Response to Reviewer #2

We thank Reviewer #2 for their valuable comments and suggestions, which have helped make the study clearer and more organised. We believe the manuscript has improved significantly after addressing these points, and we appreciate the opportunity to convey our message more effectively. We have reproduced Reviewer #2's comments below in black text and numbered the comments for clarification when addressing comments relevant to both referees. Our responses are in blue text, and any additions to the manuscript are in red text. Our reference to line numbers is based on the initially submitted manuscript.

General comments

1. The results and discussion section is incredibly dense and hard to follow. The authors have performed so many sensitivity tests and created so many acronyms for them that it is hard for the reader to understand what's going on and the aims of the paper risk being buried. Additionally, many of the plots are very busy (although I appreciate that large-format versions would be available in the final online version of the paper). At the very least, I would suggest the authors seriously consider whether all the information in the results section is necessary to be included in the main section of the paper and consider moving some plots to the supplementary info. I would suggest splitting the results and discussion, which would enable the thread of the paper to be easier to follow. The paper would also benefit from a clearer statement of the objectives and demonstrating how each analysis section is designed to address them.

We thank Reviewer #2 for their various suggestion to improve the readability, clarity and structure of this manuscript**.** We have made numerous changes to the manuscript following both reviewers' comments to make it easier to follow. Below, we highlight the major changes:

i. We have reduced the number of sensitivity tests by eliminating some redundant experiments related to the factors BA, EF, and HDI. The sensitivity analysis section has also improved the explanation of the experiments following Reviewer #1 4.d comment and the result section, where we focus our analysis on the comparison between the experiments and the control run.
ii. We focused more on the comparison between INFERNO and the inventories than on the comparison between the inventories, as this is not the main aim of the current study.
iii. We have reduced the use of acronyms. We changed the abbreviation of some PFTs and experiments, removed the abbreviation from some PFTs and infrequently used words (more information in the response to Reviewer #1 4.c).
iv. We have changed some plots, making them simpler to interpret (e.g. Fig. 3 and 8). We have also removed and added supplementary figures in the main manuscript, which better support the analysis (more information about changes to the plots can be found in the response to Reviewer #1 4.a).
v. As suggested by Reviewer #2, we have separated the results and discussion sections for better clarity.

Finally, we have now clearly stated our objectives in the last paragraph to the introduction, outlining the sections intended to address them. Below, we show the change to the last paragraph of the introduction in the revised manuscript (l65):

"[…] previous studies have primarily focused on carbon emissions from fires, whereas this study aims to evaluate the simulation of fire-derived emissions in atmospheric models. We

seek to identify areas for development and improvement by analysing the biases associated with these emissions. In Section 3, we present our results focusing on the evaluation of INFERNO against fire emission inventories, use sensitivity experiments to investigate key drivers influencing CO emissions and finally quantify the model processes contributing to the calculated model-inventory biases. [...]"

2. A large section of the paper is devoted to comparing different satellite-derived fire products against each other. While this is valuable, I think some of the finer details could be moved to a methodological note in the supplementary info. As is often the case, the authors find that the differences between the observational products are about the same as the differences between the model and the observations. With this in mind, the approach used which averages multiple fire datasets to produce a benchmark is not sufficiently justified. The authors exclude FINN (in which case, it could be removed from the analysis in the main text) but include GFED4 in the averaging even though it differs substantially from GFED5 (for example). The reasoning behind the averaging of multiple fire products needs to be more clearly explained.

We thank Reviewer #2 for pointing out the potential misalignment in our inventory comparison with our objective of evaluating INFERNO. Therefore, we have significantly reduced these comparisons in the manuscript, but still provide a small section on comparing inventories to acknowledge the spread in emission estimates in Section 3.1.

In response to Reviewer #1's suggestion, which is in line with Reviewer #2's comments, we have removed the analysis on the FINN inventory from the main manuscript. Regarding the averaging across multiple fire emission inventories, we acknowledge that this is not an optimal benchmarking method, especially considering the methodological and regional differences among these inventories. However, in this study, we wanted to encapsulate the variability in estimates of CO emissions from fires across a range of the most up-to-date emission inventories. Both GFED4s and GFAS are well-established inventories in the literature, and while they have outperformed other inventories in emission evaluations in South America (Hua et al., 2023, Reddington et al, 2019). Their biases have also been highlighted (Naus et al., 2022, Liu et al., 2020). In contrast, 3BEM-FRP and GFEDvn5 represent newer-generation inventories adjusted with finer-resolution satellite data and updated land cover inputs. Because they include smaller fires, they tend to estimate higher emissions and are expected to be more accurate than GFEDvn4s and GFEDvn5; however, they lack the extensive and long-term validation that GFED4s and GFAS have. In fact, GFEDvn5 emissions are still under development and were downloaded as Beta products.

