Author response to Referee 1 (RC1)

All line numbers in this document refer to the revised manuscript.

RC1

Thank you for providing feedback and insight. We hope to have addressed your comments in the response below, and have made some revisions to highlight the questions you posed.

RC1-1

Comment: It seems that the authors are introducing synthetic anomalies. Then, they perform a threshold selection for the information-based rule and the XGBoost. It is very important to clearly state what data has been used for training, what data for validation/grid search, and for testing.

Author response: We agree with the reviewer and believe that the reorganization and revisions state the data roles more explicitly in the reviewer's terms. Tuning (training): both methods use the entire 880 events, with the rule-based equation grid search (8,400 parameter combinations; ranges and increments now fully specified at L338–347), and the XGBoost models trained on the same benchmark. No data were held out (stated at L348–349 and L408). Validation: XGBoost model and threshold selection used five-fold cross-validation within the benchmark (~80/20 folds, L402–408); the equation's grid search used no separate validation split, which we now state rather than imply. Testing: independent evaluation comes from the continental-scale application (Section 5, L599 ff.), in which the frozen configurations were applied to all reporting devices across the United States and Canada and candidate detections were manually validated. We deliberately chose this disclosure over constructing a synthetic hold-out: a held-out fold would share the fault-generation process with the training data and would overstate independence, whereas the continental deployment tests the tuned configurations on a large corpus of real data, where aside from devices excluded due to QA/QC, faults were not known beforehand. The in-sample status of all synthetic-benchmark results is now flagged where those numbers appear (Table 3 caption; Section 4.3).

Changes in manuscript: Scoping statements added at L348–353 (grid search) and L408 (XGBoost); labelling process stated at L394–395; sweep ranges and increments completed at L338–347; caveats added to the Table 3 caption and Section 4.3 (L591–597); the evaluation protocol for the independent continental test is described in Section 5.1 (L599–609).

RC1-2

Comment: The authors should clearly explain the difference/benefits of the two methods.

Author response: We agree this needed to be more apparent to the readers. The reorganization now provides one in a dedicated comparison section. The difference is one of design intent and evaluation target, and the revision states both explicitly. The rule-based equation is a precision instrument: an interpretable flag for categorical (binned AQ ranges) differences that, through its network entropy limit, defers judgement when the network itself is disordered (Section 4.3, L593–595). The XGBoost classifier is a recall instrument trained on a broader target, every

device-hour regardless of AQI category, and detects signals too subtle to register categorically, addressing the two structural limitations of the equation stated at L355–359. Section 4.3 reports the comparison symmetrically on both targets (F_1 0.88 vs 0.42 on the broad target; 0.45 vs 0.87 on the measurable-outlier target, L577–582) and states that the two sets of scores are never combined (L588–589). The benefits of each are given with their costs: the model's added coverage depends on synthetic training data, in-sample evaluation, and the equation's flag among its features, and it was trained on only the five designed fault types (L595–598); Section 6.4 (L905–908) states that the XGBoost models are deliberately naive benchmarks kept in service of the tuneable equation, which is the focus of the study.

Changes in manuscript: XGBoost methods consolidated into Section 3.5, with the motivating limitations at L355–359; new Section 4.3 (Method Comparison, L574–598) with the two-targets statement and complementary-roles close; new Table 3; scope disclaimer in Section 6.4 (L905–908).

RC1-3

Comment: The information-theoretic approach should be compared with other simpler methods, e.g., one of their criterion, the absolute value of the difference between the sensor reading and the average of the neighborhood values.

Author response: Agreed, and the revision adds in a comparison. The reviewer's suggested baseline (the absolute difference between a device's reading and the average of its neighbourhood) is now the first of 3 baseline neighbour-comparison rules defined in (new) Section 3.6 (L418–422), alongside the neighbour z-score and the median/MAD robust z-score, each computed on the same networks and hourly measurements as our methods. The standardized rules were evaluated at their conventional cutoffs (3, the three-sigma rule as applied by Shekhar et al., 2003; and 3.5, the Iglewicz and Hoaglin, 1993 recommendation for the modified z-score) and all three additionally at the cutoffs that maximized their performance on our benchmark (L424–427), against a common, method-independent outlier definition (L428–432). Section 4.1 reports the outcome under both objectives (L514–528): under the precision-weighted objective the tuned equation leads ($F\beta{=}0.5 = 0.797$ versus 0.778 for absolute neighbour deviation), while under F_1 the ranking inverts and absolute deviation wins (0.743) at three- to four-fold higher false-positive rates. The separation is sharpest at the smoke boundary (L529–537; Table S4): the equation's false-positive rate rises modestly (0.57% to 1.30%) while absolute deviation rises roughly 25x in count (44 to 1,102). We also state where the baseline rule wins outright, proportional drift (fault type 4), where a concentration-difference rule is built for exactly that error (L535–537). Full operating points and smoke-stratified results for every method are given in Tables S3 and S4.

