We thank both reviewers for their recognition of the breadth of this work and the value of its observational contributions. Both reviewers raised concerns regarding our GEOS-Chem analysis and the emission ratio/enhancement ratio (ER/EnR) framework. We note that many of these concerns can be addressed through further clarification. For example, both reviewers pointed out that the overestimation of ozone in wildfire smoke by chemical transport models has been documented in previous studies. We want to highlight that the contribution of this work is not to reiterate the existence of an absolute ozone bias in models. Rather, our analysis investigates a deeper aspect of this discrepancy: the model's inability to reproduce the relationship between PM2.5 and MDA8 ozone observed during smoke events, where ozone increases under light smoke conditions but levels off or even declines under heavier smoke. This incapability at least partly explains the well-documented ozone overestimation in models mentioned by the reviewers. Admittedly, more work and more sites need to be tested in the model to address this weakness. We have revised the manuscript to more clearly articulate our contributions. Our point-to-point responses to individual comments are shown below in black, and reviewer comments are written in green.

Reviewer 1:

In this manuscript, Jin et al describe observations and modeling for a one-month smoke influenced period in Missoula MT. The manuscript covers a lot of ground. Some of this is done well and is a useful contribution, but other aspects are not strong. The areas that I found to be strong (or could be) are:

1. The range of VOC observations.
2. The smoke day identification scheme.
3. Cancer risk calculations (although this is not adequately described).
4. Role of rainout in removing PM_5 (barely described, but this could be a very interesting section). This last point was mentioned only briefly (line 450), but is potentially an interesting analysis. There have been some prior debates over the possibility of "low PM smoke events", and these results seem to support this. To my knowledge, there is almost nothing published on this, so a good analysis on this point would be a useful contribution.

Response: We appreciate the reviewer's identification of the manuscript's contributions. In response, we have revised the manuscript to improve the clarity and transparency of the health-risk methodology and to better contextualize the discussion of rainout and low-PM smoke. We agree that this aspect is potentially important and underexplored in the literature. At this moment, because only one day of rainy smoke-influenced days was captured during the study period, we have been careful not to overinterpret these results. We've also addressed the reviewer's concerns regarding O3, the ER/EnR analysis, and the GEOS-Chem results point-by-point below.

The aspects that I found to be weak or not useful contributions were:

1. The O_3 observations (possible bias?) and analysis.

Starting with the O_3 observations, the authors should know that the Thermo-Fischer 49i instrument used here is known to suffer from strong positive bias in smoke.

See these refs: https://doi.org/10.5194/amt-14-1783-2021 and https://doi.org/10.5194/amt-15-3189-2022

It is possible that the operators have fixed this problem, but at minimum, the authors should describe the bias issue and what, if any, steps were taken to minimize this bias. I recognize the results shown (low O_3) seem inconsistent with a positive bias, but nonetheless, the issue should get discussed.

Response:

Thank you for raising this important point. We have now revised Sect. 2.3 to acknowledge this potential artifact and cite the relevant studies.

Although the University of Montana's 2B Technologies Model 211 ozone monitor, which is generally less susceptible to these interferences, was not deployed during the September 2020 study period, both instruments were operated concurrently during July 2021, which overlaps with the peak wildfire season. We therefore compared daytime ozone measurements from the Montana DEQ Thermo Scientific 49i and the University of Montana 2B Model 211 during July 2021 and separated the overlap period into no/least BB-impacted and BB-impacted conditions (Fig. X1). We note that the two ozone measurement datasets are not collocated, and the difference between them reflects not only different instruments but also different site locations: the DEQ and UM sites are approximately 3 km apart, with the DEQ site more influenced by urban NOx emissions and the UM site more influenced by campus-scale activity. However, it could help shed light on how the potentially positive bias may affect our analysis.

The comparison shows that the relationship between the two instruments is broadly similar in the two subsets. In the scatterplot comparison (Fig. X1A), the regression slopes are close to unity, and the correlations are high in both subsets (no/least BB-impacted: slope = 0.98, $r = 0.85$; BB-impacted: slope = 1.1, $r = 0.90$). The difference in slope between the two subsets is not statistically significant at the 0.05 level ($p = 0.061$). In the bias analysis (Thermo − 2B; Fig. X1B), the mean daytime bias is +1.24 ppb for no/least BB-impacted conditions and +2.64 ppb for BB-impacted conditions. The between-period difference in mean bias is also not statistically significant ($p = 0.109$). These results suggest that, within the July 2021 daytime overlap period, we do not detect a statistically significant difference between the Thermo 49i and 2B Model 211 measurements.

However, we do observe considerable hour-to-hour variability between the two instruments (Figs. X1B and X1C), with individual differences frequently reaching about +10 to +15 ppb and

occasionally falling to about −10 to −20 ppb. Thus, while the average bias is modest, short-term agreement between the two records is imperfect. Some portion of this variability likely reflects true spatial variability in ozone between the two sites, in addition to instrumental differences.