Therefore, we have preserved the individual characteristics of each inventory throughout the study, equally weighting GFEDvn4s, GFASvn1.2, 3BEM-FRP, and GFEDvn5 Beta in our averaged dataset used in the machine learning section, to represent a best estimate of fire CO emissions. This decision reflects a cautious and inclusive approach, acknowledging both the reliability of well-established inventories and the potential improvements offered by newer ones. Ideally, we would implement a performance-based weighting or subsampling strategy using an atmospheric model and compare the results with Total Columns CO (TCCO). However, this level of analysis is beyond the scope of our study and is therefore acknowledged in the limitations at the end of the manuscript.

Here is the new addition to l133 in the original manuscript:

"For the machine learning analysis of INFERNO biases (Section 2.6) and for visualising differences in selected figures, we calculated an ensemble average (mean) dataset based on the four inventories: GFED4s, GFED5, GFAS, and 3BEM-FRP. Each inventory was equally weighted in the average. GFED4s and GFAS are well-established inventories in the literature. While they have outperformed other inventories in their emissions estimate in South America (Hua et al., 2024; Reddington et al., 2019), their biases have also been noted (Naus et al., 2022; Liu et al., 2020). In contrast, 3BEM-FRP and GFED5 (the beta version) are considered next-generation inventories. They have been adjusted to better represent small fires and include updated and more accurate land cover data (Mataveli et al., 2023). However, these newer inventories lack the extensive long-term validation that GFED4s and GFAS have undergone. Overall, using an average of these inventories represents a balance between incorporating innovative methodologies and relying on well-established datasets for this study."

3. It is unclear why total column CO observations are being used in this analysis. The modelling setup used produces CO emissions, but as it is a land-only model it is confusing to the reader to imply that atmospheric modelling has been done as well. To evaluate the model using TCCO you would need to simulate the atmosphere as well. Atmospheric transport of CO, a species with a long atmospheric lifetime, along with the topography and regional circulation of SA, means that TCCO is not necessarily co-located with peak CO emissions from fires. Additionally, some of the analysis is limited by the need to consider shorter timescales as a result of limitations of the TCCO datasets. I suggest the authors remove the TCCO analysis from the paper and stick to CO emissions. Using the emissions produced by INFERNO to force an atmospheric chemistry model, and comparing that to observed TCCO, would be a logical next step but would be beyond the scope of this paper.

We appreciate the concerns of Reviewer #1 and Reviewer #2 regarding the use of total column CO (TCCO) in our analysis. It is important to clarify that we include the TCCO not as a benchmark for evaluating emissions. The emissions and TCCO are linked as CO is emitted and then undergoes meteorological and chemical processing. So, while not directly comparable, the TCCO can provide useful information on emissions from space, increasing the robustness of the results. For instance, the TCCO trends for South America, where fire emissions dominate the CO balance (Lichtig et al, 2024), show general agreement with inventories and support a weak decrease trend in CO emissions. However, we recognise that our original wording of the introduction on line 68, where we stated, "we compare the carbon monoxide (CO) emissions from fires simulated by JULES-INFERNO with various biomass burning inventories and satellite-retrieved total column CO (TCCO)," was misleading. Therefore, while we believe the TCCO does have merit in this study, as both reviewers have questioned its application here, we have removed it from our analysis to avoid any confusion.

4. I will confess that I am not an expert in the use of ML detailed in this work. However, the way in which certain variables were excluded from the ML model seems to be limiting. My understanding of this approach is that it can handle large numbers of covariates and indeed that this is a strength of the approach. It would be interesting to see if the results are sensitive to the choice of variables excluded, or to run the analysis with all variables. The use of soil moisture to represent leaf and wood carbon in particular is a concern – these variables strongly covary, but are calculated in very different ways in the land

model. By using one set of processes to represent another, the authors may be limiting the power of the analysis especially given the goal is to evaluate model processes.

