Author response

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 L340–349), and the XGBoost models trained on the same benchmark. No data were held out (stated at L350–351 and L410). Validation: XGBoost model and threshold selection used five-fold cross-validation within the benchmark (~80/20 folds, L404–410); 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, L602 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 L350–355 (grid search) and L410 (XGBoost); labelling process stated at L396–397; sweep ranges and increments completed at L340–349; caveats added to the Table 3 caption and Section 4.3 (L594–600); the evaluation protocol for the independent continental test is described in Section 5.1 (L602–612).

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, L596–598). 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 L357–361. 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, L580–585) and states that the two sets of scores are never combined (L591–592). 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 (L598–601); Section 6.4 (L908–911) 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 L357–361; new Section 4.3 (Method Comparison, L575–601) with the two-targets statement and complementary-roles close; new Table 3; scope disclaimer in Section 6.4 (L908–911).

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 (L420–424), 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 (L426–429), against a common, method-independent outlier definition (L430–434). Section 4.1 reports the outcome under both objectives (L516–530): 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 (L532–540; 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 (L538–540). Full operating points and smoke-stratified results for every method are given in Tables S3 and S4.

Changes in manuscript: New Section 3.6 (L420–442); baseline-comparison paragraphs in Section 4.1 with the label bridge (L516–540); 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 (L365–374): 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 (L908–911): 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 (L365–374); new Table 3 with comparisons on the model's training target; scope statement in Section 6.4 (L908–911).

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 (L396–397): 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, L591–592; Table 3 caption). The pre-existing protocol details remain at L400–414.

Changes in manuscript: Labelling definition and training-target distinction added in Section 3.5.2 (L396–397); the distinction is restated where scores are reported (Section 4.3, Table 3 caption). The imbalance and cross-validation description at L400–414 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 (L708–721). 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 (L732). 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 (L723–728). New Section 6.2 (L830–841) 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 L594–595); deferral framing added at L723–728; new Section 6.2 paragraph (L830–841) with Zhang (2025) and Miller (2019); Kelp et al. (2022) cited for sparse regulatory coverage (L39); sparsity analysis of Section 5.3 (L708–721) and the density minimum with Fig. S1 (L189) 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, L485–497; 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; L594–595), 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, L779–793).

Changes in manuscript: Per-fault optima and parameter clustering reported in Section 4.1 (L485–497, Table 2); smoke-boundary stability added (Table S4; L594–595); Section 6.1 (L779–793) 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 (L357–361), 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, L644–670; 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 (L723–728). An application can consume raw hourly flags, or apply a persistence threshold matched to its cost of false alarms (Section 6.1, L779–793).

Changes in manuscript: Duration-stratified validation reported in Section 5.2 (L644–670) with Table S6 (supplement); the per-hour snapshot basis stated at L357–361; persistence framed as a tunable downstream filter (L723–728); application-matched operating points in Section 6.1 (L779–793).

RC2

RC2-1

Comment: However, it was difficult to evaluate this manuscript as it is not well organized and bloated with figures and datasets outside of typical paper organization (i.e., methods, results, discussion). There are interesting analyses in this manuscript but a reorganization of the paper is needed… [and later:] The organization of this paper is a bit messy. We are already on section 4 and it seems to be introducing new methods… I would rewrite all methods to be in the methods section. / On section 5, and new data sources and scope is introduced. I think this manuscript needs a large reorganization. It contains many interesting analyses but it is long-winded and not effectively organized. The Discussion section is introducing even more figures and analysis. There are 17 figures in this paper, which can easily be cut down to half this number given that many of the figures do not add much substantive information or could be condensed more.

