

“Understanding drivers and biases of simulated CO emissions by the INFERNO fire model over South America”

The paper provided for review characterizes INFERNO’s accuracy by comparing predictions to benchmark inventories. Comparisons are made at a monthly $0.5^\circ \times 0.5^\circ$ scale for as much of the period of 2004-2021 as each benchmark allows. Fidelity is assessed for total emissions, seasonal amplitude, and time-trend slopes. The main conclusion from the summary statistics is that predicted emissions for the center of the continent, which dominate the totals, on average are too high and excessively seasonal. Total CO is more accurate for the less influential southern and northern regions, although the latter shows perverse seasonality. A series of experiments perturb several of INFERNO’s few inputs. The experiments suggest that incorporating the Human Development Index could improve predictions. Finally, an XGBoost model of prediction errors indicates that the strongest associate of large errors is relative dominance of rainforest vegetation, with its elevated potential for complicating effects on fire from deforestation.

General comments

1. Raw total column CO is not among the appropriate benchmarks for fire emissions. First, CO emissions from fires can move and mix over South America for weeks if not months. The CO over a study cell thus does not represent the fire CO that the cell generated. Due to long persistence, even spatial average column CO is not an instructive benchmark for monthly emissions. As you note (l. 364), the Andes tend to hold and deflect low-altitude smoke. A comparison of fire locations to total CO displays coarsely that the places with heaviest smoke are west of the places with the most fire (Fig. 1, below). The map comparison is clouded by difference in study periods, spatial differences in fuel loads and combustion completeness, and non-fire CO sources. But the maps do illustrate the general spatial consequences of smoke transport in South America.