Finally, any residual interference in the Thermo 49i would be expected to bias O3 high during smoke events, not low. It therefore cannot explain the suppressed O3 observed under the heaviest smoke conditions and would, if anything, make our inferred O3 suppression conservative. In summary, we have revised the Methods text describing the O3 measurements to explicitly acknowledge this potential artifact and cite the relevant studies in Sect. 2.3 (Bernays et al., 2022; Long et al., 2021).

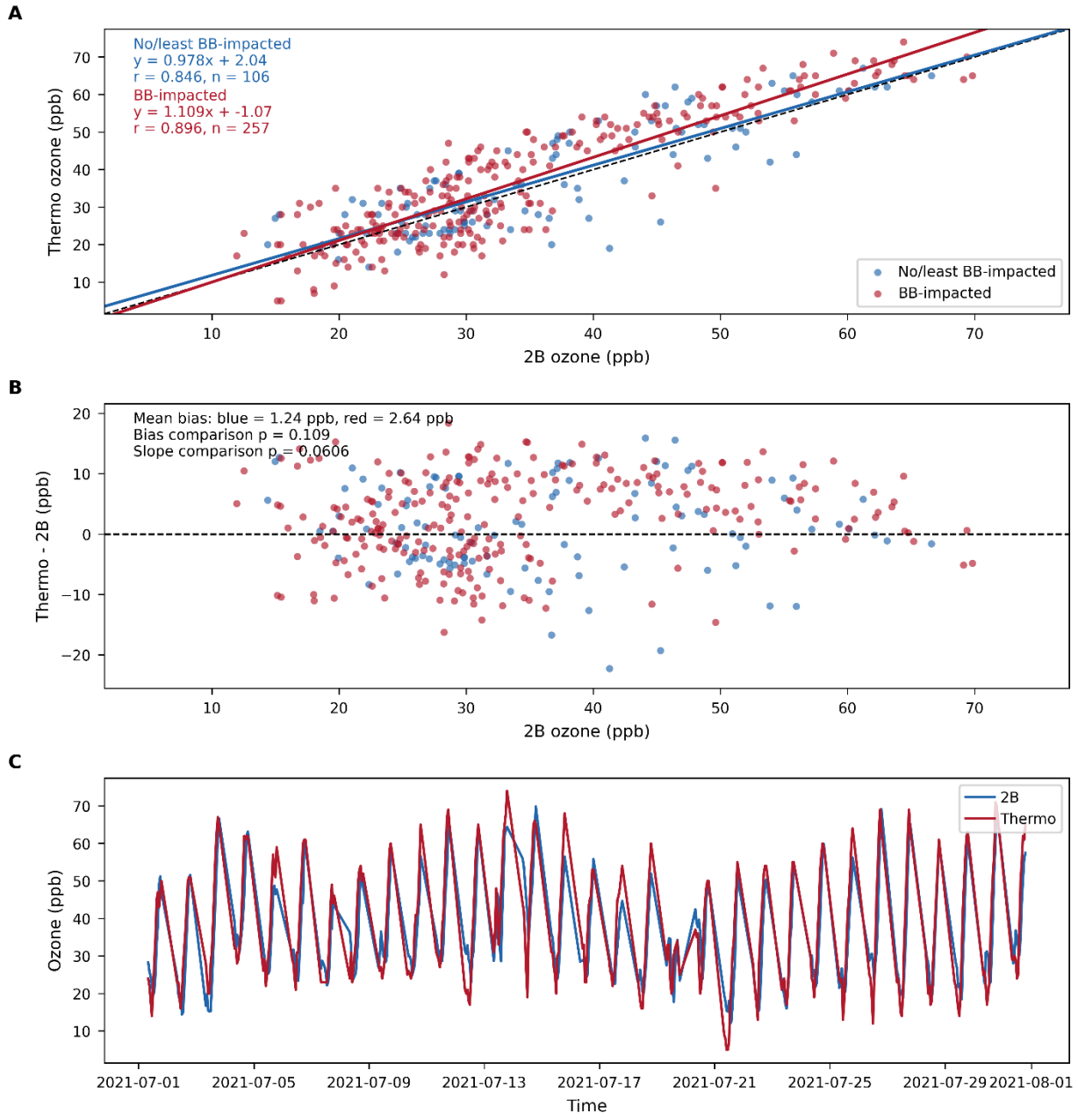

Figure X1. Daytime comparison of ozone measured by the Montana DEQ Thermo Scientific 49i and the University of Montana 2B Model 211 during July 2021 at two sites in Missoula. (A) Thermo versus 2B ozone under no/least BB-impacted (blue; PM2.5 < 12.5 µg m^{-3}) and BB-impacted (red; PM2.5 ≥ 12.5 µg m^{-3}) conditions. Colored lines show linear fits; the dashed black line shows the 1:1 relationship. (B) Thermo − 2B as a function of 2B ozone. (C) Daytime time series from the overlap period. Note the distance between the two sites is about 3 km.