Author response: The revision undertakes the reorganization requested. All methods now live in Section 3 (network construction 3.1–3.2; information theory 3.3; tuning 3.4; XGBoost 3.5;

baseline comparison methods 3.6; continental-scale evaluation 3.7), results are split into the synthetic benchmark (Section 4) and the continental-scale evaluation (Section 5), and the Discussion no longer introduces new analyses. No new data sources are introduced mid-paper. On figures: the submitted 17 are now reduced to 12 in the main text, and reference material has moved to a new supplement. We also note a tension between the feedback received to shorten the manuscript while also providing additional information, comparisons, and clarification. We resolved it by reducing the main-text figure count, moving reference material to the new supplement, and condensing the text where we were able, while the additions the reviewers requested necessarily contributed some length.

Changes in manuscript: Full restructure as described; new supplement (Figs. S1–S5, Tables S1–S6); original Figure 9 deleted; figure count 17 → 10 main text.

RC2-2

Comment: "While many parts of North America, particularly large population centres, are covered by regulatory PM2.5 monitoring networks, these stations are sparse in smoke-prone regions." → This probably needs a citation, perhaps use: Kelp, M., S. Lin, J.N. Kutz, and L.J. Mickley (2022). A new approach for optimal placement of PM2.5 air quality sensors: case study for the contiguous United States, Env. Res. Letters, 17, 034034, DOI: 10.1088/1748-9326/ac548f. Otherwise, doesn't the relatively long correlation length-scale of smoke and the fact that populated centers are monitored cover most of our bases in terms of population smoke exposure?

Author response: Added as suggested. Kelp et al. (2022) is now cited at exactly that sentence in the Introduction (L38–40).

Changes in manuscript: Kelp et al. (2022) cited in the Introduction; reference added to the bibliography.

RC2-3

Comment: 2nd paragraph: but doesn't deSouza et al 2021 show that these low cost sensors are deployed in largely less socioeconomically diverse locations? And perhaps also in less optimal locations as well?

Author response: The reviewer's point is well taken, and it is part of our motivation. The reference intended is, we believe, deSouza and Kinney (2021) for the demographic concentration of PurpleAir sensors and Kelp et al. (2022) for "optimal" locations, w.r.t. capturing the extent and timing of smoke. We also cite Choi and Hummel, 2026 in this revision, which tackles the optimal sensor placement question.

Changes in manuscript: None required — deSouza and Kinney (2021) cited at L69 and L182 since submission; Kelp et al. (2022) at L39; Choi and Hummel (2026) cited in Section 3.3.1.

RC2-4

Comment: "Even if the device is performing as intended, publicly operated sensors may produce aberrant measurements if they are sited or used with the intent to capture hyper-local conditions of interest…" → Feel as though this needs a citation, or perhaps I am not

understanding what is meant… Or perhaps this is a motivating point of your study? In any event, I think slightly more clarification is needed.

Author response: This is a motivating point of the study, and we believe the original sentence lacked enough details. The passage has been rewritten (Introduction, L58–66): privately operated sensors may produce aberrant measurements because, although performing as intended mechanically, they are sited incorrectly, exposed to an emission source at close range (backyard barbecue), or used deliberately to capture conditions of interest to those who deploy them (industrial/agricultural emission source). We define these as hyper-local conditions: impacts that may influence a single sensor. To the reviewer's specific questions, siting intent of a consumer grade sensor is generally not public knowledge.

Changes in manuscript: Introduction passage rewritten with the hyper-local definition and examples (L58–66); the definition is used consistently in validation (Section 5.1, L626) and results (Section 5.2).

RC2-5

Comment: 2.1: would be helpful to have in the SI a list of the different types of sensors.

Author response: The supplement now includes Table S1, summarizing each device category used in the study: data source, typical instruments, measurement principle, and the correction or quality control applied, consolidating the descriptions of Section 2.2 that have now also been moved to the SI.

Changes in manuscript: New Table S1 (device category summary) added to the supplement.

RC2-6

Comment: Also, no discussion of differing signal-to-noise detection ability has been mentioned. Why are regulatory grade monitors treated the same as low cost sensors? → I see that they are discussed later, probably can mention so. But the latter part of my question remains.