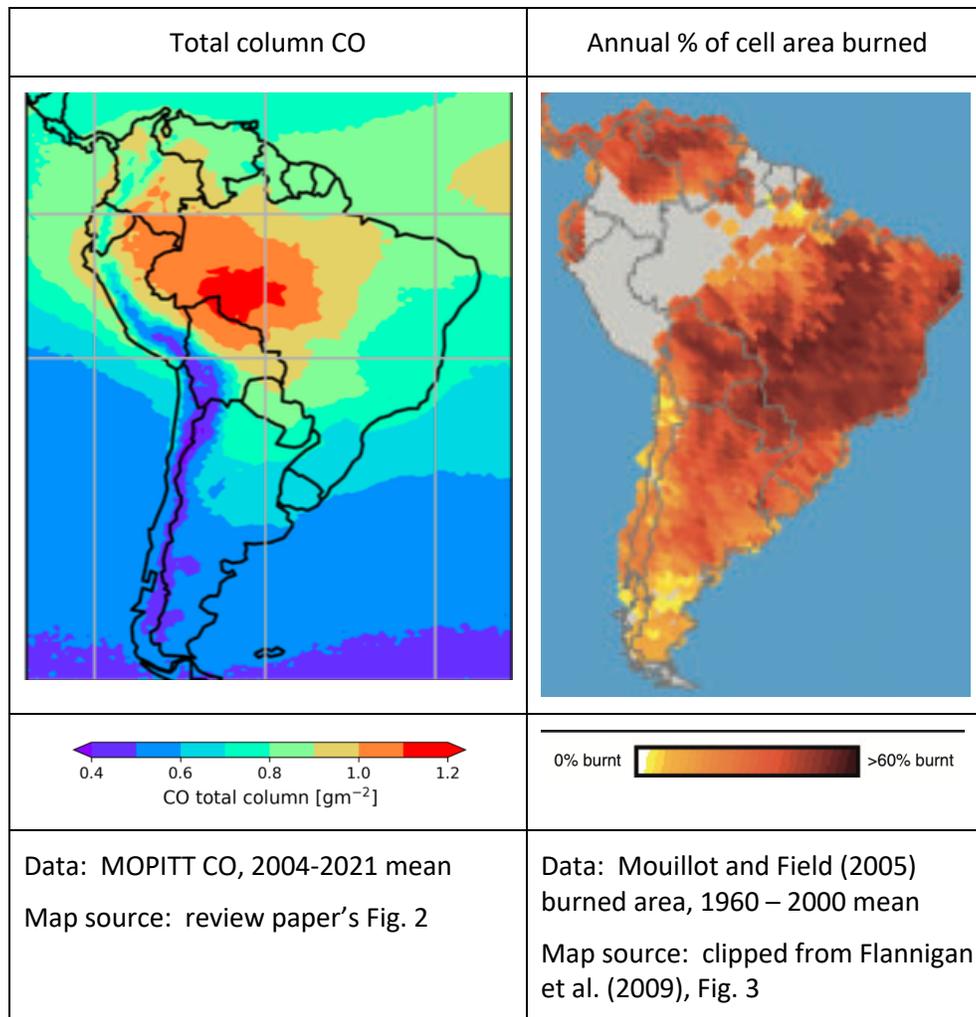


Fig. 1: Rough comparison of typical fire locations with typical CO locations for South America

Second, other CO sources need to be backed out from totals before the remaining CO can appropriately be attributed to fire. A modest offset to the severity of this issue is that in most of the Amazon basin biomass burning is the source of most CO emissions. Proportionally more is directly anthropogenic in some other parts of South America.

In support of using total column CO, this study cites four papers. Each cited study interposes a transport model in between fire inventories and column CO. The review study does not, and should. For all but the last study listed below, the models used incorporate atmospheric chemistry and multiple sources of CO.

- Hua et al. (2024) provide each of four fire inventories to CESM-CAM before comparing model results to MOPITT. (Please see source papers for unabbreviated model names, etc.)
- Liu et al. (2020) feed 5 fire emissions inventories in turn to GEOS-Chem, using the adjoint model's standard estimates of background CO. They compare model results to a ground-based PM2.5 monitor.

- Naus (2022) et al. invert TM5-4D-Var to model fire sources from MOPITT CO, then benchmark their results against one fire inventory and flask data collected by aircraft.
- Jury and Gaviria Pabón (2021) explore patterns of smoke transport in South America, including by running emissions from an imaginary fire through the relatively simple HYSPLIT transport model.

This paper's summary of the Hua and Liu studies that "inventories generally align well with CO emissions in SA" overlooks core methodology differences. The assertion does not justify this paper's method of using total column CO.

2. There are multiple issues related to statistics.

a. Shapley additive explanation (SHAP) values are described as "the expected marginal contribution of a feature... It is calculated by taking the weighted average of all possible subsets of the selected features in which the specific feature can contribute." (l. 293-295) A SHAP value is not the expected marginal contribution of a feature. That is fortunate, because comparisons of marginal contributions among strongly correlated features easily becomes misleading. The strength of SHAP scores is that they are independent of ordering, meaning *independent of marginality* – even though for ordinary least squares models they are calculated from numerous marginal contributions (Grömping, 2015). With respect to the summary in the submitted paper, this is quite different from the average of subsets of the features. SHAP values for machine learning models are conceptually parallel to Shapley values including in the sense that they evaluate feature contributions independent of marginality. A reference to scikit-learn (l. 288) suggests that this study used Python's SHAP command. A package citation would be welcome.

SHAP values can be calculated for a model whether or not the commendable complication of cross-validation is provided. The following asserted connection therefore does not hold.

"We calculate the SHAP values using a five-fold cross-validation approach... This allows us to calculate a complete set of SHAP values for the whole dataset and calculate a map of contribution[?s]." (l. 296 – 299)

b. Target Average Relative Ranges: Predictions that have perfectly accurate responsiveness to predictors reproduce only the portion of variability (r^2) that the predicting model can explain (Farmer & Vogel, 2016). With imperfect models, an unavoidable tradeoff therefore exists between reproducing realistic variability, and reproducing realistic responsiveness to predictors. An example illustrates the consequences. In a generic fire prediction model, exaggerated responsiveness to drought would, up to a point, cause average relative ranges to be closer to 100% than would a model with more realistic responses to changes in moisture. If one wants realistic reactivity to predictors, it is a failure, not a success, if INFERNO's average relative range is as large of those of the emission inventories.

