only the single pair of readings a manual tap creates, not a session. The right reading is narrower: when the engine is given readings too close together to carry information, it surrenders roughly eight points of detection rather than manufacture an arrival estimate — and it raises no false alarms doing so. That is the trade the safeguard was built to make, now measured rather than assumed. This study was measured on the prior 5-mile configuration and has not been re-run at 6 miles.

Scope and Limitations

- This result covers satellite lightning detection only, on a 6-mile near-band. Weather-service warnings, forecast signals, atmospheric instability data, and on-site sensor data are not included; on-site sensor history at the one live venue currently spans only a few weeks.
- Ground-network lightning is live in the app and is not in this backtest. The in-app display merges ground-network lightning data (Vaisala Xweather, 8-mile near-band) with the satellite data measured here (6-mile near-band). The background alert path does not — it runs satellite-only, matching this backtest exactly.
- The ground-network feed is not replayable. No historical archive of that data is available to us, and we are not aware of one being available to anyone. This is a standing limitation, not a pending task: its detection efficiency, false-strike rate, and geolocation error on historical storms cannot be measured by us or independently checked by a reader. The in-app display therefore sees more, and sees it sooner, than the numbers in this paper describe, and it may also produce alerts this backtest could not have produced.

- Sensor-failure behavior was not exercised by this backtest. In the replay harness the satellite feed is available in every window — the cached data layer reports success unconditionally — and the ground-network feed is structurally absent. No window in the 1,054-window dataset therefore represents one lightning channel failing while the other continues to report. The system’s behavior in that state is not “tested and held”; it was never in the sample space. This applies to the false-positive figure as much as to the detection rate: zero false positives across 759 certified quiet periods is a statement about windows in which both channels behaved as designed.
- Detection-rate confidence intervals are approximately ± 5 percentage points at this sample size ($n=295$ positive events) — real, but wide enough that small differences between regions or future re-measurements should not be over-interpreted.
- Lead time is computed only over events the system actually detected — it says nothing about how early a missed event might otherwise have been caught.
- The cadence study, like the rest of this document, replays satellite data only. The live in-app display merges a ground-network feed with a much shorter refresh window, whose behaviour at short polling intervals cannot be replayed and is not measured here.
- A minor, non-safety-relevant behavior was noted during verification: alert severity can briefly step down and back up during a genuinely approaching storm, since the underlying logic evaluates each data point independently with no smoothing. This does not weaken the safety guarantee (the system’s conservative floor logic remains intact at every instant) but may be worth a future refinement for perceived stability, restricted to delaying downgrades only, never delaying an upgrade.

How to Read These Results

These results are real and reproducible. Several points govern how they should be read.

- Two detection figures, always together. One counts every verified storm event; the other counts only events the satellite sensor could physically observe. Neither is presented without the other.
- Technology versus venue. The pooled figures describe the detection technology across five regions. For the deployed venue at Buck Hill Falls, the relevant figures are the Monroe County results: 61.8% of all verified events and 65.4% of those the satellite could observe.
- What was measured. This study replays satellite lightning data only. The deployed system also uses ground-network lightning, National Weather Service warnings, forecast and atmospheric instability data, and on-site sensor data where installed. Ground-network lightning cannot be replayed historically, by us or, as far as we know, by anyone.
- Which path the false-positive figure describes. The false-positive bound applies to the configuration measured: satellite lightning alone on a 6-mile near band, which is exactly what the background alert path runs. It is not a measurement of the in-app display, which also merges ground-network lightning on an 8-mile band. Because no certified quiet window contained a satellite flash within 11.6 miles, the zero-false-positive result does not test the widened band directly; false alarms there are monitored live.
- Detection and lead time are floors for both paths. Additional evidence can only raise an alert tier, never lower it, so a configuration with more sensors over a wider band detects at least as many real storms, at least as early. The same mechanism works against a false-positive rate, which is why only that figure carries a path qualifier.
- Provenance. This document, its underlying data, and its full computational provenance are retained and available for independent audit of any figure.

Appendix: Provenance

Full replay results, per-window timelines, certification records for all 1,054 windows, and the exact software version (git commit) these results were measured against are retained in the codebase and available for independent audit at any time.

Revision history

October 2, 2026. Satellite near-band widened from 5 to 6 miles following a pre-registered evaluation (lock db96703, amendment 7d2d723, results fbeab10, decision 48d3daa); every figure measured at the satellite near-band is restated at the 6-mile configuration. Pooled detection 69.8% → 72.2% (D1) and 71.5% → 74.0% (D2); D1 95% CI ±5.2 points → 66.8–77.0%; average lead time 36.3 → 36.5 minutes (median 45, unchanged); Buck Hill Falls (Monroe, PA) 58.2% → 61.8%; Palm Beach, FL 83.1% → 86.4%; El Paso, CO 72.4% → 74.1%; Travis, TX 53.1% → 55.1%; Citrus, FL 93.8% (unchanged); regional range 53–94% → 55.1–93.8%; per-region median lead Citrus 45 → 50 minutes and Palm Beach 40 → 45 minutes (others unchanged). No false-positive figure changed (0 of 759, 0 critical, 5 caution-tier windows), and the at-least-caution figure is unchanged (96.3%, 284 of 295). Reason: the background alert path’s effective proximity radius was 5 miles, not the 6 miles published; the near-band was widened to match, and every statement of a 5-mile satellite near-band or threshold now reads 6 miles. The near-band sensitivity and polling-cadence sections are retained as measured on the prior 5-mile configuration; the cadence study has not been re-run at 6 miles. Added the section “Near-band widening, October 2026.”

September 23, 2026. Detection-rate figures corrected downward following the addition of three storm-motion plausibility checks to the production engine, and re-verification of the full 1,054-window backtest against the tightened code. Pooled detection 70.5% → 69.8% (D1) and 72.2% → 71.5% (D2); average lead time 36.9 → 36.3 minutes; Buck Hill Falls (Monroe, PA) 60.0% → 58.2%; El Paso, CO 73.3% → 72.4%. Near-band sensitivity figures 77.3% → 76.9% (D1), 79.2% → 78.8% (D2), 39.1 → 38.8 minutes. Cause and affected events are described under “A correction found by tightening the engine.” No false-positive figure changed. A polling-cadence robustness study was added.

September 23, 2026 — separate correction, not caused by the above. The near-band section previously reported that “28 events changed classification, 24 gained pre-event detection.” Those two numbers were arithmetically wrong independent of the engine corrections, and would have been wrong even if the engine had never been touched. Widening a search radius can only ever add detected strikes; it can never remove one. A widening comparison therefore cannot produce a net change larger than the number of events that newly gained detection, and the published figures claimed exactly that. The correct figures for the originally published run were 24 changed and 20 gained; after the engine correction above they are 25 and 21, which is what this revision states. This is recorded as its own dated correction rather than folded silently into the others, because an arithmetic error and a measurement correction are different kinds of mistake and a reader is entitled to see which one they are looking at.

September 20, 2026. Near-band sensitivity analysis added; scope and limitation sections expanded.

July 10, 2026. Initial publication.