Author response to Referee 2 (RC2)

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

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 L68 and L180 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, L623) 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 (L9–14): 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, L249–256), 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 (L9–14; 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 (L136–137). 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 L142–144). 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) (L144–146).

Changes in manuscript: Section 2.3 expanded with the selection and delineation procedure (L136–146); 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 (L132).

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, L246–248).

Changes in manuscript: The quoted sentence reworded in Section 3.3.1 (L246–248); the later robustness statement carries its justification (L257–260).

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, L249–256). 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, L253–256): 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 (L257–266); 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, L341–345; Fig. S2 shows the three binning schemes considered).

Changes in manuscript: Bin mechanics and the boundary trade-off stated in Section 3.4 (L341–345); modified-AQI definition with Fig. S2; the binning claim scoped per RC2-11 (L249–256).

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, L507–514). 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 (L483–496): 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 (L507–514); parameter-clustering discussion at L483–496; 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; L354–416), introduced immediately after the rule-based equation whose limitations motivate it (L355–359).

Changes in manuscript: XGBoost methods consolidated into Section 3.5 (L354–416); results consolidated in Section 4.2 with new Table 3; method comparison in Section 4.3; scope statement at L553–555.

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, L827–838): 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 (L632–635). 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, L720–725) 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 (L827–838) with Zhang et al. (2025) and Miller et al. (2019); persistence-deferral framing in Section 5.3 (L720–725); case study retained in the supplement (Fig. S5).

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).