2. Regarding the O3 analysis, the authors have too few observations to draw much of a conclusion here. O3 from smoke is tricky as it can be positive or negatively influenced by smoke. One must also consider what O3 would have been on the day without the smoke.

Response: We respectfully disagree that the available sampling precludes interpretation of the qualitative O3–PM2.5 behavior during the smoke episode. Wildfire smoke can both enhance O3 under lighter smoke and suppress O3 under heavier smoke, and this behavior has been widely documented (Buysse et al., 2019; McClure and Jaffe, 2018). In our dataset, all three sites exhibit consistent behavior despite varying peak PM2.5 levels, likely reflecting differences in proximity to sources and smoke intensity. We agree that the exact turning point (i.e., the PM2.5 level at which O3 transitions from enhancement to suppression) can vary across sites and events and is best constrained using multi-year, multi-site observations (Buysse et al., 2019). We have softened and edited our language in Sect. 6 where feasible to avoid over-interpreting features supported by only a small number of days.

To address the reviewer's counterfactual point ("what would O3 have been without smoke"), we leverage the one-third of September 2020 no-smoke days at Missoula. We show that the structured enhancement–suppression pattern is apparent during smoke-impacted conditions, whereas on no-smoke days O3 variability does not exhibit the same relationship with PM2.5 (**Fig. X2**).

Finally, we emphasize that the goal of this section is not simply to add another case study site, but to evaluate to what extent a CTM can reproduce the observed smoke-intensity dependence of O3 relative to PM2.5 at a surface site. We plan to expand this model-evaluation framework to broader regions and longer time periods in future work.

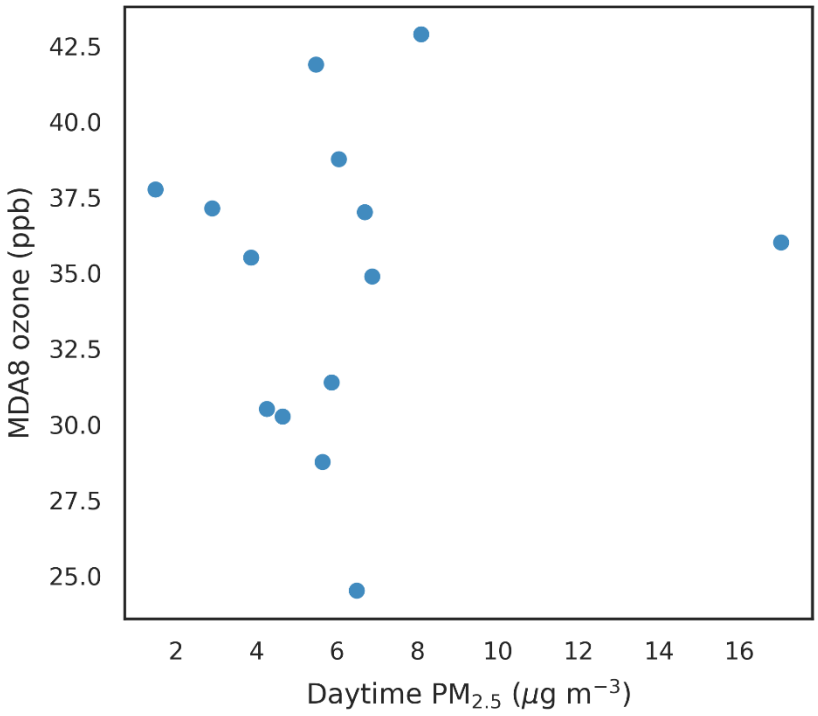

Fig. X2. Relationship between daily maximum 8-hour average ozone and daytime PM2.5 during non-BB days in September 2021 in Missoula.

3. The GEOS-Chem analysis. Regarding the CTM (GC) analysis, I am not sure what this adds to the analysis. Using a coarse resolution model (25km) to model smoke transport and chemistry is well known to problematic. First one has to get the emissions and transport correct and then the chemistry. Neither is handled well in GC for smoke plumes.

Response: we appreciate the reviewer's concerns regarding the use of a ~25 km chemical transport model (CTM) for wildfire smoke. We agree that coarse-resolution CTMs have well-known limitations for near-source plumes and the very rapid post-emission chemistry in fresh smoke. Accordingly, we do not use GEOS-Chem as a plume-resolving tool, nor do we draw conclusions that depend on near-source processes.

However, we respectfully disagree that this implies the GEOS-Chem analysis is not useful in the present context. Our study is explicitly receptor-oriented and focuses on aged, regionally mixed smoke impacting Missoula over multi-day periods, for which CTMs are commonly applied to provide a regional-scale context and a diagnostic model–observation evaluation framework. In fact, GEOS-Chem reproduces several key aspects of this episode reasonably well at the receptor site: it captures the timing/occurrence of smoke-impacted periods and the broad temporal evolution of smoke-related variability, providing meaningful regional context for transport of aged smoke.