c. Feature exclusion: Your use of machine learning to analyse prediction residuals (Lundberg et al., 2020) s lets you differentiate between the importance of omitted features and that of improvements that in theory may be possible via different handling of existing inputs. While extending your analysis may not be feasible, I wonder what might be learned by further exploration of the features with the highest SHAP scores. For example, a plot of residuals versus share of broadleaf evergreen trees might hint whether a different transformation before use in INFERNO would be better. Perhaps there are ranges of the PFT percentage where predictions are better or worse. Or if the problem is more or less linear, maybe omitted features that are mostly specific to rainforest vegetation, like patterns of deforestation, are needed. Or correlations and scatterplots of the residuals with interactions of each

other input times broadleaf evergreen tree percent might suggest that the form of a specific non-linearity need adjusting. Etc!

In running XGBoost you excluded any feature whose covariance with an included feature exceeds 0.6 (l. 564). (The methods section near l. 284 is a better place to describe such choices.) The consequences of the paring rule used seem to hamstring the analysis. Large numbers of features typically do not hamper resolution of machine learning algorithms (Breiman, 2001). Interpreting models based on data with numerous features is difficult, especially in the presence of strong multicollinearity. But paring does not obviate the difficulty. Omission amounts to throwing away all unique information content in excluded predictors. Paring predictors may appropriately improve ease of use for equations whose primary purpose is predicting (Hastie et al., 2009, p. 58). But if the modeling goal instead is causal inference, as here, then subsetting based on a numeric rule limits the questions that can be asked of the data (Judd et al., 2017, p. 132).

An example illustrates that dropping an important but highly cross-correlated feature can lead to conclusions of dubious usefulness. The feature selection in this paper assumes that soil moisture represents the importance of leaf carbon as a diagnostic predictor in INFERNO. Within an ESM, soil moisture and carbon stock values have very different derivations. If the real problem happened to be well-shared or held entirely by leaf carbon, then knowing that soil moisture is critical when soil carbon is omitted from the available predictors helps little. The art is to pare while retaining statistical access to the realities one wants to understand. I suggest trying runs of XGBoost with all features, and/or analyzing results of the current runs with considerable attention to the implications of initial paring.

Less critical concern: The choice of a VIF cutoff is a matter of professional opinion rather than a consequence of an objective discontinuity. A citation therefore is needed on l. 284. Your description is slightly muddled (l. 281-2). A “correlation between pairs” describes how much of the variability of one feature can be explained by the [one] other, not by “the other features”. A VIF, on the other hand, does compare one feature to all others.

d. Trend confidence intervals (might need only clarification): This paper compares trends found in the emissions inventories to trends in INFERNO’s predictions (section 3.1.2). Fig. 4 shows that you have appropriately calculated trends one cell at a time (e.g., Andela et al. (2017)), which avoids pseudoreplication. There are two ways to calculate trend confidence intervals for large areas. I do not see the needed documentation of which you used. The conservative and prudent approach is to sum each year’s emissions for the large region, then calculate a trend and its confidence interval by treating each year as one observation. Confidence intervals will be large, and should be. The alternative is to treat each cell’s trend as independent and average the trends across space. Averaging across cells requires relatively complex adjustments for spatial autocorrelation in order to provide accurate confidence intervals. Please address my uncertainty whether sound confidence intervals support, for example, the precision - and implication of comfortable fidelity - in l. 409-410’s assertion that “most inventories and INFERNO suggested a negative trend of approximately $-1.0\%yr^{-1}$ ”.

You calculate trends for two time periods, one only 8 years long. High interannual variability in fire incidence is globally generic, especially in forests. A result is that for trends annual fire incidence has a low signal to noise ratios. It is improbable that any real and persistent trend would become evident with statistical confidence in only 8 years. The widespread contrasts even in sign for specific locations between the two rows of maps in Fig. S5 illustrate this point. If the total column CO dataset is dropped from this study, the exacerbated problem of the trends being calculated on the shorter time period may take care of itself.

e. Cross-inventory averaging (less critical concern): In places you average across four inventories and used the averages as the benchmark (e.g. l. 484). Where this is done, and why, needs to be indicated more clearly.