Author response: Devices of all types are grouped together into networks in this study, although the devices included have been calibrated or corrected to be intercompared broadly (PurpleAir has EPA correction, state sensors have been calibrated). The motivation is stated at the outset of the paper (L8–14): calibration and correction equations improve the accuracy of a given sensor's measurement, but they are device-specific, and the maintenance, siting, and operation of many deployed devices is unknown, with the need for tractable, generalizable methods of identifying unreliable sensors across device types. Since comparisons occur in AQI bin space, which absorbs small noise (Section 3.3.1, L251–258), and the entropy and information statistics are functions are built to handle some measure of disagreement, we treat them intercomparable. Signal-to-noise differences between device types exist, and this is intended to complement, not replace calibration and correction methods.

Changes in manuscript: Table S1 added (fleet summary); the design rationale is carried by passages present since submission (L8–14; Section 3.3.1) and the device-response discussion in Section 6.3.

RC2-7

Comment: 2.2: “We define a smoke event as…” → This is not specific enough. Do you define the smoke event based on HRRR-Smoke? Based on NOAA HMS? Based on elevated CO from satellites? … need more detail on how you chose them and delineated the times. See: Liu, T., … (2024).

Author response: The events were identified by manual judgement by looking at all available sensor data from the sources used in this study (PurpleAir, AirNow, etc.), not a programmatic definition, and the revision now says so explicitly (L138–139). This data was contextualized using satellite fire detections from HMS, smoke plumes from HMS, known fire incidents from NIFC. Defining smoke events based off models (HRRR) could be very inaccurate, and NOAA HMS has its own limitations. The events were drawn from either known events (LA wildfires) or by reanalysing air quality impacts over the study period in different regions. The time periods were selected solely to see PM2.5 sensor networks undergo elevated conditions from a nearby or regional smoke source and then return to lower ambient concentrations. We now state why we deliberately did not derive events from a satellite smoke product. Such products can indicate smoke aloft rather than surface impacts and translating them into surface smoke days requires regionally calibrated filtering, as the reviewer’s suggested reference itself demonstrates (Liu et al., 2024, now cited at L144–146). Systematic smoke-event identification is a research problem in its own right and outside our scope; for our purpose, smoke-impacted periods in which to detect outliers, we believe hand-curated events are sufficient, and per-event characterization (satellite fire detections, daily smoke plume coverage, and network AQI time series) is provided in the companion materials (Illson, 2026b) (L146–148).

Changes in manuscript: Section 2.3 expanded with the selection and delineation procedure (L138–148); Liu et al. (2024) cited; Table 1 identifies the smoke source (WF/Rx) per event.

RC2-8

Comment: Fig 1, Table1: how is the extent determined in fig 1, what are the conditions of the Rx fire treatments? I downloaded the companion repository (Illson, 2026b) and it is fairly unstructured with little documentation… I think there needs to be more information on these fires and data you are using…

Author response: Extent was determined by analysis not to necessarily capture the full range of the smoke event, but to grab a sample of data with which to insert these synthetic outliers to calibrate our method. So, the analyst derived a bounding region to subset a group of air quality devices, shown in the paper in Figure 1 and in the companion repository. The conditions of the Rx treatments would require knowing the burn plan and day-of-burn meteorology, which are likely not attainable in specifics. The fire detections shown in the plot, timeseries of mean AQI for the study region, and HMS plume evolution over time should hopefully contextualize the events we chose. Note, we did not define smoke events as being related to a particular fire or set of fires, but rather a geographic area that had observational sensor networks experience elevated PM2.5, with contextual evidence supporting that the source would be smoke in origin, either from wildfire activity, or Rx fires, which were surmised by being out of climatological fire season, and having small and short-lived satellite fire detections nearby. We hope the textual

revisions will allay the reviewer's concerns, and if further action is required the companion repository, or paper can be adjusted with more information as needed.