Here, we extend our earlier CO-constrained ground evaluation (Jin et al., 2023) by leveraging co-located surface observations of PM2.5, O3, and VOCs, enabling a more comprehensive multi-pollutant assessment of CTM performance during smoke. Such ground-based evaluation of speciated VOCs during wildfire impacts remains relatively limited in the CTM literature because long-term VOC observations are scarce. Within this framework, GEOS-Chem exhibits a systematic low bias in the magnitude of smoke enhancements across multiple species; this multi-species bias pattern is consistent with underestimated BB emissions (magnitude and/or speciation) and/or missing or underrepresented processes affecting smoke mass and composition (e.g., wet scavenging and/or reactive VOC chemistry). In addition, because a long-standing CTM challenge is reproducing observed O3 behavior in smoke, we use this observationally rich episode as a bounded diagnostic benchmark to test whether GEOS-Chem reproduces the observed smoke-intensity dependence of O3 relative to PM2.5, and to identify where the model succeeds or fails. We believe, that these analyses and associated findings, in particular, the additionally identified model weaknesses in BB environments are important contributions to the literature.

The rapid OH chemistry in smoke plumes after emissions is another well known problem for models. The fact that CTMs often dramatically over-predict O_3 from fires has been shown previously:

http://dx.doi.org/10.1016/j.atmosenv.2016.06.032

https://doi.org/10.1016/j.scitotenv.2018.05.048

It seems fairly obvious that this is the wrong tool to apply to the problem.

Response: We acknowledge the reviewer's statement about "rapid OH chemistry in smoke plumes after emissions." While the reviewer does not provide additional detail, we interpret this comment as referring to a known limitation of coarse-resolution CTMs in representing early post-emission chemical processing. As discussed in the last comment, our focus here is on diluted, aged smoke, which is less affected by the sub-grid issues.

We also appreciate the reviewer pointing to prior studies documenting substantial positive model biases for fire-related ozone (e.g., Baker et al., 2016; Baker et al., 2018). Again, our objective here is not to simply restate that "CTMs overpredict fire O3". The value of this work is partly in answering why the absolute smoke O3 concentration is wrong: testing whether CTMs can reproduce the non-monotonic relationship reported in observations.

In the revised manuscript, we have 1) cited the studies noted by the reviewer as evidence of previously documented CTM O3 bias in smoke and 2) revise the text to emphasize that our contribution is not a general statement that "CTMs have O3 bias," but the identification of a

more specific model deficiency: failure to reproduce the observed non-monotonic O3 response under smoke influence. These revisions are reflected in Lines 500–520.

Lastly, I have some concerns with the ER/EnR analysis. First, its not clear what the authors have done nor what is the goal of this analysis. Is it to show emission ratios (not emissions) or estimate photochemical age or what ? In any case the uncertainties are very large and not discussed.

Response: Thanks for raising this concern. We rewrote this section in the revised manuscript to clarify the objectives and outputs of this analysis. Specifically, (i) the ln(EnR)–photochemical-age framework is used to extract chemical-processing information from fixed-site observations of aged smoke by estimating an effective first-order loss rate (reported as $k_{\mathrm{OH,eff}}$ for selected primary VOCs) and an approximate time-zero enhancement ratio (reported as ER_0). (ii) Separately, we report event-integrated EnRs of PM2.5 and VOCs relative to CO as a descriptive metric for comparison with prior studies; we do not interpret these as emission factors.

We have also expanded the uncertainty discussion and quantified the inherent method error using two consistency/closure checks. First (observation closure), we test internal consistency using the same species that define the photochemical clock (benzene and toluene): applying the ER-retrieval framework to the observations returns time-zero ERs for benzene and toluene within ~20% of the assumed values used in the aging calculation. This provides an empirical estimate of methodological uncertainty arising from background selection, mixing, and clock assumptions. Second (model synthetic-truth test), we apply the identical ER/EnR retrieval workflow to GEOS-Chem time series at the receptor site and test whether the method recovers the benzene/toluene ER implied by the model. In this synthetic test, the retrieved ER_0 agrees with the model-implied (GFAS-based) value within ~10%, providing an additional bound on intrinsic method performance when the "true" ER is known.

In addition, we expand the sensitivity analysis in the Supplement. EnR uncertainty is mainly driven by the background definition, whereas photochemical age uncertainty is dominated mainly by the assumed initial benzene/toluene ratio used in the clock framework. Sensitivity tests show that changing the background with a looser or stricter definition alters integrated EnRs by ~20%, while changing the assumed initial benzene/toluene ratio from field-based values to a lab-based value shifts inferred time-zero ERs for benzene, toluene, and C8 aromatics by ~10–40% (Fig. X3) (Gkatzelis et al., 2024; Koss et al., 2018; Permar et al., 2021). These quantified uncertainties are now reported explicitly and do not change the conclusions of the analysis.

