# FORECAST RESIDUALS FOR THE DIAGNOSIS OF HIDDEN TRAFFIC EVENTS: THE RANKING OF NORMAL MODELS DEPENDS ON THE ALARM RULE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Hidden traffic events such as lane closures, demand surges, signal faults and sensor failures are commonly detected by forecasting normal traffic and thresholding the residual. We ask what the two stages of such a detector should condition on: the normal model that produces the residual, and the rule that reads it over sensors and time. The study uses a simulated signalised corridor in which every event episode has an event-free twin sharing all random variables, so that the event's effect on each sensor is known exactly, and radar measurements from Interstate 24 with a located event log. The central finding is that forecasting accuracy alone does not determine detection quality: the ranking of normal models depends on the residual score and the alarm rule. Under an instantaneous threshold a forecaster that conditions on recent observations predicts the event it is meant to expose. On the corridor the seasonal prior is not improved upon by the selected hybrid under this score, while under a rule that integrates the residual the accurate penalised autoregression decisively becomes the best linear pairing, at a longer delay and a longer time in alarm; the freeway ordering does not reverse. The relation between accuracy and detection differs between model families under one rule, and a nonlinear network and a zero-shot foundation model are poor inputs under either rule for different reasons. Two results of narrower scope follow: adding the nearest sensor to the detector's input raises recall substantially in this benchmark, and the normal model that is best for detection under an instantaneous rule is the worst among the tested strategies for reconstructing the event-free counterfactual. The generator, configurations, splits, full sweep tables and scripts accompany the submission; the corpus and checkpoints are to be released on publication.

## 1 INTRODUCTION

A lane closure, a stalled vehicle or a signal fault changes the flow at nearby detectors minutes before anyone reports it, and automatic incident detection has tried to read those changes for fifty years (Parkany & Xie, 2005). The dominant modern approach is *forecast-then-detect*: fit a model of normal traffic, compute the residual between its prediction and the measurement, and raise an alarm when the residual is large (Chandola et al., 2009; Choi et al., 2021). In the linear case the question this raises is old: statistical process control found that a chart on the residual of an autoregressive forecast can be slower to signal a sustained shift than a chart on the raw observations, because the forecast absorbs the shift (Alwan & Roberts, 1988; Wardell et al., 1994). Forecast-then-detect systems nonetheless use the most accurate forecaster available, encouraged by traffic-forecasting results in which a long temporal context of the target sensor carries most of the predictive information and spatial breadth adds little (Li et al., 2018; Yu et al., 2018; Anonymous, 2026). Whether accuracy is the right criterion for a detector's normal model, and for learned and pretrained forecasters of a spatial field, has not been tested with ground truth, because real data does not say which sensors an event moved.

This paper tests it. In a microscopic simulation of a signalised corridor, every event episode is paired with a twin that shares every random variable and differs only in the intervention, so $\Delta x = x^{\text{event}} - x^{\text{normal}}$ is known exactly; public radar data from Interstate 24 (Coursey et al., 2024), with the management centre's event log, tests what transfers. The central finding is that forecasting accuracy



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

alone does not determine detection quality: the ranking of the tested normal models depends on the residual score and the alarm rule. Under an instantaneous threshold, in a family of autoregressions on the seasonal residual recall falls as validation error falls, in a family of autoregressions on raw history it rises, because there the accuracy comes from representing the diurnal cycle. The seasonal prior is not improved upon by the selected hybrid under this score; under a cumulative-sum rule the first family's relation reverses and an accurate hybrid autoregression decisively becomes the best linear pairing on the corridor. The process-control mechanism, absorption of the shift by the adaptive forecast, is measured by injected steps and coefficient sums, and it does not account for every model: a nonlinear network retains as much of a step as the ridge and still detects less under the integrating rule.

**Contributions.** This is a diagnostic and benchmark study; no architecture is proposed. What it adds to the process-control result is: (1) a controlled comparison of three normal-model families, a nonlinear network and a pretrained foundation model under two residual scores, two families of alarm rule and two spatial aggregations, on a spatial traffic system with exact ground truth (Figure 2, Tables 1 and 2); (2) a quantitative evidence-to-null diagnostic that combines retained injected-step area with the clean-day accumulated-score tail, alongside evidence that attenuation alone does not explain every model (Section 3.4, Figure 5); (3) joint accounting of recall, delay and alarm duration, under which the comparison between rules changes (Figure 4, Table 2); (4) a test of each pairing after a change of signal plan and of demand with thresholds recalibrated (Table 3) and a statement of which findings the freeway data supports (Table 4); and (5) a verified paired benchmark on which counterfactual reconstruction is scored exactly (Section 5), with a supporting result on spatial context (Section 4).