I question including GFED4s in averages. Averaging implies that you have equal confidence in the accuracy of each included inventory. The description of GFED5 (Chen et al., 2023) dwells on the reasons for each change in the newer version. If you nevertheless feel GFED4s is worthy of inclusion in averages, that deserves justification.

Does including FINN enhance this study? Many papers already agree with you that FINN is out of line with other emission inventories (l. 483). You sensibly leave it out of some averages. But this begs the question of why FINN is included in the first place. What have we learned about INFERNO at the end that only FINN could teach us? If you agree but want to document your thoroughness, say you tried FINN and it didn't work well, then be done with it.

f. Geographic cutpoints (less critical concern, l. 88): Why geographic blocks rather than (groups of) PFTs, or the ecotypes shown in Figure S1? Why 3 instead of 2 or 5 regions? Replicating the division used in Li et al. (2024) (l. 90) might be useful if comparison to that work's findings were central to the analysis of this study's findings, but instead the reference is not subsequently mentioned. Beyond the connection to the other study, why are the specific latitude choices optimal? I see that separating the northern region helps for the analysis of seasonal cycle. The southern break, however, is troubling, mainly because it chops in two the cerrado, one of the continent's most naturally fire-prone regions. A diagonal line would more effectively segregate the rainforest and/or deforestation zones. Ideally the zones would be more balanced with respect to size or total fire emissions, so that each would deserve more nearly equal attention.

3. The paper provides too little assessment of the study's results in the context of other research.

Examples: What do your results suggest about the effectiveness of various aspects of INFERNO's fundamental structure? Is it workable that INFERNO takes no account of deforestation rates (Li et al., 2013), which you suggest are crucial (l. 439 – 447; the region with high unexplained error in Fig. S11; BDT's high Shapley value in Fig. 9)? Does this work recommend any adjustment to the response shape for fuel load or soil moisture (Teckentrup et al., 2019, Fig. 5)? How do your findings support or differ from other ESM fire model sensitivity analyses? Overall, the analysis feels partly digested. Comparisons to previous work will help clarify and simplify what we have learned, and perhaps need to change, about INFERNO as a result of your efforts.

4. More editing could make this paper a smoother reading experience. I am disappointed still not to have a good grip on what you think should be done to INFERNO going forward and why, or what insights you have provided for people who will use and interpret INFERNO's predictions. Neither do I feel clear about the robustness of the apparent conclusions.

- Every figure would be improved by simplifying. If all the current detail really is needed as reference, move the details to the supplemental.
 - It is not apparent to me what your clear, simple intended take-away message is for any of the figures. For Figure 1, for example, what do you want a reader to notice about the contrast between the two maps – or is the message related to the distribution of PFTs and similarly apparent in each map?