We revised the manuscript accordingly to clarify the objectives of the ER/EnR analysis in the main text and to expand the discussion of associated uncertainties and closure tests in Supplement Sect. S1.

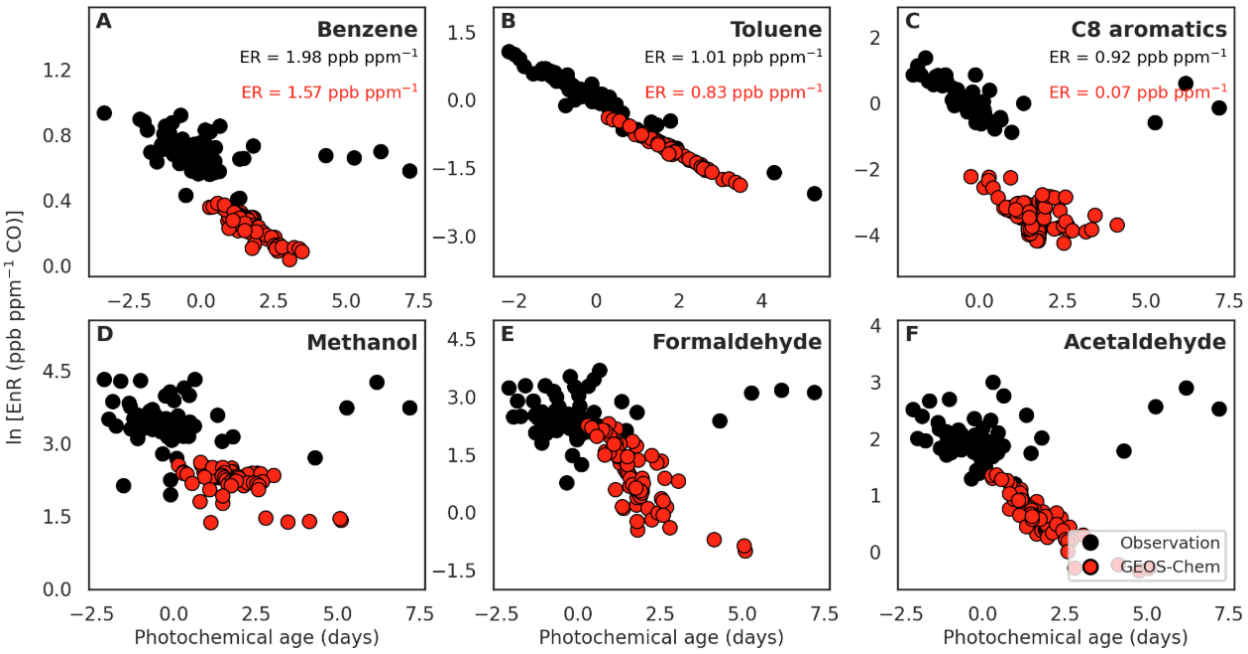

Figure X3. Observed and modeled enhancement ratios (EnR) of key volatile organic compounds (VOCs) versus photochemical age. Panels show ln(EnR) versus photochemical age for (A) benzene, (B) toluene, (C) C8 aromatics, (D) methanol, (E) formaldehyde, (F) acetaldehyde, where EnR is defined as ΔVOC/ΔCO (ppb ppm^{-1} CO). Observations are shown as black circles and results from GEOS-Chem driven by GFAS fire emissions are shown as red circles. Points represent 3 h binned data. Similar as Figure 5, but the initial toluene-to-benzene ratios have been changed from field-based to lab-based values.

Detailed comments:

Line 46-48"However, analogous…" ?? Aren't there lots of observations that could be used to evaluate CTMs?

Response: Indeed, observations of PM2.5 mass and O3 are many. However, co-located speciated VOC and HAP measurements during smoke periods remain limited, particularly during sustained smoke events. Such data can be used to constrain smoke composition and reactivity in the CTM. We revised the sentence to: "However, analogous ground-based validation for smoke-related VOCs/HAPs—especially during sustained smoke episodes—remains limited, motivating this work."

50: I think the wildfire smoke in Missoula was actually worse in 2021.

Response: We agree that Missoula experienced severe smoke impacts in 2021, and we revised the text to avoid implying that 2020 was uniquely the "worst" smoke year for this site. Our focus on September 2020 is instead motivated by its relevance as a major western U.S. smoke event with continental-scale influence, together with the availability of our most comprehensive near-surface chemical measurements in Missoula, including speciated VOC and hazardous air

pollutant observations required for the analyses in this paper. Although 2021 may have had a longer smoke season in Missoula, September 2020 captured a particularly intense and well-characterized period of sustained regional smoke exposure, making it the most suitable case study for this work.

L101: "Based on climatology.." ??