We would like to thank reviewers #1 and #2 for their suggestions to run the XGBoost model using all variables without removing any correlated features. After further review of the literature, we found that gradient boosting models are well-suited for handling correlated features(Power et al., 2024). Based on the reviewers' comments, we have now used all INFERNO input variables within an XGBoost model, increasing the number of selected features used from 14 in the original analysis to 20. The results were compatible between the two XGBoost models using different numbers of inputs. The variables that ranked highest in the original feature importance analysis (based on SHAP values) retain their relative positions when using the full 20-input variables approach, as Fig. R1 illustrates. This approach also highlights the importance of other variables, such as Tropical Broadleaf Evergreen Trees. In the original manuscript, the contribution of this tree PFT was difficult to distinguish from its highly correlated soil moisture. Variables that also correlate with soil moisture, such as wood carbon, did not rank as high, which is another piece of new information that the new approach brings. Additionally, the model performs better, increasing the proportion of variability explained from around 64% to 67% based on the coefficient of determination R^2.

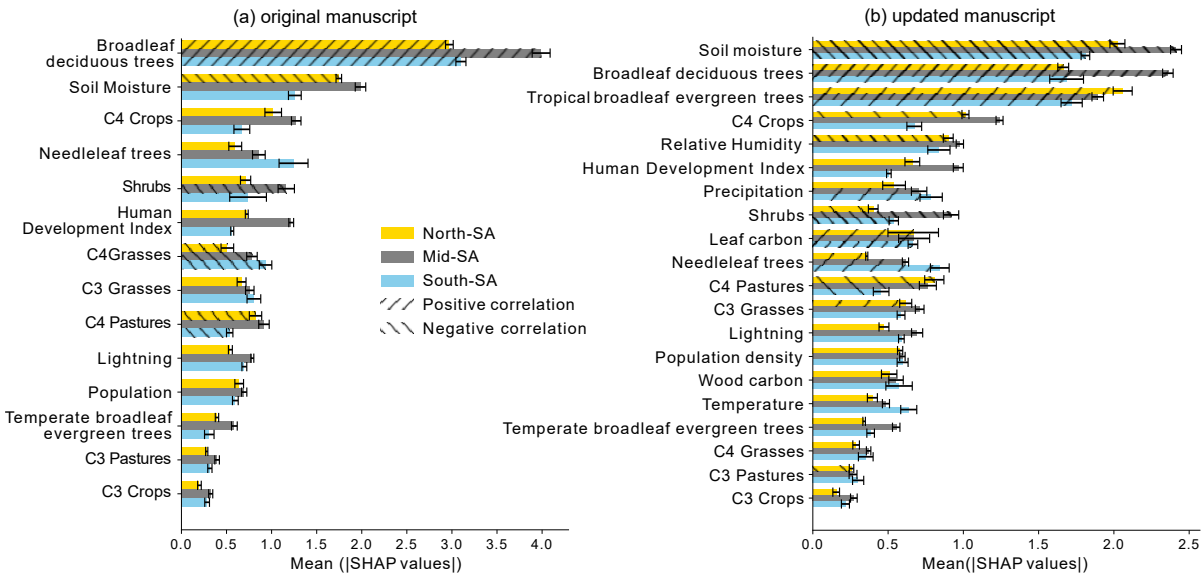

Figure R1. feature contribution to the XGBoost predicted biases using absolute SHAP values from the (a) original manuscript and (b) updated manuscript. (b) was updated in Fig. 9 of the original manuscript, Fig. 8 in the updated manuscript.

We add the following to the methodology section on l279:

"We selected 20 inputs for the machine learning model, comprising prescribed data and JULES outputs used by INFERNO to calculate emissions. [...]. We enable correlated features in the machine learning model, as in a gradient boosting model, any redundant information is automatically disregarded. This happens because the decision trees are built by splitting features in a series of dependent trees, so they can not make identical splits using correlated features (Power et al., 2024)."

5. The paper contains a number of typos and instances of awkward wording, which are fixable with thorough proof-reading. I do not believe it is the responsibility of peer

reviewers to perform the work of a sub-editor, though I have pointed some examples below.

We have now gone through the manuscript and updated the text where appropriate.

Specific comments

We thank the reviewer for going further in their role by pointing out typos and sentences that required rewording and editing. Below, we have checked (✓) the comment that has been corrected. Further descriptions of the change have been added to the comment where needed:

- ✓ Some of the authors' names seem to be misspelled in the submission, based on the professional profiles found on the websites of their universities ('Veláquez-García' = Velásquez-García, 'Chiperffield' = Chipperfield).
- ✓ Line 7: Specifically, this should say 'carbon monoxide (CO) emissions estimates'.
- ✓ Line 10: Remove 'categories of'.
- ✓ Lines 16-18: It is not immediately clear what the percentages in this sentence refer to.