## 2 SETUP

**Residuals.** Let $x_{i,t}$ be the measurement at sensor $i$ and interval $t$ (flow, occupancy or speed over 30 s). A normal model produces $\hat{x}_{i,t|t-1}$ from information available at $t-1$; the *forecast residual* is $e_{i,t} = x_{i,t} - \hat{x}_{i,t|t-1}$ and the *standardised residual* divides it by an estimate of its scale. The *seasonal prior* is a per-sensor historical average by day type and time of day; its predictive mean does not use recent observations, though the standardised detector built on it does, through the scale of Section 3.5. The *autoregression* predicts from the last $L$ observations of the sensor with calendar covariates: a ridge regression for its closed form (context $L$ in 30 s steps, penalty as a multiple of the number of training rows $N$), and a patch-mixing network (Tolstikhin et al., 2021; Anonymous, 2026) as the nonlinear alternative. The *hybrid* is a ridge autoregression on the seasonal residual. All models are fitted on event-free days only and strictly causally; validation error is the mean absolute one-step error on held-out event-free days, in vehicles per 30 s.

**The paired benchmark.** The corridor is a ten-junction signalised arterial in SUMO (Lopez et al., 2018) with three lanes each way, side streets, a parallel service road, fixed-time signals with a 90 s cycle and 82 detector stations reporting every 30 s, under a calibrated diurnal demand with day-to-day variation. Five event classes are interventions on the vehicle set, the controller or the measurement: capacity reduction and blockage, a demand pulse, a frozen signal phase and a sensor fault. Every event episode is generated with a twin from the same seed in which the intervention is absent; the twins are identical before onset and in placebo runs, verified to the bit. The twin's exact effect $\Delta x$ is what the counterfactual reconstruction, the definition of the sensors an event moves, the severity axis and the simulator's verification rest on; every detection result uses only the intervention's onset, duration, class and location. The corpus holds 500 pairs stratified over class, onset regime, location and ten simulation seeds, each seed a demand realisation shared by its 50 events, so the test half covers five held-out demand realisations with 50 events each and every corridor interval over events is conditional on those five; and 100 event-free days, split in half for thresholds and for the realised false-alarm rate and disjoint from the 68 days on which normal models are fitted. Appendix O labels every result by whether any event episode entered its selection or training (none for the fixed representations, the baselines, TimesFM and the real data).

**Real data.** FT-AED (Coursey et al., 2024) provides 30 s radar measurements from 49 stations along 18 miles of I-24 on twenty weekday mornings, with 44 crash reports and 19 labelled anomalies. Its supplementary log supplies milemarkers and classes; further days are analysed at a strictly

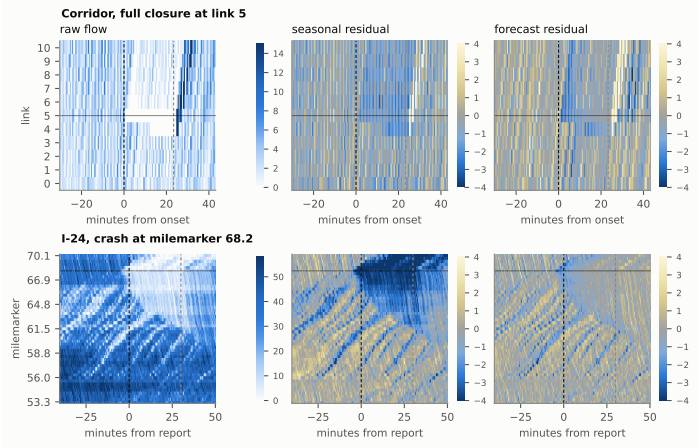



Figure 1: Raw flow, standardised seasonal residual and standardised forecast residual as space-time fields around one event; the affected band is a deficit of flow, shown as negative; the diverging map is monotone in lightness, so a deficit prints dark and a surplus light. Top: a full closure of link 5 on the corridor. Bottom: a crash at milemarker 68.2 on I-24, the crash with the largest response among the 38 located. The seasonal residual holds a sustained band at the affected sensors for the event's duration; the forecast residual shows an onset transient and fades while the event continues.

causal 120 s aggregate (Appendix J). Detection uses the benchmark's window from 20 minutes before to 15 after a report. Section 3.7 reports both network-level recall and a secondary analysis requiring the first alarming neighbourhood to lie near the logged location.

**Protocol.** Four things about a detector are kept separate: its *score*, instantaneous when it is a function of the current residual field and integrated when it accumulates past residuals (a cumulative sum); its *persistence rule*, the run of exceedances that raises an alarm; the *aggregation window* a learned detector reads; and the *alarm accounting*, excursions per day or time in alarm. The default instantaneous score for the non-learned methods is the *neighbourhood statistic*, the absolute standardised residual summed over the stations of three consecutive links and maximised over junctions; with the two-of-three rule, fixed before any result, it is the instantaneous statistic below, the rule is named wherever it changes, and CUSUM and the GLRT of Appendix E are the explicit exceptions. On the corridor an event is detected if the detector fires between onset and end; false alarms are excursions on event-free days. Thresholds are fitted per hour of the day on tuning days, the *realised* rate is measured on disjoint held-out days, the target is swept, and recall is compared at a realised two false alarms per day by interpolation between swept points, the operating point of every result unless stated. On the freeway, main-text performance comparisons use interpolated recall at matched realised alarm rates; nearest-point paired differences and bootstrap intervals, which need per-event flags and are computed at the swept point nearest that rate, are reported in Appendix J and do not quantify uncertainty in the interpolated difference. Delay is the first alarm's time from onset over detected events. Every corridor table scores from step 128 of each episode; Table 1 uses all 500 events and the others the test half (Appendix B).