Response: Here, by "climatology" we mean the air-quality climatology of PM2.5, i.e., the multi-year statistical distribution of daily-mean PM2.5 in Missoula during the wildfire season (June–October) over 2010–2024. We revised the text at first mention to explicitly define "climatology" in this sense and updated the Fig. 1 caption to briefly state how the distribution is constructed (median/IQR and whisker definition; handling of missing data) in the corresponding text. It now reads: "Figure 1 illustrates the climatology of PM2.5, defined here as the multi-year statistical distribution of daily-mean PM2.5 during the wildfire season (June–October) over 2010–2024 in Missoula."

130; Define ncps

Response: We now have defined ncps at first use as normalized counts per second. The revised text reads: "Sensitivity of all compounds used in this study ranges from 0.9 to 14 ncps $ppbv^{-1}$, where ncps denotes normalized counts per second."

146: The Thermo instrument seems to have the strongest bias. Please discuss data and any remediation that was taken.

The details are provided in response to Comment 1.

150: As noted above, I am not sure what is gained from the GC results here.

Response: We respectfully disagree that the GEOS-Chem results do not add value. The model evaluation is important not only for ozone, but for the broader aged-smoke mixture, including CO, PM2.5, speciated VOCs, and O3. GEOS-Chem reproduces the timing and transport of the major smoke intrusions reasonably well, but it underestimates the magnitude of smoke enhancements, with low biases of about 30–40% for CO and PM2.5 and about 60–90% for most VOCs. It also fails to reproduce the observed non-monotonic O3–PM2.5 relationship, indicating an important limitation in its representation of smoke-influenced ozone chemistry. These results show that current CTMs may capture regional smoke occurrence while still misrepresenting key reactive chemistry and smoke composition, with direct consequences for exposure assessment. In our case, these biases lead GEOS-Chem to underestimate the smoke-attributable cancer risk by about 40% and the chronic non-cancer hazard index by about an order of magnitude.

196: I don't understand the 1.5x. Does this mean for PM with (say) a bg of 6 ug/m3, the criteria is 15 (9+6)? Similarly for CO, with a bg of 100 ppb, would the criteria be (approx.) 250 ppb (100+150)? Please clarify. This seems ok for PM, with a relatively low bg, but not CO, as it has a high bg.

Response: We thank the reviewer for pointing this out. We agree that the "1.5× background" wording can be misleading when applied to individual species with very different background levels and variability. To avoid over-interpreting this threshold, we want to clarify that smoke identification is based on multiple lines of evidence considered together, rather than a strict stand-alone cutoff for any single species such as CO. Our classification instead relies on the combined use of PM2.5, HMS/satellite smoke information, and other supporting tracers to identify smoke influence more robustly.

215: It seems you are looking for a yes/no answer, but in reality you are probably getting a range of smoke influence.

Response: We agree with the reviewer that smoke influence is better represented as a continuum rather than as a strict yes/no condition. In the original manuscript, we therefore interpreted smoke influence using multiple indicators and graded confidence levels (low, medium, and high), rather than a binary classification.

For example, we describe medium-confidence smoke influence as follows: "At intermediate daily mean $PM_{2.5}$ between 20 and 35 μg m^{-3} (n = 5; 5–6, 11–12, and 23 September), $\Delta PM_{2.5}$ still overlaps at least two gas-phase tracers but not all, signaling moderately aged or dispersed smoke and therefore medium diagnostic confidence."

We describe low-to-medium-confidence smoke influence as follows: "When the daily-mean $PM_{2.5}$ falls below 20 μg m^{-3} (n = 4; 7, 21–22, and 24), $PM_{2.5}$ overlaps with at most one gas tracer, and HMS does not indicate smoke on 24 September, so these low-load cases warrant only low-to-medium confidence."

We have now revised the text to more clearly emphasize the subset of days on which all indicators agree, and designated these as high-confidence BB-impacted days. The revised text reads: "When daily mean $PM_{2.5}$ exceeded 35 μg m^{-3} (n = 7; 13–19 September), the $\Delta PM_{2.5}$ approach aligned with HMS and with all gas-phase tracers considered here (i.e., CO, ACN, furan, and MA), indicating concentrated smoke that coupled large aerosol mass with gaseous enhancements. A k-means clustering analysis (k = 3: smoke, no-smoke, uncertain) applied to pairwise combinations of CO, $PM_{2.5}$, and each tracer corroborated this classification, reinforcing that at extreme aerosol loadings, routine ground-level $PM_{2.5}$ alone can reliably identify smoke during summer in Missoula. We therefore classify these days as a high-confidence BB period".

We hope this revision addresses the reviewer's concern by clarifying that our framework does not treat smoke influence as strictly binary, while still identifying the subset of days with the strongest and most consistent evidence for BB impact.

256: Strongly recommend to use local standard time or GMT.

Response: Our study period is entirely in September, when Missoula is on MDT (UTC−6). We therefore keep local time (MDT) throughout and have revised the manuscript to explicitly label MDT (UTC−6) at first mention (see Line 260)

270: missing: Does this mean horizontal or vertical?