The percentage mentioned in the abstract refers to the mean bias percentage. Since this is a measure of bias or relative difference, we have chosen to refer to it directly as such. Below is an example of how the percentages are currently presented in the abstract:

"[...] in the Arc of Deforestation (southern Amazon), INFERNO tends to overestimate CO emissions by around 70%. [...]. An experiment shows that INFERNO produces up to 100% higher CO emissions in the Amazon region when using drier meteorological reanalyses"

and refer to the metric in the brackets. Note that instead of comparing the experiments and the control based on their performance (i.e. comparing with the inventories), we are now comparing the simulations directly, which is why these numbers have changed.

- ✓ Lines 26-27: Unclear wording ('the success of fire-prone ecosystems is enhanced').

We meant "succession"; however, we removed this sentence while rewriting a part of the manuscript to make the objective of this study clearer (Review #2.1) and passed the manuscript through proofreading.

- ✓ Line 45: Suggest 'represent' or 'simulate' instead of 'understand'.

We have chosen "represent" as a more accurate word for the message.

- ✓ Line 61: Where does the figure that SA represents 15% of annual fire carbon emissions come from? Over what time period?

Thanks for highlighting the missing reference. Here, we have added the corresponding text citation for this contribution as:

"[...], as the region contributes with around 15% of annual global fire carbon emissions (van der Werf et al., 2010)."

- ✓ Line 311: Should be $Tg\ yr^{-1}$.

We have corrected this throughout the manuscript.

- ✓ Figure 2: The colour bar used for the annual mean emission plots is not perceptually uniform and features large non-uniform breaks at arbitrary intervals. Please use an

appropriate colour bar – if these plots were made using Matplotlib, which they appear to have been, there are several available (see https://matplotlib.org/stable/users/explain/colors/colormaps.html). In addition, for the CO total column plots, the colour bar used is diverging which is inappropriate for displaying a continuous variable which is not a difference.

Thank you for bringing this to our attention. We aim to ensure clarity for a diverse range of readers, so we have updated the colour bars as suggested.

- ✓ Line 364: 'Andes' rather than 'Andean (mountain range)' is sufficient.
- ✓ Line 365: 'Accumulates'.
- ✓ Line 395: 'Despite the peak of precipitation being…', also there appears to be an incomplete citation here '(Grimm)'.
- ✓ Table 3: The caption should explain why some of the numbers are in bold type. There is also a more general question here about whether short-term trends of <10 years are meaningful given interannual variability and the complex political environment in South America (which the authors describe well in this section from a fire perspective).

We concur with Reviewers #1 and #2 that the 8-year period presents significant uncertainty due to the limited time sample and the high variability in emissions. This uncertainty is exacerbated by our use of annual CO emissions to calculate trends. Consequently, we have decided to exclude the short-term trend from our study.

Regarding Table 3, we added information in the footnote of the table describing the meaning of the trends in bold:

Note: The magnitudes of trends are highlighted in bold when the Mann-Kendall test indicates a significant trend at the 95% confidence level (p-value < 0.05). […].

- ✓ Figure 6: This figure would benefit from hatching/stippling to denote where the biases are statistically significant according to an appropriate test.

Thanks for the suggestion, we included the hatching not only in that figure but also in other places where we wanted to highlight the significance of the calculated test. We decided using the hatching where the test was not significant, since this can shade the valuable information that significant values present.

- ✓ Line 509: 'Differed'.
- ✓ Line 515: 'Has', also 'simulated' rather than 'estimated'.
- ✓ Line 523: 'Increased' or 'enhanced', not 'extenuated' (this does not mean 'extended').
- ✓ Figure 8: Should be 'Spatio-temporal'.

This figure was changed to a simpler, more effective figure that shows the experiments' changes relative to the control run.

Lines 591-592: There are words missing and therefore the sentence does not read correctly. '…landscape fragmentation, which represents both….and can lead to…fire suppression effects'.

This sentence belongs now to the discussion session, as it was separated from the results section following Reviewer #2.1's comment. This sentence has been reworded as:

"In a fire model, crop representation needs to include the agriculture management cycle (Li et al., 2013), the influence of socioeconomic factors on management practices (Li et al., 2013),

agricultural expansion, and landscape fragmentation (Silva-Junior et al., 2022), among others. Crops representation would also need to include the agricultural role in fire suppression (Haas et al., 2022)."