- To my thinking, Fig. S1 belongs in the main paper. Fig. 1 could move to the supplemental.
- For each figure that currently has more than one map, I suggest including only one. The one might be an example, with all others in the supplemental if needed. If the message is some contrast, show only a single pair of maps.
- The black polygons in Fig. 5 are unreadable (and here and in Fig. 4 there are two sets of black polygons, one for nations). Since I'm not sure of the figure's message I don't know what to suggest instead. A single full size map? A blown-up example area? A map only of points whose trend is statistically significant?
- Many paragraphs, especially in the results section, are simply difficult to follow. Others ramble and want tighter phrasing.
- The use of abbreviations is immoderate. The simplest illustration is to suggest that you compare the portion of sentences that contain at least one abbreviation from any page in this paper to a page of another paper you admire.
 - Most of the paper's abbreviations would better be replaced with words: ML, ARR ('relative seasonality?'), BA, MB, all the PFT names, SA ('northern region' for 'North-SA?'), TCCO ('column CO?'), and more. Many readers will understand ENSO without a second thought, and many may be as familiar with CO or even MOPITT. But the more obscure the abbreviation, the more likely a reader must stop to decode. One result is potential for confusion. Another is that we will the less enjoy and absorb the material offered.
 - Self-generated abbreviations are the worst. I strongly recommend replacing every abbreviation in Table 2 with a descriptive name, eliminating the codes from the paper. As it is, a reader trying to puzzle through lines 508-512, for example, probably is sentenced to scrolling pages back to find Table 2 more than once. By then, flow, and concentration on your message, are thoroughly disrupted.
 - MATOPIBA, FARC, and any other abbreviations with no subsequent reference can be eliminated. The combined needle-leaf PFT predictor that you created for the machine learning equations (l. 276) appears not be mentioned again, and if so there is no need to include the NT label. I did not check whether the same opportunity applies to '(Sh)'.
 - Aerosol Optical Depth (l. 446) is never decoded. A systematic check is needed that each abbreviation is decoded at first use – one more reason to shorten the list!
- I regret that I continue to find it nearly impossible to follow what you are doing and why with each experiment. Below is one possible approach to reorganizing the portions of the methods section that describe INFERNO and the experiments.
 - a. Create a mini-section (around one paragraph) for each of the 7 experiment groups. Start each mini-section with a specific motivating research question based not on the model but on the real world. It's hard to provide an example because the text is currently so difficult to understand. Here's an attempt: The current assumption is that if a fire occurs, then at least x% of the site's fuel mass will burn up, and at most x%. (It appears that) you consider instead that a fire might consume any portion of the fuel, from virtually none to all.

- b. Next in each mini-section, explain how the existing understanding about reality works its way into INFERNO's calculations. Include the least amount of detail that is necessary for a clear description of the experiment.
- c. Third, describe what you will change in each experiment to explore your alternative description(s) of the real world. In the invented example, I think you will test allowing the model to use consumption completeness values from 0 to 100%.
- In the results, when you extract insights from each experiment in turn, relate the findings to the real world through the experiment's research question.

It seems likely that this or a similar organizational approach could obviate much of what now is in section 2.4.1, the description of INFERNO. As it is, without context it is difficult to know what to try hardest to remember as one reads. For all details not directly relevant to an experiment and maybe some that are, let an interested reader look in existing INFERNO documentation. A tighter focus may also reduce the number of equations needed in this paper.

- Many additional and salutary simplifications are readily available. Here are only a few examples:
 - Table 1: Is the NDT PFT applicable to this study? Any that are not can be omitted, with acknowledgement that the table is partial.
 - l. 153: "The JULES-ES configuration for ISIMIP3a was utilised in this study. This configuration is described in Mathison et al. (2023)..." to "We used the JULES-ES configuration for ISIMIP3a (Mathison et al., 2023)" (then maybe explain why).
 - To make the inventory names easier to read, you might omit 'vn': GFED5 for GFEDvn5, and probably entirely drop the version references for GFAS and FINN after the first mention.
 - The numbers for RH_low and RH_high could replace the variable names, simplifying the Eq. 5.
- Approaching the realm of picky details:
 - Many italicized unit designations are missing spaces between the abbreviated words: g kg⁻¹ instead of gkg⁻¹ (e.g. l. 124)
 - Choose one version of a word or abbreviation and stick to it: burned or burnt, per unit as either a negative exponent or following a slash (e.g. "/ m²" in Fig. 3 v. "yr⁻¹" in l. 326), year abbreviated as y or yr (both in Table 3).
- A thorough sweep would pick up many typos. There also are many opportunities to resolve awkward wording, especially choices of prepositions. Following are examples, but the point is to identify more:
 - l. 5: "... South America (SA);, a region ..."
 - l. 9: "most of the fire-active zones in ..."
 - l. 111: "While" appears to be an extra word.
 - l. 124: "The EFs are consistently derived or partially derived from..." – wording is hard to follow
 - l. 139: Due to its wide swath, the global coverage is achieved in 12 hours