Response: Thank you for noting the ambiguity. We meant both processes (i.e., horizontal transport (advection) and vertical mixing/entrainment). We have revised the sentence to clarify that the strong correlation indicates that smoke was relatively well mixed across the Missoula valley at ground level. Now it reads: "Strong correlation (r^2=0.9; Fig. S2) among $PM_{2.5}$ measured at the Boyd Park and other tracers measured at the UM campus indicates spatially coherent smoke influence across Missoula at the surface"

275: Rain removal is very interesting and should be elaborated on.

Response: We thank the reviewer for highlighting this point. Now it reads: "A preceding rain event on 20 September coincided with a rapid decrease in 24-h mean $PM_{2.5}$ from 40 to 10 μg m^{-3}, whereas most gas-phase species remain relatively less decreased, with CO declining from 500 ppb to 350 ppb and total VOCs from 50 to 30 ppb. This pattern suggests preferential removal of particles by precipitation scavenging and may provide a physical explanation for "smoke-present but low-PM2.5" days. It also highlights that smoke identification based on column indicators (e.g., HMS) may not always reflect surface PM exposure when precipitation occurs."

Figure 3: Is time axis LST, or ??

Response: Thank you for pointing this out. The time axis in Fig. 3 is in local time, specifically Mountain Daylight Time (MDT, UTC−6), because the study period is entirely in September 2020. We have revised the figure caption and relevant text to state this explicitly and avoid ambiguity.

368/Section 6: I found this analysis to be weak. There are too few points to make a significant conclusion on the PM-O3 relationship. The large over-prediction in modeled O3 has been shown. It is very hard to model smoke O3 and this has been shown by others.

Response: We agree that the number of observations in this case study is limited and does not support a strong quantitative conclusion. However, our intent was not to identify a precise PM2.5 threshold at which O3 shifts from enhancement to suppression, as this behavior has already been

discussed in previous studies of BB-impacted environments (Buysse et al., 2019; McClure and Jaffe, 2018).

Rather, our purpose is to document the O3–PM2.5 relationship observed at Missoula and other sites, and then to evaluate whether the CTM captures this regime transition. This is important because our results suggest that the model's inability to reproduce this behavior may help explain its tendency to overpredict O3 under smoke-affected conditions, beyond uncertainties in emissions and transport alone. Further evaluation across more sites and years would be valuable, but is beyond the scope of the present study, which focuses on September 2020. Accordingly, we have softened the text to emphasize that the purpose of the model comparison is to test whether the CTM captures the PM-O3 relationship, which, to the best of our understanding, has not been shown in the literature.

399/Section 7: This is an important section. It needs a better description of the methodology.

Response: We thank the reviewer for emphasizing the importance of Sect. 7. We agree that the methodology for the health-risk analysis needed to be described more clearly in the main text. In the revised manuscript, we added an explicit cross-reference at the start of Sect. 7 to direct readers to the full methodological description in the Supplement. The revised text now reads: "Figure 7 summarizes the upper-limit chronic inhalation risks from wildfire smoke, assuming that a 2020-style wildfire season were to recur annually over the next century. Methodological details are provided in Supplement Sect. S2."

408: Please restate this, not clear as written: "If the BB impacted…."

Thank you for noting this lack of clarity. We have revised the sentence to make the comparison explicit. Specifically, we refer to a hypothetical non-smoke baseline applied to the same set of BB-impacted days. The revised sentence now reads: "If the same BB-impacted days had instead experienced typical non-smoke background concentrations, the corresponding PM2.5-attributable excess cancer risk would be ~15 cases per million people."

416: HI = 3. Need methodology.

Response: Thank you for noting this. The calculation is also described in the Supplement (Sect. S2). To make this easier to locate from the main text without adding a lengthy methods description in the Results, we added a direct cross-reference at first mention of HI in Sect. 7.

Now it reads: "Non-cancer chronic risk is expressed as a hazard index (HI), with the dominant effects in this study associated with respiratory and cardiopulmonary systems. The full calculation framework is provided in Supplement Sect. S2."

Reviewer 2:

This manuscript documents the impacts of wildfires on air quality at a ground site in Missoula, Montana, in 2020. The study integrates both ground-based measurements and atmospheric modeling and clearly represents a substantial amount of work. While the manuscript has considerable breadth, it lacks sufficient depth. Both the ground observations and the disagreements between measurements and models are the same as already shown in the literature.

Response: We respectfully disagree that "both the ground observations and the disagreements between measurements and models are the same as already shown in the literature." The novelty here is not another smoke case study with routine pollutants, but rather the documentation of the smoke's impact in a city during a continental smoke event that lasted ~2 weeks in September 2020, and detailed model evaluation for such an event. The presented measurements are a chemically comprehensive, co-located, near-surface observational record at community receptor sites during a sustained smoke period, including CO, PM2.5, NOx, and O3, and >70 speciated VOCs. Near-surface datasets with this level of chemical speciation and time resolution, where exposure occurs, are uncommon.