- ✓ Line 602: Remove 'feature'.
- ✓ Line 619: This sentence is incomplete: 'Furthermore, agricultural expansion and landscape fragmentation.'. What about them?
- ✓ Line 645: 'TCCO' (although I suggest removing this variable entirely).
- ✓ Line 658: This should be '(broadleaf deciduous trees (BDT) and broadleaf evergreen tropical trees (BET-Tr))'.
- ✓ Line 662: 'Sensitivity'.
- ✓ Line 663: 'Small' rather than 'short'.
- ✓ Line 664: 'PFT'.
- ✓ Line 664-665: Awkward phrasing. Try 'Both improving PFT accuracy and incorporating representation of human land-use management, through variables such as land fragmentation, might help reduce biases'.
- ✓ Line 677: 'Cut'?
- ✓ Line 677: It should be noted that this code is only accessible to people with a Met Office account; for me it returned a login page. If the underlying model code is not publicly accessible, a statement is required to explain why; additionally, there is no link provided to the model output, which the journal also requires (or, in the absence of this, a statement explaining why the data are not being made publicly available).

We thank Reviewer #2 for pointing out the missing information required to access the model code. We also wanted to add that it is in our plan to share the model results via an open access platform such as Zenodo if this manuscript is accepted. To provide clear, detailed information for accessing the code, we have added the following in "Code and data availability" l 677:

The JULES-ES control configuration (based on JULES version 7.5) is stored at [https://code.metoffice.gov.uk/trac/roses-u/browser/d/l/3/2/3/trunk, last access:11 March 2025]. JULES and associated configurations are freely available for non-commercial research use, as set out in the JULES user terms and conditions [http://jules-lsm.github.io/access_req/JULES_Licence. pdf, last access: 10 November 2025]. For a comprehensive guide to accessing, installing, and running the configurations, we direct the reader to Appendix A in Wiltshire et al. (2020). Note that to view and use the JULES-ES source code, access will be required via the Met Office Science Repository Service [https://code.metoffice.gov.uk/trac/home, last access: 10 November 2025], and is available to those who have signed the JULES user agreement. The easiest way to access the repository is to complete the online form to register at [http://jules-lsm.github.io/access_req/JULES_access.html, last access: 10 November 2025].

- ✓ Line 678: 'Are downloaded'; also, I suggest putting the dataset links in brackets.
- ✓ Line 686: This doesn't make sense: 'Some assessments were done using the deforestation front for 2020 provided at and the ecoregion provided at'.
- ✓ There are some typos in the reference list, and the DOIs are inconsistently stated with some references missing DOIs (e.g. Magahey and Kooperman 2023) and others having double entries e.g. line 872: https://doi.org/https://doi.org/10.1007/s10531-019-01720-z. Presumably these are artefacts introduced by reference management software but the reference lists that these produce should always be checked manually.

We thank Reviewer #2 for their detailed review. We have checked the reference list to avoid duplicate DOIs or missing information.

Added References

Lichtig, P., Gaubert, B., Emmons, L. K., Jo, D. S., Callaghan, P., Ibarra-Espinosa, S., Dawidowski, L., Brasseur, G. P., and Pfister, G.: Multiscale CO Budget Estimates Across South America: Quantifying Local Sources and Long Range Transport, Journal of Geophysical Research: Atmospheres, 129, e2023JD040 434, https://doi.org/10.1029/2023JD040434, 2024.

Power, J., Côté, M.-P., and Duchesne, T.: A Flexible Hierarchical Insurance Claims Model with Gradient Boosting and Copulas, North American Actuarial Journal, 28, 772–800, https://doi.org/10.1080/10920277.2023.2279782, 2024.

Reddington, C. L., Morgan, W. T., Darbyshire, E., Brito, J., Coe, H., Artaxo, P., Scott, C. E., Marsham, J., and Spracklen, D. V.: Biomass burning aerosol over the Amazon: analysis of aircraft, surface and satellite observations using a global aerosol model, Atmospheric.

Chemistry and Physics, 19, 9125–9152, https://doi.org/10.5194/acp-19-9125-2019, 2019

Wiltshire, A. J., Duran Rojas, M. C., Edwards, J. M., Gedney, N., Harper, A. B., Hartley, A. J., Hendry, M. A., Robertson, E., and Smout-Day, K.: JULES-GL7: the Global Land configuration of the Joint UK Land Environment Simulator version 7.0 and 7.2, Geoscientific Model Development, 13, 483–505, https://doi.org/10.5194/gmd-13-483-2020, 2020.