Changes in manuscript: Extent basis now stated in the text: "These site extents are analyst-defined bounding regions used to subset devices for analysis and do not necessarily delineate the full extent of smoke" (Section 2.3, following the Fig. 1 reference, L~217–219); Rx attribution criterion added (out-of-season timing with small, short-lived satellite fire detections; L~209–211); the site-selection sentence reworded so the claimed extent is that of impacted devices evident in the observational data, not of the smoke itself (L134).

RC2-9

Comment: Fig 4: this figure is quite nice

Author response: Thank you. We have retained it (now Figure 3).

Changes in manuscript: None (renumbered only).

RC2-10

Comment: 3.1: "while providing measures of disorder that remain meaningful during smoke events." → I don't know what meaningful means

Author response: "Meaningful" was doing undefined work in that sentence. It has been reworded to state what we intend. The measures retain a consistent interpretation as networks vary in size and device type, including during smoke events (Section 3.3.1, L248–250).

Changes in manuscript: The quoted sentence reworded in Section 3.3.1 (L248–250); the later robustness statement carries its justification (L259–262).

RC2-11

Comment: "This categorization avoids overemphasis on minor numerical differences, while capturing meaningful shifts in health-relevant air quality." → I don't really buy this. If you want to say that it makes the mathematical optimization easier then that is perfectly fine to say. But public health researchers would argue that any concentration of PM2.5 is 'meaningful'…

Author response: Agreed, and we take the reviewer's point. Any concentration of PM2.5 is meaningful from a health perspective, and we did not intend to claim otherwise. However, the recommended actions for the public experiencing smoke vary by AQI category as opposed to each additional µg/m3 of PM2.5. The original wording was not explicit that "health-relevant" referred to the categories of public air quality indexes, not to the health effects of particulate matter, which are outside the scope of this work. The sentence has been rewritten to state what we actually rely on (Section 3.3.1, L251–258). Categorical bins suppress minor numerical differences that could be background noise between sensors, and they simplify the comparison our detection rule optimizes. We now claim only that this binning is sufficient for broad outlier detection, with the practical benefit that flagged disagreements correspond to the larger category differences.

Changes in manuscript: The quoted sentence rewritten (Section 3.3.1, L255–258): the health claim removed; the passage now states the optimization and noise-suppression rationale and scopes the binning claim to broad outlier detection and public-facing categories.

RC2-12

Comment: L323–332: no citations are given and the discussion here is fairly vague, referring to information theory as an approach or an idea, rather than just identifying specific applications or models or tests, etc.

Author response: That paragraph has been rewritten to anchor claims in cited applications rather than describing information theory in the abstract. Entropy-based design and evaluation of environmental monitoring networks (Keum et al., 2017), mutual-information sensor placement (Krause et al., 2008), value-of-information placement design in PM2.5 networks (Choi and Hummel, 2026), and entropy-based outlier detection in wireless sensor networks (Zhang et al., 2010). The paragraph then states how our use differs, characterizing the instantaneous state of a network and per-device agreement.

Changes in manuscript: Final paragraph of Section 3.3.1 rewritten with four citations (L259–268); vague and uncited claims removed.

RC2-13

Comment: "neighbouring devices whose measurements fall into AQI bin x" → are these the categorical AQI bins? Are the bins not equal in size? For example, you can have a monitor that is off by 50 AQI points but contained in the same category… but a monitor that is off by 5 that switches categories is flagged? Does that mean the latter example has greater entropy…? / Ok, I think the following paragraphs addressed it a little bit, but the idea of the bins is still unclear… only using the binned AQI categories or are you also using the AQI scores in the bins as well?

