Change-Point Detection for Sensor-Rich Chemical Processes

George Weale
University of California, Santa Barbara
Santa Barbara, CA, USA
gweale@ucsb.edu

***Abstract*—Large chemical processes produce multivariate sensor streams whose distributions change during grade transitions, throughput changes, ambient disturbances, equipment degradation, and faults. A monitor that treats every departure from a fixed baseline as a fault creates an excessive alarm burden; one that smooths away regime changes can miss meaningful loss of control or quality. This paper develops an online change-point-detection framework for sensor-rich chemical processes. The method combines causal data handling, operating-regime context, multivariate residual features, Bayesian online run-length inference, and an alarm policy that separates statistical evidence from engineering diagnosis. Evaluation uses controlled process simulations and permissioned, de-identified industrial historian windows, with event-level detection recall, false alarms per operating hour, detection delay, calibration, and robustness to missingness and sensor drift. The resulting evidence bundle makes each alert traceable to its causal signal window, operating context, data-quality state, model version, and threshold policy.**

***Index Terms*—statistical process monitoring, change-point detection, industrial time series, Bayesian online change-point detection, multivariate analysis, fault detection**

I. Monitoring Objective

Online monitoring must identify distributional changes in multivariate chemical-process signals early enough to support operator review while maintaining a transparent alert burden across legitimate operating regimes. The formulation defines the causal data boundary, operating context, feature construction, run-length inference, alarm semantics, baseline comparisons, and event-level evaluation criteria that connect statistical evidence to an operationally meaningful alert.

Statistical process monitoring has long used multivariate projections and residuals to identify deviations [2], [1]. Industrial changes are harder than a single binary classification problem because their start times can be uncertain, their signatures can be delayed across sensors, and normal transitions can resemble faults [3], [4]. The monitor therefore evaluates timeliness and alert burden together and stratifies performance by operating context.

The system detects statistical change while keeping root-cause classification, control action, and safety diagnosis within authorized engineering processes. Each alert initiates review through a retained evidence bundle rather than acting as an autonomous plant decision.

II. Related Work and Motivation

Multivariate statistical process monitoring commonly projects many correlated measurements into score and residual spaces, then watches those spaces for departures from a reference operating condition [2], [1]. This approach is valuable because an individual measurement may remain within an ordinary range while a combination of measurements becomes unusual. Batch and continuous-process monitoring have also shown that time alignment, operating phase, and transition behavior matter to the meaning of a chart limit [4], [9]. Fault-detection research distinguishes quantitative residual generation, multivariate statistics, and later diagnosis or isolation work [10]. That distinction matters here. A change point is not necessarily a fault: it can be a legitimate throughput transition, a known setpoint change, a sensor-maintenance event, or a disturbance requiring review. Treating every departure as a fault inflates the apparent sensitivity of a detector while hiding the practical cost of false alarms. Classical CUSUM methods and abrupt-change detection provide useful baselines because they make the trade between timeliness and alert burden explicit [6], [7]. Bayesian online change-point detection offers a natural way to retain uncertainty about the current run length rather than force each sample into a fixed normal-or-fault label [5]. The contribution is a chemical-process monitoring framework that combines causal data handling, context-conditioned features, a persistent-alert rule, and event-level metrics. It evaluates whether an alert is timely and useful to an operator, not only whether an offline classifier separates hand-labeled samples. Figure 1 illustrates multichannel regime-change evidence and transient uncertainty.

III. Data Contract and Causal Preprocessing

A. Time-Series Boundary

At timestamp t, let $\boldsymbol{x}_t \in \mathbb{R}^p$ contain process measurements and $\boldsymbol{u}_t \in \mathbb{R}^q$ contain known setpoints or manipulated variables. Each channel has a tag identifier, engineering unit, timestamp source, measurement type, physical range, and quality flag. The monitor must use only information available at or before t; centered rolling windows and future-value interpolation are forbidden in online evaluation.