Changes in manuscript: New Section 3.6 (L418–440); baseline-comparison paragraphs in Section 4.1 with the label bridge (L514–537); smoke-stratified false-positive comparison in Section 4.1.3; supplement Tables S3 and S4.

RC1-4

Comment: Why have the authors chosen the XGBoost for binary classification? They should motivate their choice and compare it with some other well-known methods.

Author response: The motivation for XGBoost is now consolidated in Section 3.5 (L363–372): gradient-boosted decision trees are a well-established class for tabular classification (Friedman, 2001; Bentéjac et al., 2021), and they fit the structure of this problem specifically. On comparing with other well-known methods: we did not add a benchmark against other ML classifiers, and we want to explain. This is not an ML-comparison study, and the focus is the information-theoretic features and how they can be applied, with a well-known ML classifier (XGBoost) serving as a supporting benchmark. Section 6.4 states this scope explicitly (L905–908): the models are deliberately naive, trained on synthetic faults, and a properly annotated model trained on real outliers could plausibly outperform every method presented here. Benchmarking additional classifiers against the same synthetic labels would expand that provisional comparison without changing any conclusion of the paper.

Changes in manuscript: Motivation paragraph consolidated in Section 3.5 (L363–372); new Table 3 with comparisons on the model's training target; scope statement in Section 6.4 (L905–908).

RC1-5

Comment: What is the labelling process for the supervised binary classification?

Author response: The reviewer is right that the labelling process was not stated explicitly enough at first. In the original manuscript, the training protocol (class imbalance of ~1% positive, ~1:90; proportional minority-class weighting; five-fold cross-validation) was described, but the definition of a positive label was implied rather than given. We have added it in Section 3.5.2 (L394–395): labels follow directly from the fault-injection process. Every device measurement-hour within a perturbation window (was perturbed) is a positive example, every other device measurement-hour a negative one, with no manual annotation involved. This training target is intentionally broader than the measurable outlier definition used to evaluate the rule-based equation, a distinction now carried through the results (Section 4.3, L588–589; Table 3 caption). The pre-existing protocol details remain at L398–412.

Changes in manuscript: Labelling definition and training-target distinction added in Section 3.5.2 (L394–395); the distinction is restated where scores are reported (Section 4.3, Table 3 caption). The imbalance and cross-validation description at L398–412 is unchanged from the original submission.

RC1-6

Comment: What is the effect of a sparse sensor network? Are the anomalies also identifiable in the presence of smoke? During smoke events, are the outliers identifiable only in the case of a dense network?

Author response: These questions are the motivation of the paper, and the revisions made aim to reorganize the existing evidence and add new results that tackle that question more directly. Regarding sparsity, this was quantified in the original submission and remains in Section 5.3 (L710–718). The local smoke events with poor misclassification rates occur in networks with roughly half the neighbour density of other smoke conditions (12.8 vs 23.6 average neighbours per device at 7500 m). Events flagged only by the 12500 m networks misclassified at higher rates than those caught at 7500 m (50% vs 0%), indicating that reliable classification depends on

nearby context, aka sensory density. Neighbours added at distance may introduce uncertainty rather than confidence. Network construction also imposes a density minimum, so it is tuneable (thresholds chosen so at least half of devices have ≥3 neighbours; Fig. S1). We note these rates rest on a small sample (n=3 events flagged only at 7500 m, n=10 only at 12500 m, and n=32 local smoke events overall) and on the four parameter sets (Table S5), so we treat them as indicative rather than definitive. The manuscript states this relationship warrants further investigation with a larger validation dataset (L729). On identifiability in the presence of smoke, yes, and the revision adds more evidence. Smoke-stratified performance for all methods (Table S4) shows the equation's false-positive rate nearly stable across the smoke boundary (0.57% to 1.30%) and the XGBoost model's essentially flat (0.26% ambient, 0.25% in smoke), where the comparison outlier methods degrade. On whether outliers are identifiable only in dense networks, no, but rather than guess in sparse or disordered conditions, the framework is designed to defer. The conditional nature of the rule-based equation, specifically on entropy, withholds classification when the network itself is disordered, and the revision now states that the same deferral can be tuned temporally through persistence or through a minimum neighbour consensus (L720–725). New Section 6.2 (L827–838) makes the underlying structure more explicit. Wide/long transported plumes disperse and produce strong network consensus and are the easier case, while local smoke, impacting only part of a network or a single sensor, is by construction the most disordered and hardest case for any snapshot-based consensus method, which is precisely the regime this framework was designed to address.