Author response: Yes, all comparisons use the categorical bins only. Entropy and information are computed on bin occupancy, and the bin deviation is measured in bin space; raw AQI values within a bin are not used. On the reviewer's example, an important clarification, a monitor off by an AQI value of 5 that crosses one category boundary is not flagged by the detection rule. The bin deviation must reach B/β, or roughly two to four bins at the used configurations. A bin change will increase entropy only so much as the number of bins present in the network at the time of observation. If that bin change is very unique in an ordered network, the method's bin-deviation parameter is meant to catch those to determine if the bin change is far enough away categorically to be of interest, which is tuneable for the operator. It can have a high information value, yet still not constitute an outlier under the full rule set. The problem the reviewer identifies is nonetheless real at bin boundaries, and we accept it as the cost of the categorical approach. The revision states this explicitly and describes the modified 12-bin AQI, whose overlap bins were added precisely to soften the category breakpoints (Section 3.4, L343–347; Fig. S2 shows the three binning schemes considered).

Changes in manuscript: Bin mechanics and the boundary trade-off stated in Section 3.4 (L343–347); modified-AQI definition with Fig. S2; the binning claim scoped per RC2-11 (L251–258).

RC2-14

Comment: Figure 7: font sizes are too small. Text is small, hard to see the lines.

Author response: That figure (Figure 5 under the revised numbering) has been regenerated with larger type.

Changes in manuscript: Figures regenerated with larger type.

RC2-15

Comment: “In terms of spatial configuration, the optimal proximity varied widely for each fault type, ranging from 5,000 to 20,000 m, implying that specific fault detectability may be at least partially scale-dependent.” → I think this is important to actually tease out… Why do these sets of parameters lead to the best performance for these specific error types?

Author response: The revision adds more interpretation. The per-fault pattern follows from how the faults were constructed, that the magnitude of the induced perturbations mirrors the failure modes they were designed to mimic, and from first principles, more extreme perturbations should be easier to identify (Section 4.1, L509–516). The five fault types differ significantly in how far they move a device from its network. Fault type 1, which was a very high shift fault causes readings well above the detection threshold, while outage, flatline, drift, and sinusoidal faults move readings a median of one to two bins, at or below the threshold by construction. So the low per-fault scores on the subtle faults reflect the subtlety of those failure modes rather than parameter idiosyncrasy. On the parameter side, the optima cluster interpretably (L485–497): entropy limits sit high in the tested range across all fault types (an ordered network is a precondition), information limits sit low, and the wide spread in optimal proximity reflects that different failure modes manifest at different spatial scales relative to their neighbours.

Changes in manuscript: Per-fault interpretation added in Section 4.1 (L509–516); parameter-clustering discussion at L485–497; operational implications in Section 6.1.

RC2-16

Comment: Don’t find Figure 9 very useful

Author response: We have removed it in the revision.

Changes in manuscript: Original Figure 9 deleted; subsequent figures renumbered.

RC2-17

Comment: I find the XGBoost sections fairly haphazard. There are no tables or figures and it seems to come out of nowhere.

Author response: This was addressed as a revision. All XGBoost methods are now consolidated in Section 3.5 (motivation, features, labelling, and training; L356–418), introduced immediately after the rule-based equation whose limitations motivate it (L357–361).

Changes in manuscript: XGBoost methods consolidated into Section 3.5 (L356–418); results consolidated in Section 4.2 with new Table 3; method comparison in Section 4.3; scope statement at L556–558.

RC2-18

Comment: The high misclassification rate during local smoke events is concerning especially given that that is a benefit of using low cost sensors. There is no discussion of what the correlation length scale of wildfire or Rx fire smoke might look like, but they are generally long traveled plumes (at least for wildfires). Making the local classification of smoke an easier task, at least from first principles.

Author response: The reviewer's first-principles reasoning is correct for transported wildfire smoke, and the revision now states the correlation-length structure directly (Section 6.2, L830–841): large plumes travel regional to continental distances (Zhang et al., 2025), impact entire networks at once, and produce strong consensus between sensors that there is a smoke event. Classification there is comparatively easy, which our smoke-stratified results reflect. The elevated misclassification rate is specific to local smoke, and follows from how those events are defined in this manuscript (L635–638). Part of a network impacted while most is not (prescribed and small fires produce less smoke that disperses over shorter distances and interacts strongly with terrain; Miller et al., 2019). These are, by our definition, the most disordered network states and the hardest case to separate true outliers from single sensors reading real smoke. Fig. S5 (supplement) shows an example. The revision also addresses that (Section 5.3, L723–728) local smoke and transition periods are disproportionately short-lived, so increasing consecutive hours flagged helps filter out these high false positives. These are also tunable through entropy parameters or a minimum neighbour count in the network.