**Calibration stability.** A separate event-free pairing-selection pilot used 56 fitting, 14 fitting-validation, 14 probe and 28 normal-test days. With 42 calibration days, one replication reached 0.589 recall but exceeded both burden targets (3.57 alarms/day, 2.40% alarm time); another reached 0.568 recall and narrowly met them (1.92/day, 0.95%). An 84-day calibration amendment with fresh seeds selected the same window-16 alarm family and reached 0.575 recall, 1.69 alarms/day and 0.87% alarm time. These runs motivate further calibration checks; one passing amendment does not establish that 84 days reliably controls burden.

Table 1: Recall at a realised two false alarms per day, all 500 corridor episodes from step 128, flow, the neighbourhood statistic with the two-of-three rule (and, in the second column, with a rule of four consecutive exceedances, under which the raw flow does not reach the rate), thresholds on 50 event-free days and the realised rate on 50 others, interpolated at the reporting rate. The forecast residual is the 64-lag ridge autoregression's at penalty $N$; "standardised" is the level-dependent scale of Section 3.5. The selection and test halves agree to within 0.04 on every row; the threshold-fitting spread is 0.01 to 0.02 for the unscaled rows and 0.07 for the standardised seasonal residual (Appendix B).

| Representation | Recall | Rule 4/4 | Delay | Closure | Blockage | Pulse | Signal | Fault |
|---|---|---|---|---|---|---|---|---|
| Raw flow | 0.082 | — | 434 s | 0.01 | 0.03 | 0.12 | 0.09 | 0.13 |
| Seasonal-prior residual, unscaled | 0.511 | 0.862 | **166 s** | 0.51 | 0.51 | 0.68 | 0.48 | 0.39 |
| Seasonal-prior residual, standardised | **0.758** | **0.963** | 281 s | 0.95 | 0.89 | 0.72 | 0.62 | 0.48 |
| Forecast residual | 0.285 | 0.483 | 264 s | 0.39 | 0.38 | 0.42 | 0.26 | 0.03 |
| Standardised forecast residual | 0.365 | 0.731 | 257 s | 0.50 | 0.49 | 0.48 | 0.33 | 0.02 |

# 3 THE NORMAL MODEL AND THE RULE THAT READS IT

## 3.1 FOUR REPRESENTATIONS, ONE RULE

Table 1 scores four representations of the same 500 episodes through the instantaneous statistic. Residualising by any model detects far more events than the raw measurement, and among the residuals the ordering is the reverse of the forecasting ordering: at the reporting point the standardised seasonal residual, from the worse forecaster by a fifth, detects about twice as many events as the standardised forecast residual. In the plotted comparison of Figure 3, the corridor and the released month of I-24 under the instantaneous statistic with the two-of-three rule, the seasonal residual stays above the forecast residual across the evaluated alarm rates. Figure 1 shows why: at the affected link of a closure the forecast residual falls to a tenth of its onset value within 25 minutes while the event is still active, whereas the seasonal residual holds, because once the event has run for a few intervals the autoregression's context contains the reduced flow, and the model predicts it.

## 3.2 ACCURACY AND DETECTION WITHIN AND BETWEEN FAMILIES

Two experiments vary the adaptation alone. In the first, the autoregression forecasts from a context ending $h$ steps earlier: recall rises with the lead and no lead reaches the seasonal prior, because a finite lead only postpones contamination. In the second, the normal model is selected on detection rather than on forecast error: three families crossed with six context lengths and seven ridge penalties, the autoregression on raw history, the hybrid, and the seasonal prior alone, selected on one half of the events and reported on the other (Figure 2, top left). The two families relate accuracy to detection differently under the same rule. In the hybrid family the rank correlation between validation error and test recall is $+0.94$: recall falls as the penalty lightens and the context lengthens, because a longer context explains more of the residual's persistent component, which is what a sustained event looks like, and no hybrid improves on the seasonal prior alone. In the raw-history family it is $-0.76$: there accuracy comes from representing the diurnal cycle, which is structure worth removing, and the accurate configurations are the better detectors, though none reaches the seasonal prior. The right criterion for the normal model is therefore not accuracy but what the accuracy is made of.