This coverage enables constraints that routine monitoring cannot provide, especially surface evaluation of VOC performance in CTMs. On the modeling side, our objective is to use these multi-species observations to diagnose likely drivers of model error, and to test whether CTM-based systems reproduce key observed behaviors, including the O3–PM2.5 response across smoke conditions, which has not been explicitly evaluated in CTMs using co-located surface observations.

Additionally, because health risk is driven by composition rather than PM mass alone, our chemically resolved observations enable (i) species-resolved health-risk attribution (i.e., identifying which pollutants drive cancer risk and HI beyond PM2.5 mass) and (ii) an exposure-relevant model evaluation, testing whether CTMs reproduce the health-relevant mixture that underpins the derived risk metrics, which is not possible with routine PM-only monitoring.

1. Figure 3: Why does maleic anhydride exhibit such strong diurnal variation?

Response: Maleic anhydride (MA) is not primarily a directly emitted BB tracer like CO or acetonitrile. It is largely a secondary oxidation product of furanoids (Coggon et al., 2019). Therefore, MA is expected to show a stronger diurnal structure than primary tracers because its abundance reflects photochemical production tied to daytime oxidant availability.

2. Emission ratio estimation: The approach used to estimate the emission ratios is confusing and appears circular. For example, the measured emission ratio from WECAN campaign is used to calculate the photochemical clock, and the photochemical clock is then used to infer the emission ratio. In addition, given that the lifetime of toluene is only about 2 days, the

toluene/benzene ratio is not an appropriate tracer for inferring a photochemical age of ~7.5 days. In such an aged plume, primary toluene would be largely depleted, and background toluene concentrations would likely have a substantial influence on the toluene/benzene NEMR.

Response: We thank the reviewer for this important comment. We agree that the presentation of the emission-ratio (ER) analysis was not sufficiently clear and could give the impression of circular reasoning. We rewrote Section 5 to clarify both the purpose and the limitations of this approach.

First, our intent was not to claim that the toluene/benzene framework provides a fully independent constraint on ERs under all conditions. Rather, it is used as a consistency-based approach that combines an assumed initial ratio, differential OH loss, and downwind observations to infer photochemical aging and assess whether the inferred ERs remain physically reasonable. To address the reviewer's concern more directly, we added/clarified evaluation tests in which (1) the approach is applied to model output, where the initial ERs are known, and (2) near-source observed benzene/toluene values are used as an internal consistency check. These tests provide an estimate of the method uncertainty and help demonstrate where the approach performs reasonably well and where it does not.

Second, we agree that the toluene/benzene ratio becomes increasingly uncertain for the most aged plumes. In particular, for photochemical ages of ~7.5 days, primary toluene is expected to be strongly depleted and the inferred age/ER becomes much more sensitive to background toluene. We therefore revised the manuscript to explicitly state that the oldest data points (>5 days) are subject to large uncertainty and should be interpreted cautiously. Our quantitative interpretation is focused primarily on the younger subset of the data (<4 days), for which toluene remains substantially above background in many cases and the method is more defensible.

Finally, we now discuss more clearly that additional uncertainty arises from the assumed background concentrations. The corresponding uncertainties have been reflected in the supplemental materials, and we have softened the conclusions accordingly. We hope this can resolve the reviewer's concerns.

3. Figure 6B: There are insufficient measurement data points to support the proposed non-monotonic relationship between O3 and PM2.5. However, I agree with the broader conclusion that the model overpredicts O3.

Response: We agree that the Missoula dataset alone is too limited to define a precise quantitative threshold at which O3 transitions from enhancement to suppression as PM2.5 increases. Our intent here is therefore not to derive an exact functional relationship, but to highlight the non-monotonic behavior evident in the observations. This interpretation is supported by the two

additional western U.S. sites shown in the same figure set and is also consistent with previous literature documenting O3 suppression under heavy smoke.

More importantly, we wish to clarify that the key contribution of this analysis is not simply the observation of a non-monotonic O3–PM2.5 pattern, nor merely the finding that the model overpredicts O3, both of which have been discussed previously. Rather, our contribution is to show that the CTMs fail to reproduce the observed non-monotonic relationship between wildfire smoke intensity and MDA8 O3. This missing suppression under heavy smoke may help explain, at least in part, the well-documented positive O3 bias in wildfire-influenced model simulations. We have revised the text in Lines 500-520 to make this point more clearly.

4. Health risk discussion: The discussion of chronic and acute health risks is outside my area of expertise. I recommend that this section be carefully evaluated by reviewers with relevant expertise.

Response: We appreciate the reviewer's note. The health-risk calculations follow established inhalation risk-assessment frameworks, and the full methodology and toxicity reference values are documented in the Supplement (Sect. S2).