- l. 144: Rather than decode all the abbreviations, which otherwise is needed, you might drop the description of all three products and simply say you chose the product that uses both thermal and near infrared bands.
- l. 146: "...and an optimal estimation-based algorithm to retrieve..." "and optimization to retrieve"
- l. 173: "which models the PFTs competition"
- l. 240: missing 'd'
- l. 241: mismatch of singular pronoun with plural antecedent
- l. 244: 'Man-Kendall' has a typo
- l. 314. "This" could refer to either of the two numbers in the previous sentence, the average relative range or the portion of total CO. The simplest way to prevent problems related to antecedents is to eschew generic pronouns.
- l. 327: typo in the version number for FINN, and l. 498, typo in an experiment abbreviation - though I hope both have become moot.
- l. 363: extra space between 'respectively' and comma
- l. 364: typo in 'not.'
- l. 416: "The emission also decreases..." to "Emissions decrease..."
- l. 439-440: "[I]nfractons against flora" looks like perhaps a literal translation. I do not know what it means and would prefer that the paper tell me the concept instead.
- l. 449: To what does "here" refer, and is it needed?
- l. 452: To reach the widest readership, I suggest you assume many are unfamiliar with each ecotype in the sentence. A possible alternative is to describe what those regions have in common, either as a characteristic or as a general location within the continent.
- l. 454: "large-scale interannual forcing" – plain English, please. Lack of rain?
- l. 456: "Derived" needs rewording, because deforestation did not emit the air.
- l. 486: Sentences with parenthetical alternatives are hard to digest without reading the sentence twice, so I do not recommend the admittedly common practice. Any kindness that helps a reader consider your message rather than trying to puzzle out that message is good. Similarly, on l. 555 put numbers immediately adjacent to each element of the list rather than grouping numbers and names separately.
- l. 489: "reduce CO emissions through the territory in about 60%" 'By' about 60%?
- l. 492: "produces the underestimation of the seasonal cycle amplitude on the three subregions" ?causes underestimation of seasonal cycle amplitude in all three regions
- l. 508: "Ignition" should not be capitalized
- l. 532: typo in 'Fig 6o'
- l. 844: Typo in article name. Please carefully read though the bibliography for any other issues.

Line-specific comments:

l. 66: "This study decisively addresses this gap by rigorously assessing the performance of the model..." Isn't the qualitative judgment of 'decisively' for your readers to judge rather than you? The assertion sounds overstated, which almost cheapens your paper.

l. 150: “Fair” is generically a judgment call in the eyes of the beholder, and therefore best used very sparingly in academic papers. Here instead be more tangible about what is gained by your choice.

near l. 183, Table 1: Why do some PFTs have emission factors of 0? Do those PFTs really need to be in the table at all? Before presenting Table 1, please tell the reader why we are seeing the information. For example, will you be changing each default value in an experiment? This suggestion is subordinate to restructuring the experiments introduction as recommended above.

l. 234: Soil underlies virtually all land, so specify throughout that the PFT is ‘bare’ soil

l. 238: Introduce each descriptive statistic first in terms of what you are trying to describe in the real world. In this instance, average relative range describes something like how well the model predicts the variability of a cell’s emissions.

(l. 238 con’t.) The description of average relative range’s calculation seems to imply that the metric is a unified summary across the entire study area. You subsequently report relative ranges for particular ecotypes (e.g. l. 353). The calculations for areal subsets should be mentioned in the methods section.

l. 244, 246, 251: Please see first comment re l. 238. What aspect of fidelity to the real world (and not simply the model) do you want each metric to describe?

l. 285: Rather than a list of variables whose names are ambiguous out of context, describe the outcome of the step – for example, which variables you changed, or say you optimized all in the procedure’s default list, or give some comparable description of why your choices matter.