Data are first resampled to a declared causal grid. A missingness indicator $m_{t,j}$ is retained for each sensor j. Short gaps may be forward-filled only when the sensor physics and update interval justify the operation; longer gaps produce an explicit masked feature rather than a plausible invented value. Scaling parameters are estimated only on a designated reference segment, using either mean and standard deviation

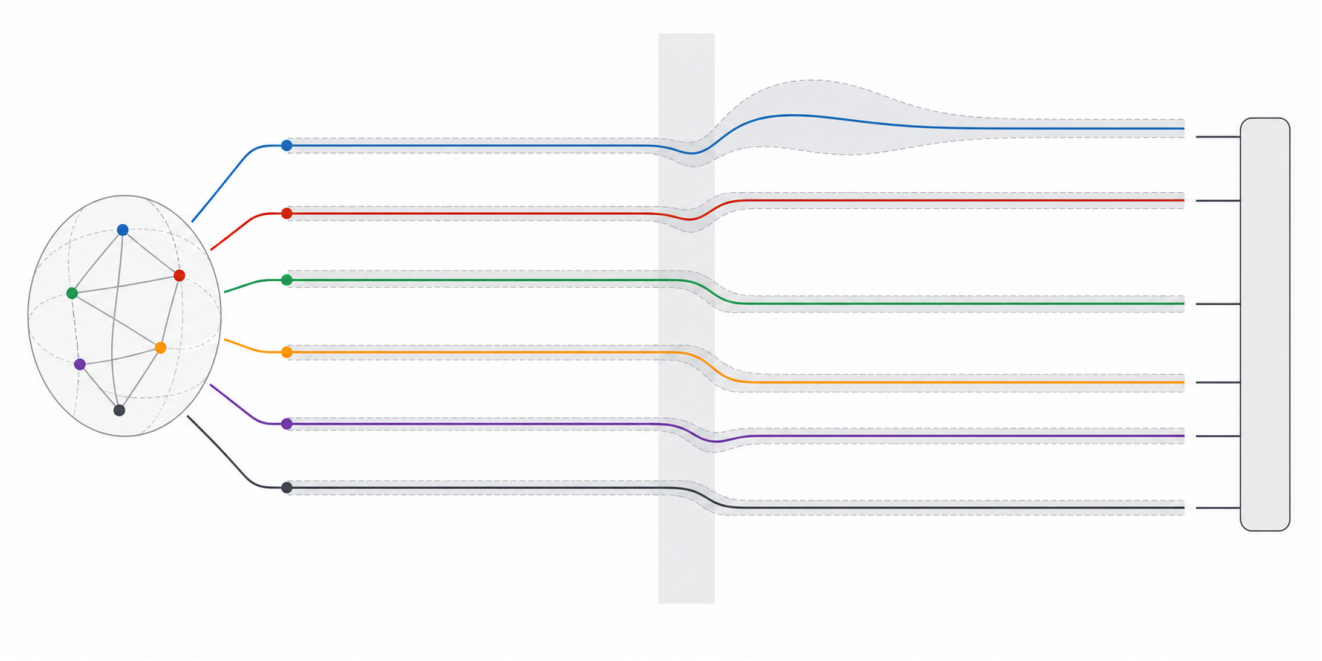

Fig. 1. Multichannel regime-change evidence. Sensor trajectories shift near the shaded interval while uncertainty expands transiently for one channel.

or robust median-based alternatives. The chosen estimator and the reference segment must be independently documented as an intended normal regime or reported as an assumed baseline.

B. Operating Context

Known operating context reduces the chance that a commanded grade or throughput transition is misread as an unexplained fault. Let c_t include available mode labels, setpoint bands, production rate, and relevant utility state. If labels are absent, a context estimator may cluster a reference period, but its labels are treated as latent approximations rather than ground truth. Context must be held fixed or estimated causally during a test window; it may not be reconstructed after reviewing the event labels.

IV. Feature and Change-Point Model

A. Multivariate Evidence Features

The detector operates on a causal feature vector $\boldsymbol{z}_t$ rather than raw values alone:

$$\boldsymbol{z}_t = [\tilde{\boldsymbol{x}}_t,\ \tilde{\boldsymbol{x}}_t - \tilde{\boldsymbol{x}}_{t-1},\ \boldsymbol{r}_t,\ \boldsymbol{m}_t,\ c_t]\,. \tag{1}$$

The residual vector $\boldsymbol{r}_t$ may be obtained from a context-conditioned principal-component model, a state-space model, or another documented forecaster. For a principal-component baseline with loading matrix P_k, score and residual spaces are computed from the reference-scaled signal. The Hotelling score statistic and squared prediction error are retained as interpretable baseline features, not assumed sufficient for all changes [2].