The mechanism the process-control literature describes is measured directly (Figure 5, Appendix E). An injected step of one or five residual standard deviations at one sensor shows that adaptive prediction attenuates a sustained shift: the 64-lag ridge retains a fifth of the step after fifteen minutes and 3% at thirty, the hybrid with a 128-step context 0.66 and 0.48. For a linear autoregression the residual under a constant shift has a decaying transient and a steady-state component $\delta(1 - \sum_\ell a_\ell)$; the fitted coefficient sums are 0.989 for the ridge and 0.838 for the hybrid, so the ridge's steady state is about 1% of the shift and the hybrid's 16%, and at thirty minutes the ridge has reached its asymptote while the hybrid, with a 64-minute context, has not. What an integrating rule can recover is this retained evidence, transient and steady state together, relative to the distribution of its accumulated score on event-free days, which sets the threshold; the measurements do not separate the two components' shares of the hybrid's gain, and no decomposition is claimed.

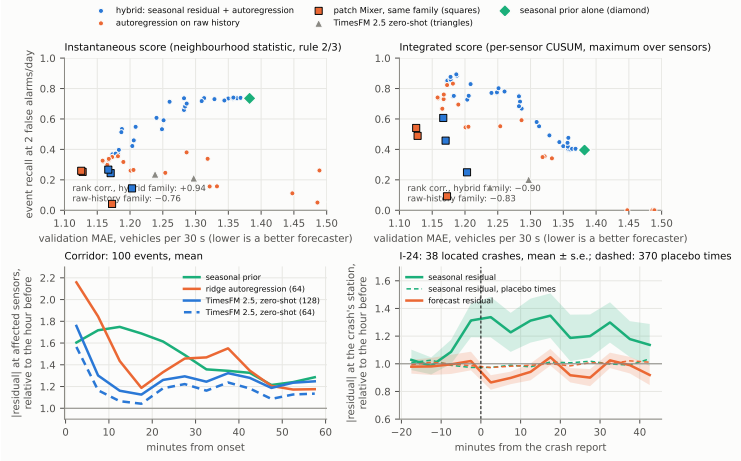



Figure 2: Top: event recall against validation forecast error for the 85 linear normal-model configurations, the six patch-Mixer models (squares) and TimesFM 2.5 zero-shot (triangles), corridor test half, level-dependent scale, under the instantaneous statistic (left) and the per-sensor cumulative sum maximised over sensors (right); both panels share their axes, and 17 raw-history configurations with validation error above 1.5 vehicles per 30 s, all with recall at most 0.33 under the statistic and zero under the cumulative sum, lie outside them. Bottom left: the absolute residual at the sensors an event moves, relative to the hour before onset, for the seasonal prior, the ridge autoregression and TimesFM 2.5 zero-shot, 100 corridor events. Bottom right: the absolute flow residual at the station nearest each of 38 located I-24 crashes, relative to the hour before the report, mean and standard error, with the same profile at 370 matched placebo times (dashed).

### 3.3 The rule that reads the residual

Re-scored with a two-sided cumulative sum per sensor in place of the instantaneous statistic, the same 85 linear configurations change order (Figure 2, top right): the hybrid family's correlation becomes $-0.90$ and the raw-history family's $-0.83$, the seasonal prior alone falls to 0.40 because the integrator accumulates the day's demand variation, and the best configuration is an accurate hybrid autoregression, at 0.89 against 0.74 for the seasonal prior under the statistic. Table 2 gives the two controls on which this comparison rests. The headline contrast changes the temporal and the spatial aggregation together; held fixed at either spatial aggregation, the seasonal prior leads under the instantaneous score and the hybrid under the cumulative sum, so the reversal belongs to the temporal rule. Resampling the five held-out demand realisations rather than individual events gives a seasonal-minus-hybrid recall difference of 0.004 [0.000, 0.012] under the instantaneous score and -0.492 [-0.500, -0.484] under CUSUM; these cluster intervals preserve the paired event structure and make the effective sample size explicit. The comparison also depends on how alarms are counted (Figure 4): at a matched excursion count the integrating pairing is in alarm three and a half times as long, matched on time in alarm its lead narrows to 0.84 against 0.73, and paired over commonly detected events the seasonal statistic is two and a half minutes faster. A persistence rule of three or more consecutive exceedances, chosen on the selection half, is a shorter integrator: it lifts every instantaneous pairing on the corridor without changing their order, the seasonal prior with the statistic to 0.95 at 95 s and half the time in alarm, above the hybrid with CUSUM on all three measures, and leaves the integrating pairings unmoved. On I-24 the run rules do not reproduce the corridor's gain and can remove the seasonal advantage (Section 3.7). Table 2 (bottom) sets the selected pairings side by side: the reversal survives the spatial controls, the cumulative sum buys its recall with a larger alarm burden and a longer delay, and a short seasonal run rule performs strongly on all three measures. Appendix D scores established baselines on the same fields: the autoencoder gains threefold from the seasonal-residual input, the graph deviation network, whose own forecast re-adapts, gains less, and a ridge gated to freeze its context when a change is suspected is worse than either headline pairing. The learned-detector row in Table 2 is a supervised diagnosis result, not an apples-to-apples replacement for the event-free fixed statistic: its training uses labelled simulated episodes. We therefore use its 0.96 recall primarily as evidence about spatial context and early