l. 325: Fire frequency is a general term for a concept, but when associated with a specific number is spatially undefined. I think you mean fire return interval. The study you cite uses the two terms appropriately. Also, Júnior et al. calculate frequencies for only a portion of Brazil’s cerrado, and a portion whose fire management is not necessarily typical of the whole. It therefore seems inappropriate to describe the cited frequencies as describing to the whole ecoregion.

l. 369: Most of INFERNO’s emissions from South America are generated in the central region. Fig. 3 shows that while the relative discrepancies are obviously larger in the outer regions, the central region has the largest absolute discrepancies - even before taking into account that region’s greater size. Thus the central region is key to improving INFERNO’s predictions for South America. You recount differences among the benchmark inventories in detail in this paragraph, but say nothing about why INFERNO is higher than all but (unreliable) FINN or column CO. Previously (l. 359) you propose that the source of inaccuracy may be “the model resolution and simplified process representation” – a degree of vague generality that is distressing. How can you use the seasonal cycle analysis to assess and/or refine this possibility? The only inputs to INFERNO with strong seasonality are several in the fuel multiplier, and soil moisture in combustion completeness. INFERNO’s seasonal amplitude ought to be smaller than that of the benchmarks (please see note 2b above), but is larger. To which seasonable inputs is INFERNO excessively sensitive?

l. 385: You seem to speculate that high gross primary productivity in September and October in the grasslands of northern South America drives the unrealistic late-season peak in INFERNO emissions. Those months are late in the rainy season. INFERNO lacks a mechanism to account for curing to recognize that grasses barely burn when green. Is that, together with the accumulation of grass growth over the course of the wet season, what you think causes the prediction error pattern?

near l. 400, Figure 3: Please see note above about complexity of figures. Also, the horizontal lines for the means (?medians) are too hard to see. Please include the percentile cutpoints for the colored

portion of the boxes in the caption. Please make the panel letters bigger. I would find panel (a) easier to digest with a linear y-axis, with different y-axis ranges for each region.

near l. 435, caption for Table 3: Please note that the time period for mean emissions and seasonal amplitude is the full study period. Be explicit what the parentheses mean, which I assume to be standard errors of the mean, and bold, which I assume means statistical confidence that a value differs from zero. Why are standard errors for trends not included?

(Table 3, con't) Here or somewhere else, I would like to see simple correlations of cell-month values between INFERNO and each inventory.

(Table 3, con't) The time scales differ between Figure 3 and Table 3 – monthly in one, annual in the other. Unless there is a strong reason, make them consistent. After adjusting for time scales, is the data in the rows for emissions and seasonal amplitude in Table 3 the same as displayed in Figure 3? If not, I'm confused. If so, duplication in the table is a distraction. If you feel strongly that the numbers need to be documented in tabular as well as graphical form, move them to the supplemental. After that and after removing data related to column CO, the remaining data in Table 3 might more effectively be presented graphically.

~ l. 448, Fig 5: Might these maps be more informative as differences from INFERNO? Maybe as a single map, INFERNO minus the mean of 4 inventories?

l. 448: Where do we get a mean of 61% per year (or 85% on l. 451) when the map color scale tops out at 15 and much of the southern region has light colors?

l. 452: Please give the baseline for the 208% change. Is it the 4-inventory mean?

near l. 487, Fig. 6: I find the version in S6 plus a map of the baseline case to be more meaningful.

near l. 507, caption for figure 7: The word "type", like "class", "group", etc., conveys no inherent meaning. Please choose a more informative label. While this figure needs further simplification, currently it is the clearest of the figures. I suggest you muse about how to draw greater focus on the central region throughout the paper, to highlight its disproportionate contribution to continental totals and tendencies.

l. 539: "This can cause a spatiotemporal overestimation of around 50%." I'm not sure what you mean in referring to a spatiotemporal overestimation. Is 50% the maximum for any cell-year, for instance?

l. 570: Hyperparameter tuning does not select model parameters. Do you mean "best hyperparameters"?

near l. 590, Fig 9: I strongly suggest no abbreviations in the legend(s). By rearranging, you may not need to show the legend labels twice.

l. 630-631: "INFERNO was able to reproduce emissions in key active fire zones, such as deforestation fronts (e.g. Arc of Deforestation)..." To your satisfaction, really?? More generally, this section summarizes your discoveries about which subsets of predictions tend most or least to be accurate. It says almost nothing about what the accuracy patterns mean. *Why* predictions are worse or better when and where they are, and ideally therefore how to improve the model, are more useful insights. I think you have a basis to make reasonable speculations.