Feature selection is frozen before test evaluation. If a model uses a learned representation, its fit set, update policy, missing-data behavior, and random seed must be retained in the experiment manifest.

B. Bayesian Online Run-Length Inference

Let r_t denote the number of observations since the most recent change point. Following the online change-point formulation of Adams and MacKay [5], the recursion updates a joint run-length distribution:

$$p(r_t, \boldsymbol{z}_{1:t}) = \sum_{r_{t-1}} p(r_t \mid r_{t-1}) \tag{2}$$

$$\times\, p(\boldsymbol{z}_t \mid r_{t-1}, \boldsymbol{z}_{t-r_{t-1}:t-1}) p(r_{t-1}, \boldsymbol{z}_{1:t-1}). \tag{3}$$

The hazard function $H(r)$ specifies $p(r_t = 0 \mid r_{t-1}) = H(r_{t-1}+1)$ and must be set from a declared expected regime duration or varied in sensitivity analysis. The predictive model may be a context-conditioned multivariate Student-t distribution with shrinkage covariance; a full unrestricted covariance is not used when the effective sample size cannot support it.

The posterior change evidence is

$$\gamma_t = p(r_t = 0 \mid \boldsymbol{z}_{1:t}). \tag{4}$$

An alert is emitted only if $\gamma_t \geq \tau$, the condition persists for d causal samples, and an alert cooldown has expired. The threshold τ, persistence d, and cooldown are selected on a validation set or set by an explicitly declared false-alarm target. They are not tuned on the final test events.

TABLE I
ALERT POLICY PARAMETERS THAT MUST BE DECLARED BEFORE FINAL TESTING

Parameter	Meaning
τ	Posterior change-evidence threshold in Eq. (4)
d	Consecutive causal samples required before alerting
Cooldown	Minimum time between alerts in one context segment
Hazard model	Expected regime duration or sensitivity grid
Missing-data rule	Permitted gap length and feature-mask behavior
Escalation rule	Human recipient and required evidence bundle

V. ALARM SEMANTICS AND EVIDENCE BUNDLE

An alarm record contains: event time; posterior evidence trajectory; context; contributing residual and raw-signal features; data-quality flags; model version; threshold policy; and a pre-event baseline window. Feature contribution may be reported as a normalized residual ranking or as a local likelihood contribution, but it must be labeled *diagnostic evidence*, not causal proof.

The triage labels are REVIEW, EXPECTED TRANSITION, INSTRUMENT CONCERN, PROCESS CONCERN, and UNRESOLVED. They are assigned through a reviewed annotation process. The model's statistical alert and a later engineering classification are stored separately to prevent a model output from being mistaken for an operator finding.

VI. EVALUATION DESIGN

A. Data Sources and Splits

The first evaluation tier uses a controlled benchmark, such as the Tennessee Eastman process simulation [3], augmented with documented operating transitions and sensor-dropout tests. Simulation supplies known disturbance timing and supports repeatable stress cases. A second tier uses de-identified historian windows only with permission, an approved data-governance boundary, and an annotation process that prevents label leakage.

Train, validation, and test splits are made by complete run, campaign, or contiguous operating window. Randomly shuffling timestamps is prohibited because it leaks temporal structure and lets adjacent observations appear in both fitting and evaluation. All detector comparisons receive the same causal preprocessing and test windows.

B. Baselines and Ablations

The minimum baselines are: univariate cumulative-sum (CUSUM) charts [6]; principal-component monitoring using T^2 and SPE; a fixed-window likelihood-ratio detector; and Bayesian online change-point detection with a simple independent emission model. The full model is ablated by removing context, residual features, missingness indicators, and persistence logic. A more complex representation is retained only if it improves fixed evaluation outcomes without materially worsening calibration or alert burden.