Table 2: Top: recall of the seasonal prior / the hybrid (128 lags, penalty $10N$) with the spatial aggregation held fixed, corridor test half, two excursions per day. Bottom: selected pairings and their trade-offs, corridor test half: recall, delay and fraction of event-free time in alarm at two excursions per day, and recall at a matched time in alarm (the seasonal statistic's 0.34%). The CUSUM rows use the two-of-three rule on the cumulative sum, under which they are unchanged from one-of-one; the dash means the matched-time result is not reported for the consecutive-run row. The ordering of the linear rows is specific to this corridor (Table 4); performance after a change of regime is in Table 3.

| Spatial aggregation | Instantaneous: seasonal / hybrid | CUSUM: seasonal / hybrid |
|---|---|---|
| Maximum over sensors | 0.31 / 0.19 | 0.40 / 0.89 |
| Neighbourhood pooling | 0.73 / 0.51 | 0.25 / 0.69 |

| Rule | Normal model | Recall | Delay | Time in alarm | Recall at matched time in alarm |
|---|---|---|---|---|---|
| Instantaneous, 2/3 | seasonal prior | 0.73 | 298 s | 0.34% | 0.73 |
| Instantaneous, run $\geq 3$ | seasonal prior | 0.95 | 95 s | 0.18% | — |
| CUSUM | hybrid 128, $10N$ | 0.89 | 444 s | 1.19% | 0.84 |
| CUSUM | ridge 64, $N$ | 0.74 | 331 s | 0.90% | 0.67 |
| Learned, 8 min, 2 nbrs | seasonal residual | 0.96 | 194 s | 0.38% | 0.95 |

warning. The label-free graph-deviation control isolates the representation effect under the same event-free training protocol: replacing raw flow by the seasonal residual raises recall from 0.276 to 0.478, while the detector trained without twin negatives retains 0.943 versus 0.953 recall with twins (Appendix D, Appendix F). Thus the headline learned gain is not attributed solely to supervision, and the twin construction chiefly improves localisation rather than detection.

### 3.4 A NONLINEAR NETWORK AND A FOUNDATION MODEL

The patch-mixing network at three context lengths (squares in Figure 2) is the most accurate model in its family and among the worst detectors under the statistic, with recall between 0.05 and 0.27. TimesFM 2.5 (Das et al., 2024), a 200-million-parameter pretrained forecaster run zero-shot, has recall of 0.21 and 0.24 at two contexts against 0.33 for the ridge and 0.73 for the seasonal prior; its residual at the affected sensors is back near baseline ten minutes after onset, while the seasonal prior's stays elevated for twenty (Figure 2, bottom left). Under the cumulative sum the two fail for different reasons. TimesFM attenuates the injected step within five minutes, size-dependent as a linear model cannot be, and stays at 0.15 to 0.20. The Mixer retains as much of a step as the ridge and reaches 0.61 at a delay of sixteen minutes; a supported contributor to its shortfall is the event-free distribution of its cumulative sum, whose daily maximum has a heavier tail than the ridge's and sets a higher threshold, while its marginal tails, run lengths and autocorrelation match the ridge's (Appendix E). Attenuation therefore does not by itself explain every model's detection performance; a full parameter-adapted foundation model was not run. An event-free affine calibration of TimesFM's output on 25 days worsens held-out MAE (1.238 to 1.294) and changes recall from 0.237 to 0.265 with overlapping intervals (Appendix E), so the zero-shot result is not explained by a simple output bias. A simple diagnostic combines the retained area of the injected step with the 90th percentile of the clean-day daily CUSUM maximum: its evidence-to-null ratios are 0.20 for the seasonal prior, 1.30 for the ridge, 2.03 for the hybrid, 1.20 for the Mixer and 0.12 for TimesFM (Appendix E). It predicts the broad linear and foundation-model ordering, but not the Mixer's full event deficit, so it is a measured diagnostic rather than a general criterion.

### 3.5 THE RESIDUAL SCALE

Standardising the seasonal residual by the parametric scale $\hat{\sigma}_i = a_i\sqrt{\ell_{i,t} + 1}$, with $\ell$ the trailing ten-minute mean flow, raises recall by half (0.49 to 0.74 on the test half); a per-sensor constant scale gives none of the gain, so it comes from the level dependence, which shrinks the scale when flow is suppressed. Recent observations thus enter the seasonal detector through the scale and not through the mean. The scale is not the source of the seasonal prior's lead, which holds at a constant scale, and the reversal under CUSUM holds at both scales (Appendix E); on the freeway the scale lowers recall.