Changes in manuscript: Smoke-stratified evaluation added (Table S4; Section 4.2 flat-FPR statement at L591–592); deferral framing added at L720–725; new Section 6.2 paragraph (L827–838) with Zhang (2025) and Miller (2019); Kelp et al. (2022) cited for sparse regulatory coverage (L39); sparsity analysis of Section 5.3 (L710–718) and the density minimum with Fig. S1 (L187) carried from the original submission.

RC1-7

Comment: Is the method proposed optimal or nearly-optimal for all scenarios, i.e., smoke and non-smoke events?

Author response: No, it is designed to be stable and tolerant across non-smoke and smoke events, but we cannot definitively say it is optimal. This is why the framework is described as tuneable. Empirically, no single parameter set was optimal for more than one fault type, and the optimal spatial proximity varied across the full tested range (Section 4.1, L483–496; Table 2 lists the per-fault optima). The parameters do cluster, indicating partial generalizability rather than a universal optimum. What does hold across smoke and non-smoke conditions is stability of the error behaviour. The equation's false-positive rate moves only from 0.57% to 1.30% across the smoke boundary and the XGBoost model's is essentially flat (Table S4; L591–592), so no scenario-specific retuning is required to keep false alarms controlled. The revision states that the choice of parameters and persistence thresholds depends on the application (Section 6.1, L776–790).

Changes in manuscript: Per-fault optima and parameter clustering reported in Section 4.1 (L483–496, Table 2); smoke-boundary stability added (Table S4; L591–592); Section 6.1 (L776–790) added to state the tuning levers and application trade-offs explicitly.

RC1-8

Comment: Is the method intended for large anomalies, e.g., lasting more than one hour? What if there is just a single measurement that deviates from normality? Is it flagged? These scenarios are also relevant in cases where data is fed to downstream applications.

Author response: The methods operate per-hour by design, but we show how persistence (anomalies lasting more than one hour, for n consecutive hours) greatly reduces the false positives. This would be a key consideration for how to implement this method. For example, spatially resolved surface products that interpolate between sensors may want to disregard noisy results when it's a single sensor, whereas public informational sites may want to retain anomalous sensors for a few hours to see if they are picking up a real signal.

In its most basic form, we use each hour as an independent snapshot (L355–359), so there is no minimum anomaly duration, and a single deviating measurement is assessed and can be flagged in that hour. The practical question is how much confidence a single-hour flag deserves, and the revision quantifies this with a duration-stratified validation (Section 5.2, L641–665; Table S6, supplement). Single-hour flags are noisy (31–51% misclassification), improving to 6% for the information-theoretic method (15% for XGBoost) by 3 consecutive hours, and to 95–98% genuine outliers beyond 13 hours.

Notably, many brief flags are not failing sensor errors. Over half of the 2–5 hour events were validated as genuine "hyper-local" air quality phenomena rather than sensor faults.

For downstream applications, this is why persistence is implemented as an optional downstream filter rather than built into detection (L720–725). An application can consume raw hourly flags, or apply a persistence threshold matched to its cost of false alarms (Section 6.1, L776–790).

Changes in manuscript: Duration-stratified validation reported in Section 5.2 (L641–665) with Table S6 (supplement); the per-hour snapshot basis stated at L355–359; persistence framed as a tunable downstream filter (L720–725); application-matched operating points in Section 6.1 (L776–790).

Additional changes (author-initiated)

While revising Section 4.2, we identified an error in the originally reported XGBoost false-positive rate of 15.2%. That value was computed from the wrong column of an internal results table and does not describe the classifier's performance. Recomputed over the full evaluation dataset against the model's stated training target (every device-hour within a perturbation window), the false-positive rate is 0.26% at the 0.95 probability threshold. The corrected value appears in Table 3 (L571–572), which replaces the affected prose in Section 4.2; no other reported quantity was affected.

Additional changes in the final revision round, following co-author review: the device descriptions of Sect. 2.2 were condensed to a summary with full detail moved to the supplement (Sect. S1 and Table S1); the standalone false-positive-rate figure was consolidated into supplement Table S4, with the key values retained in Sect. 4.1.3; the duration-stratified validation table moved to the supplement as Table S6; and the local-smoke case study moved to the supplement as Fig. S5, reducing the main text to 10 figures and 3 tables. Two corrections

were also made: the neighbour-density values in Sect. 5.3, previously mislabelled as percentages, are mean neighbour counts per device (12.8 versus 23.6 at 7,500 m).