TABLE II
EVALUATION SAFEGUARDS AGAINST OVERSTATED PERFORMANCE

Risk	Required safeguard
Temporal leakage	Contiguous run-level splits and causal features only
Threshold overfit	Tune on validation runs; lock before test evaluation
Ambiguous labels	Retain annotator rationale and unresolved category
Easy simulated faults	Include transitions, drift, missingness, and small shifts
Metric gaming	Report false-alert burden and delay beside event recall

For a reference event interval $[t_e, t_e + w]$, an alert counts as a detection if it occurs within the interval and is not already matched to an earlier event. Metrics are

$$\mathrm{Recall}_{\mathrm{event}} = \frac{\#\text{detected events}}{\#\text{reference events}}, \tag{5}$$

$$\mathrm{Precision}_{\mathrm{event}} = \frac{\#\text{matched alerts}}{\#\text{all alerts}}, \tag{6}$$

$$\mathrm{Delay}_e = \min\{t - t_e : \mathrm{alert}(t) = 1\}, \tag{7}$$

with delay reported only for detected events. The report also includes false alerts per operating hour, context-stratified recall, posterior calibration, and performance under artificial missingness and sensor bias. Aggregate accuracy is excluded because an imbalanced stream can make a detector that never alerts appear accurate.

VII. ANNOTATION AND OPERATOR-FACING DECISION POLICY

Change-point evaluation depends on an event definition that is more careful than a binary fault label. The annotation record contains an earliest plausible onset, a latest plausible onset, the available context, a reviewer rationale, and a confidence category. A transition that was commanded and documented can be labeled expected even when its sensor pattern is abrupt. A change with competing explanations remains unresolved rather than being forced into a fault class. This preserves ambiguity instead of converting it into artificial training certainty. Each alert is reviewed through a compact evidence bundle: the causal signal window, context variables, missingness flags, posterior trajectory, feature ranking, threshold version, and prior alerts during the cooldown. The reviewer may mark the evidence insufficient. That outcome reveals a data-quality or observability limitation rather than a detector failure that can be fixed by lower thresholds. Annotation decisions remain distinct from model outputs so that later reviewers can assess whether the model merely reproduced a label convention. An operator-facing policy maps alert class to a bounded action. A REVIEW alert requests inspection of the evidence bundle without changing a controller. Repeated low-confidence alerts may trigger an instrumentation-health check. Only an authorized plant procedure specifies escalation to a control-room response. This boundary supports evaluation of timeliness and interpretability while preserving process-safety analysis and human responsibility.

A. Latency, Data Quality, and Calibration

Online usefulness is constrained by the data path as much as by the detector. The evaluation record includes sensor scan interval, timestamp source, transport delay, resampling delay, and maximum tolerated late-arrival time. A detector evaluated on perfectly ordered data can appear faster than an implementation that waits for historian aggregation. Event delay therefore has two components: statistical detection delay measured from the causal input stream and end-to-end notification delay measured from the available operational timestamp. The two are not interchangeable. Data-quality conditions are tested explicitly. The benchmark injects short gaps, stale readings, gradual bias, and isolated spikes under a fixed schedule. The monitor either preserves an alert decision with a quality flag, abstains, or marks the bundle insufficient; it does not silently convert invalid measurements into a confident probability. Calibration is assessed by grouping posterior evidence into bins and comparing the frequency of annotated change events within those bins while retaining uncertainty in human event labels. The report includes a failure atlas rather than only an aggregate metric table. Each representative false alert, missed event, delayed detection, and unresolved event has a compact causal trace with the selected context and data-quality state. This practice supports improvement of the representation and annotation boundary without hiding adverse cases behind an average. It also ensures that a claimed improvement can be judged against the operational objective that motivated it.

VIII. Benchmark Design and Falsifiability

The benchmark locks the process simulation version, sampled variables, fault and transition scenarios, timestamp grid, reference-event intervals, missing-data perturbations, and train-validation-test windows before detector outputs are reviewed. It includes routine grade or throughput changes, gradual sensor bias, short missing-data bursts, and small persistent disturbances. These cases distinguish sensitivity to any change from useful discrimination between expected and unexpected changes. A permissioned second tier replays de-identified historian windows in shadow mode. The model receives only signals and context available online. Independent reviewers annotate whether an alert coincides with a known transition, an instrument issue, a process concern, or an unresolved event. This tier measures alert burden, delay, evidence completeness, and reviewer disagreement without post-event label leakage. The monitor loses its operational case if it produces materially more false alerts per operating hour than a simpler baseline at comparable event recall; if a context-free model performs equally well; if posterior calibration fails across ordinary transitions; or if feature-attribution records cannot be reconciled with the raw data bundle. Such outcomes favor a simpler charting method or a narrower use case and do not justify retuning the test threshold after the result is known.