Table 3: Recall retained after a change of regime with thresholds recalibrated, 100 events per corpus, two excursions per day; normal models fitted on the original regime and not refitted; thresholds on ten event-free days of each regime and the realised rate on ten others (Appendix E adds delays, per-class rates and refits). The hybrid's control cell is 0.835, rounded up.

| Pairing | Original regime | Signal cycle 110 s | Demand $\times 0.85$ | Demand $\times 1.15$ |
|---|---|---|---|---|
| Seasonal prior, statistic 2/3 | 0.81 | 0.22 | 0.74 | 0.65 |
| Seasonal prior, statistic 4/4 | 0.99 | 0.04 | 0.92 | 0.95 |
| Hybrid 128, statistic 4/4 | 0.93 | 0.02 | 0.86 | 0.90 |
| Ridge 64, CUSUM | 0.62 | 0.63 | 0.81 | 0.62 |
| Hybrid 128, CUSUM | 0.84 | 0.61 | 0.85 | 0.53 |

### 3.6 Performance after a change of signal plan or of demand

Table 3 scores the pairings on corpora generated under a 110 s signal cycle in place of 90 s and under the demand profile scaled by 0.85 and 1.15, with the normal models fitted on the original regime and not refitted, and with thresholds fitted on event-free days of the shifted regime; what is measured is therefore the recall each pairing retains after threshold recalibration, and not its performance in the interval between an unannounced change and recalibration, which was not tested. Under the plan change every instantaneous pairing collapses: the seasonal prior with the statistic falls from 0.81 to 0.22 under the two-of-three rule and from 0.99 to 0.04 under four-of-four, against a same-sized control, because a shifted table leaves a periodic residual of an event's size on every event-free day and the run rule counts it. The ridge with CUSUM retains its recall, since its residual stays centred and integrable; the hybrid with CUSUM declines, because the seasonal residual it integrates carries the plan change as a periodic bias; refitting the prior on fifty shifted days recovers most of the seasonal pairing's recall (Appendix E). Under a drift of demand the seasonal prior under the run rule retains nearly all of its recall, since the level-dependent scale absorbs a proportional change, and the hybrid with CUSUM declines at the higher demand, where the level change enters its integrand as a bias. Among the tested pairings only the ridge with CUSUM retains its recall under both changes without refitting the forecasting model; adaptive signal control was not tested. A complementary normal-day transfer condition keeps the 90 s cycle and changes only the per-junction offsets: the standardised innovation scale remains 1.006, against 1.562 when the cycle changes to 120 s (Appendix E). This supports a narrower interpretation: cycle length is the tested source of plan fragility, while offset-only coordination changes preserve the normal model's calibration; event recall under the offset condition and changes of green split or topology remain untested.

**Intervention validity.** A separate route-choice test recorded no applied route changes: all 56 event trajectories equal their normal twins. Its zero selected-detector hits therefore cannot establish failed transfer to turning events. That mechanism requires a corrected intervention experiment.

### 3.7 Real data

Table 4 states which corridor findings the freeway data supports; every freeway comparison in the main text is recall interpolated at a realised two false alarms per day, and the nearest-point paired inference is confined to Appendix J. On the released month the seasonal residual detects 0.45 of the 63 events, the forecast residual 0.32 and the raw flow 0.14, at two false alarms per eight-hour morning. On the 78-day corpus with 705 logged events, under the strictly causal 120 s aggregate at the benchmark's report window, the figures are 0.155, 0.135 and 0.063: a modest descriptive advantage, carried by crashes and incidents, with no statistical significance claimed at the matched rate. The location-aware secondary analysis (fixed threshold, two alarms per day; not the interpolated matched-rate estimate above) gives the same qualitative ordering but lower absolute recalls (0.041 versus 0.016 within 0.5 miles and 0.061 versus 0.034 within 1 mile for seasonal versus forecast residuals), because it excludes alarms elsewhere on the network. The mechanism is visible at the crash's own station (Figure 2, bottom right): over the 38 located crashes the seasonal residual rises by a third from five minutes before the report and stays elevated for thirty, while the forecast residual never rises, and at matched placebo times both stay at baseline. The reversal under the cumulative sum does not transfer: after centring and scaling each flow residual on event-free tuning data before applying the CUSUM, the strictly causal corpus recalls are 0.175 for the seasonal

Table 4: What transfers to I-24, at matched realised false-alarm rates (Section 3.7 and Appendix J give the numbers).

| Corridor finding | I-24 |
|---|---|
| Seasonal residual leads the tested forecast residual under the instantaneous statistic, rule 2/3 | Supported on the released month; a modest descriptive advantage on the corpus |
| Residual profile at affected sensors consistent with forecast adaptation | Supported at the located crashes' stations, with a placebo control |
| Cumulative sum reverses the seasonal–forecast ranking | Does not transfer; the corrected CUSUM is a near tie |
| A run of consecutive exceedances gives a large gain | Does not transfer; the run rules remove the seasonal advantage on the corpus |
| A small local neighbourhood helps | Qualitative support; fine differences unresolved with 63 events |
| Level-dependent scale improves detection | Does not transfer; the scale lowers recall |

residual and 0.162 for the forecast residual. At the separately evaluated nearest operating points, the paired seasonal-minus-forecast difference is $+0.034$ [cluster bootstrap $-0.040$, $+0.111$]; this interval does not describe the interpolated difference. Nor does the run rule's gain: on the released month the seasonal residual's recall is 0.44 and 0.50 under three-of-three and four-of-four against 0.45 under two-of-three, and on the corpus the two residuals are level under both run rules (0.13 against 0.13), so a longer run rule does not reproduce the corridor's gain and removes the seasonal advantage (Appendix E).