Thank you for the opportunity to participate in your study.

References

- Andela, N., Morton, D. C., Giglio, L., Chen, Y., van der Werf, G. R., Kasibhatla, P. S., et al. (2017). A human-driven decline in global burned area. *Science*, 356(6345), 1356–1362. <https://doi.org/10.1126/science.aal4108>
- Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3). <https://doi.org/10.1214/ss/1009213726>
- Chen, Y., Hall, J., Van Wees, D., Andela, N., Hantson, S., Giglio, L., et al. (2023). Multi-decadal trends and variability in burned area from the fifth version of the Global Fire Emissions Database (GFED5). *Earth System Science Data*, 15(11), 5227–5259. <https://doi.org/10.5194/essd-15-5227-2023>
- Farmer, W. H., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, 52(7), 5619–5633. <https://doi.org/10.1002/2016WR019129>
- Flannigan, M. D., Krawchuk, M. A., De Groot, W. J., Wotton, B. M., & Gowman, L. M. (2009). Implications of changing climate for global wildland fire. *International Journal of Wildland Fire*, 18(5), 483. <https://doi.org/10.1071/WF08187>
- Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2), 137–152. <https://doi.org/10.1002/wics.1346>
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed). New York, NY: Springer. Retrieved from <https://hastie.su.domains/ElemStatLearn/>
- Hua, W., Lou, S., Huang, X., Xue, L., Ding, K., Wang, Z., & Ding, A. (2024). Diagnosing uncertainties in global biomass burning emission inventories and their impact on modeled air pollutants. *Atmospheric Chemistry and Physics*, 24(11), 6787–6807. <https://doi.org/10.5194/acp-24-6787-2024>
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: a model comparison approach to regression, ANOVA, and beyond* (Third Edition). New York: Routledge, Taylor & Francis Group.
- Jury, M. R., & Gaviria Pabón, A. R. (2021). Dispersion of Smoke Plumes over South America. *Earth Interactions*, 25(1), 1–14. <https://doi.org/10.1175/EI-D-20-0004.1>
- Li, F., Levis, S., & Ward, D. S. (2013). Quantifying the role of fire in the Earth system – Part 1: Improved global fire modeling in the Community Earth System Model (CESM1). *Biogeosciences*, 10(4), 2293–2314. <https://doi.org/10.5194/bg-10-2293-2013>
- Li, F., Song, X., Harrison, S. P., Marlon, J. R., Lin, Z., Leung, L. R., et al. (2024, May 15). Evaluation of global fire simulations in CMIP6 Earth system models. <https://doi.org/10.5194/gmd-2024-85>
- Liu, T., Mickley, L. J., Marlier, M. E., DeFries, R. S., Khan, M. F., Latif, M. T., & Karambelas, A. (2020). Diagnosing spatial biases and uncertainties in global fire emissions inventories: Indonesia as regional case study. *Remote Sensing of Environment*, 237, 111557. <https://doi.org/10.1016/j.rse.2019.111557>
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>

Naus, S., Domingues, L. G., Krol, M., Lujikx, I. T., Gatti, L. V., Miller, J. B., et al. (2022). Sixteen years of MOPITT satellite data strongly constrain Amazon CO fire emissions.
<https://doi.org/10.5194/egusphere-2022-450>

Teckentrup, L., Harrison, S. P., Hantson, S., Heil, A., Melton, J. R., Forrest, M., et al. (2019). Response of simulated burned area to historical changes in environmental and anthropogenic factors: a comparison of seven fire models. *Biogeosciences*, 16(19), 3883–3910.
<https://doi.org/10.5194/bg-16-3883-2019>