IX. Limitations and Safety Boundary

No statistical monitor can infer a physical root cause from covariance change alone. Sensor drift, poor calibration, feed variation, and faults can produce overlapping patterns. Rare dangerous events may be underrepresented, while an old reference period may not represent current normal operation. Bayesian posterior values also depend on the hazard and emission model; they are model-based evidence, not literal probabilities of a process fault.

The tool supports a broader monitoring workflow by making changes inspectable, time-stamped, and comparable under a fixed decision policy. Shutdown, safety action, and quality release remain governed by authorized plant procedures. Deployment requires cybersecurity review, alarm-management integration, human-factors testing, and accountable process-safety governance.

Data Governance and Reproducibility Record

Historian data are treated as governed evidence rather than a generic machine-learning corpus. The experiment manifest identifies the approved data boundary, de-identification treatment, tag dictionary, engineering units, timestamp source, sampling policy, retention period, and access roles. Raw historian extracts remain immutable and separately access-controlled; derived feature sets carry a hash of the extraction query, preprocessing revision, and reference-window identifiers. This permits a reviewer to establish which observations were available to the detector without exposing unnecessary operational detail. Each experiment run retains the causal preprocessing configuration, feature definition, context policy, model revision, random seed, hazard setting, threshold, persistence rule, cooldown, and clock used for delay calculation. Annotation records include reviewer role, event interval, rationale, confidence, and whether the label was available before or after the alert time. These artifacts distinguish a genuine online result from a post hoc reconstruction that has benefited from future information or a revised label convention. The audit packet contains representative successes and failures in equal form: a causal trace window, data-quality state, posterior evidence, alert decision, and any human review outcome. Sensitive tag names may be replaced by stable aliases, but the alias map and units remain available to authorized reviewers. A model-change log records retraining, threshold changes, and added context variables; each new version is evaluated against a locked comparison set rather than merged silently into prior evidence. This preserves traceability while keeping the human authorization boundary intact.

X. Conclusion

This paper frames change-point detection as a rigorous online-monitoring problem rather than a generic anomaly-classification exercise. It integrates causal preprocessing, multivariate evidence features, run-length inference, interpretable alarm records, and event-level validation. Context-aware alerts, explicit latency accounting, fixed baselines, and a balanced failure atlas make detection timeliness and alert burden comparable within one auditable framework.

REFERENCES

[1] S. J. Qin, "Statistical process monitoring: Basics and beyond," *Journal of Chemometrics*, vol. 17, no. 8–9, pp. 480–502, 2003, doi: 10.1002/cem.800.

[2] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979, doi: 10.1080/00401706.1979.10489779.

[3] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Computers & Chemical Engineering*, vol. 17, no. 3, pp. 245–255, 1993, doi: 10.1016/0098-1354(93)80018-I.

[4] P. Nomikos and J. F. MacGregor, "Multivariate SPC charts for monitoring batch processes," *Technometrics*, vol. 37, no. 1, pp. 41–59, 1995, doi: 10.1080/00401706.1995.10485888.

[5] R. P. Adams and D. J. C. MacKay, "Bayesian online changepoint detection," arXiv:0710.3742, 2007. [Online]. Available: https://arxiv.org/abs/0710.3742

[6] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1–2, pp. 100–115, 1954, doi: 10.1093/biomet/41.1-2.100.

[7] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[8] D. C. Montgomery, *Introduction to Statistical Quality Control*, 8th ed. Hoboken, NJ, USA: Wiley, 2020.

[9] T. Kourti and J. F. MacGregor, "Process analysis, monitoring and diagnosis, using multivariate projection methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 28, no. 1, pp. 3–21, 1995, doi: 10.1016/0169-7439(95)80036-9.

[10] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis. Part I: Quantitative model-based methods," *Computers & Chemical Engineering*, vol. 27, no. 3, pp. 293–311, 2003, doi: 10.1016/S0098-1354(02)00160-6.