## 4 Spatial context in the detector

A learned detector whose only variable is the sensor set tests how much spatial context helps: each node encodes its last $T$ steps of the standardised seasonal residual with a shared two-layer perceptron, attends over a permitted set of $N$ other sensors tagged with their signed road distance, and produces per-node detection, localisation and class logits; the corridor-level score is the soft maximum over nodes, so no node's prediction depends on any sensor outside its set, and the parameter count is the same at every breadth (Appendix F). Sets are chosen by rule (none; the $N$ nearest, upstream or downstream; random; all 81); training is on simulation seeds 0 to 3, early stopping on seed 4, results on seeds 5 to 9, over three to five network-initialisation seeds.

In this benchmark local spatial context helps, and the gain arrives with the first sensor (Figure 6): at an eight-minute history one neighbour raises recall from 0.77 to 0.94, a paired difference of $+0.14$ over events that holds for every initialisation seed, in the classes whose signature is not at one sensor, the frozen signal and the sensor fault; it is a finding within the benchmark, conditional on the five held-out demand realisations, and not an effect size across independent traffic days. The distinctions beyond that are small or exploratory: two neighbours give 0.95 and four to eight 0.91 to 0.92, and no breadth is claimed as optimal (Appendix G). The plane of history against breadth separates two effects: a single sensor with 32 minutes of history reaches 0.89, against 0.92 for four neighbours with eight minutes, but the four-neighbour column detects one to two minutes sooner at every history length, because a neighbour's evidence is available now and a sensor's own history later. On I-24 three to five stations beat one and nine or all 49 are no better, with the differences not individually resolved on 63 events (Appendix H).

## 5 Detectability, and reconstruction

What the tested detector can detect at this alarm budget was measured on a 610-pair slice varying severity, duration and fault magnitude over four demand regimes (Appendix K); it is that detector's envelope, not an identifiability limit: a $0.4\sigma$ fault goes from 13% zero-shot to 93% retrained. One- and two-lane obstructions placed mid-block are rarely detected (2 to 13% of episodes), because the surviving lane discharges at saturation flow, while the full closure is always detected; a closure shorter than the detection delay is missed four times in five; and the observation interval should match the signature's timescale. Reconstruction of the event-free counterfactual inverts the require-

ment (Appendix L): scored against the exact twin at the sensors an event moves, among the tested strategies the seasonal prior is the worst (error 5.1, against 4.4 for the observations themselves), because day-specific demand variation exceeds the event's effect, and a rollout of the autoregression from the last context before the alarm is the best (2.1); one autoregression used two ways, with the alarm as the switch, serves both tasks (Appendix M).

## 6 Related work

**Residual charts and forecast-then-detect.** Sequential change detection (Page, 1954; Basseville & Nikiforov, 1993; Poor & Hadjiliadis, 2009) and statistical process control established the linear case: Alwan and Roberts (Alwan & Roberts, 1988) charted ARMA residuals, Wardell et al. (Wardell et al., 1994) showed such charts can be slower than a chart on the raw observations, and Apley and Shi (Apley & Shi, 1999) derived the matched detector; Section 1 lists what this paper adds. Deep anomaly detection adopted forecast-then-detect with recurrent (Malhotra et al., 2015; Hundman et al., 2018), graph (Deng & Hooi, 2021) and attention (Xu et al., 2022) forecasters, on benchmarks and metrics since found flawed (Wu & Keogh, 2022; Kim et al., 2022; Schmidl et al., 2022).

**Traffic incident detection and forecasting.** The California (Payne & Tignor, 1978) and McMaster (Persaud & Hall, 1989) algorithms threshold occupancy without a normal model (Parkany & Xie, 2005); learned detectors evaluate on PeMS or FT-AED (Coursey et al., 2024; Sen et al., 2024). Graph networks for traffic forecasting (Li et al., 2018; Yu et al., 2018) assume spatial dependence carries predictive information; models without message passing match them (Liu et al., 2023; Anonymous, 2026; Das et al., 2024), and ours is the abnormal-regime counterpart. Synthetic controls (Brodersen et al., 2015; Cao et al., 2021) estimate what a series would have done without an intervention; the rollout of Section 5 is that construction started at the alarm.

## 7 Limitations

The corridor is one synthetic topology under fixed-time control; one change of signal plan and two drifts of demand were tested with thresholds recalibrated (Section 3.6), and neither a change of topology nor an unannounced change with the whole detector frozen. The turning experiment applied no intervention and provides no test of transfer to that mechanism. The freeway supports some findings and not others (Table 4); its labels are report times, not onsets. The location-aware secondary analysis constrains the first alarming neighbourhood to the logged milemarker; its instantaneous seasonal lead is descriptive and absolute recall is low and remains limited by the 0.6-mile station spacing and the report-time window. Corridor test results cover five held-out demand realisations with 50 events each, and every event-bootstrap interval is conditional on those realisations; the new cluster bootstrap over the five demand realisations is reported in Appendix B, but it remains conditional on those five clusters and is not a substitute for additional independent traffic days. The standardised seasonal residual's threshold-fitting spread (0.07) is several times the others'. The independent replication exposed material burden variation with 42 calibration days; the 84-day amendment passed both burden limits, but further seed blocks are needed to quantify the remaining variance.