Changes in manuscript: New Section 6.2 paragraph (L830–841) with Zhang et al. (2025) and Miller et al. (2019); persistence-deferral framing in Section 5.3 (L723–728); case study retained in the supplement (Fig. S5).

RC3

Review preamble: "the manuscript already appears publishable with minor revisions" — acknowledge the repost note ("I initially misunderstood the review process…") graciously.

RC3-1

Comment: A fundamental question is the lack of a clear comparison baseline. Additional justification may be beneficial regarding why the proposed framework is particularly necessary under smoke conditions, unless there are cases where conventional approaches perform well under normal air-quality conditions but fail during smoke events.

Author response: This was a flaw of our study, so the revision adds in a comparison, while attempting to not further bloat the paper. New Section 3.6 (L420–442) defines three baseline neighbour-comparison rules. The absolute neighbour deviation, the neighbour z-score, and the median/MAD robust z-score. Each was computed on the same networks and hourly measurements as our methods, evaluated at their conventional cutoffs and additionally at the cutoffs that maximized their performance on our benchmark, against a common, method-

independent outlier definition. Section 4.1 reports the outcome under both objectives (L516–540): 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. To the reviewer's specific question, as to whether there are cases where conventional approaches perform well under normal air-quality conditions but fail during smoke, that case which prompted this work is now shown directly. Absolute neighbour deviation is capable in ambient air (44 false positives) but fails when smoke arrives (1,102), while the tuned equation's false-positive rate moves only modestly across the boundary (0.57% to 1.30%) (Section 4.1.3, Table S4).

Changes in manuscript: New Section 3.6 (L420–442); baseline-comparison paragraphs in Section 4.1 (L516–540); smoke-stratified false-positive comparison in Section 4.1.3; supplement Tables S3 and S4.

RC3-2

Comment: A stochastic spatiotemporal modeling framework may also be considered as a comparison approach… Including a brief discussion of physically informed spatiotemporal transport formulations in the Discussion section may help further contextualize the limitations and assumptions of the current network-based approach. For example: Choi, B., and Hummel, M. A. (2025)…

Author response: We appreciate the suggestion, and had we been aware of this work, it may have led to a different methodological approach. Given time and energy constraints, this had to be confined to an acknowledgement in the discussion. Section 6.4 now contains a physically-informed spatiotemporal paragraph (L911–915). Our networks encode proximity alone, whereas formulations such as the stochastic advection–diffusion model of Choi and Hummel (2025) represent pollutant transport explicitly, and could supply physically grounded expectations of inter-device agreement as a plume moves through a network. This helps the discussion contextualizes the assumptions and limitations of distance-based network generation as the reviewer suggested.

Changes in manuscript: Physically-informed future-work paragraph added to Section 6.4 (L911–915) with Choi and Hummel (2025) cited.

RC3-3

Comment: (Suggestion) Reducing redundant explanations in several sections may improve the flow of the manuscript.

Author response: Agreed. The reorganization consolidated all methods into Section 3 and removed redundant explanations across sections, the original Figure 9 was deleted, and reference material moved to the new supplement, reducing the main-text figure count from 17 to 12. We note the tension between the requests we received: shorten the manuscript while also providing additional information, comparisons, and clarification. This resolved by reducing figures, moving reference material to the supplement, and condensing the text where we were able.

Changes in manuscript: Reorganization and redundancy reduction as described in the response to Reviewer 2's first comment; figure count 17 → 12; new supplement.

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 (L573–574), 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).