## 8 Conclusion

Normal-model rankings depend on both the residual score and the alarm rule. With spatial aggregation fixed, the seasonal prior leads the selected hybrid under the instantaneous score, although the demand-cluster comparison does not resolve that difference; under CUSUM the hybrid leads decisively. Injected steps and clean-day accumulated-score tails explain much of this contrast: prediction attenuates sustained changes, while normal fluctuations determine how much retained evidence an integrating rule can recover. Step retention alone does not explain the nonlinear models' performance. Rankings also change with signal timing, even after threshold recalibration. I-24 supports the instantaneous ordering and adaptation mechanism, but not the CUSUM reversal, run-rule gain or benefit of level-dependent scaling. Calibration variability and the invalid turning test limit broader transfer claims. Spatial context and counterfactual reconstruction provide supporting findings within the paired benchmark.

## AI USE STATEMENT

We used generative AI tools to assist with writing, retrieval and discovery, and research ideation and execution. Writing assistance included drafting and restructuring passages, improving clarity, and editing the manuscript. Retrieval and discovery assistance included identifying related work, finding primary sources, and summarising relevant literature. Research assistance included refining hypotheses, providing feedback on experimental design, developing and checking analysis code, auditing experimental outputs, and interpreting results. AI use was limited to these roles; generative AI was not used to produce the datasets or to supply mathematical proofs.

Checks of AI-assisted work included comparing numerical claims with recorded experimental outputs, replaying saved detectors, checking analysis code and data provenance, and consulting primary sources for related work. The authors take responsibility for the final text, citations, methods, results, and claims, including material developed with AI assistance.

## REFERENCES

Layth C. Alwan and Harry V. Roberts. Time-series modeling for statistical process control. *Journal of Business & Economic Statistics*, 6(1):87–95, 1988.

Anonymous. Time versus space: A systematic empirical study of information allocation in large-scale traffic forecasting, 2026. Under review; anonymised for double-blind review.

Daniel W. Apley and Jianjun Shi. The GLRT for statistical process control of autocorrelated processes. *IIE Transactions*, 31(12):1123–1134, 1999.

Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott. Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, 9(1): 247–274, 2015.

Danni Cao, Jianjun Wu, Xianlei Dong, Huijun Sun, Xiaobo Qu, and Zhenzhen Yang. Quantification of the impact of traffic incidents on speed reduction: A causal inference based approach. *Accident Analysis & Prevention*, 157:106163, 2021. doi: 10.1016/j.aap.2021.106163.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

Kukjin Choi, Jihun Yi, Changhwa Park, and Sungroh Yoon. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE Access*, 9:120043–120065, 2021.

Austin Coursey, Junyi Ji, Marcos Quinones Grueiro, William Barbour, Yuhang Zhang, Tyler Derr, Gautam Biswas, and Daniel Work. FT-AED: Benchmark dataset for early freeway traffic anomalous event detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 15526–15549, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *AAAI Conference on Artificial Intelligence*, 2021.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2022.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Xu Liu, Yuxuan Liang, Chao Huang, Hengchang Hu, Yushi Cao, Bryan Hooi, and Roger Zimmermann. Do we really need graph neural networks for traffic forecasting? *arXiv preprint arXiv:2301.12603*, 2023.

Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using SUMO. In *IEEE International Conference on Intelligent Transportation Systems*, 2018.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *European Symposium on Artificial Neural Networks (ESANN)*, 2015.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.

Emily Parkany and Chi Xie. A complete review of incident detection algorithms and their deployment: What works and what doesn't. *New England Transportation Consortium, Report NETCR37*, 2005.

Harold J. Payne and Samuel C. Tignor. Freeway incident-detection algorithms based on decision trees with states. *Transportation Research Record*, 682:30–37, 1978.

Bhagwant N. Persaud and Fred L. Hall. Catastrophe theory and patterns in 30-second freeway traffic data: Implications for incident detection. *Transportation Research Part A*, 23(2):103–113, 1989.

H. Vincent Poor and Olympia Hadjiliadis. *Quickest Detection*. Cambridge University Press, 2009.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022.

Sai Shashank Sen et al. IncidentNet: Traffic incident detection, localization and severity estimation with sparse sensing. *arXiv preprint arXiv:2408.00996*, 2024.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. *Advances in Neural Information Processing Systems*, 34, 2021.

Don G. Wardell, Herbert Moskowitz, and Robert D. Plante. Run-length distributions of special-cause control charts for correlated processes. *Technometrics*, 36(1):3–17, 1994.

Renjie Wu and Eamonn J. Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly Transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *International Joint Conference on Artificial Intelligence*, 2018.