Overview

This paper compares fire emission parameters derived from twelve different fire emission models forecasting a 2019 wildfire event in the United States. Parameters are compared among models and observed data. By doing so, the authors aim to derive meaningful insights into the current progress of fire emission models and hence suggest ways to further improve the efficacy of such models. The parameters compared in this paper include biomass burning organic carbon emissions, smoke AOD (magnitude and spatial coverage), surface PM$_{2.5}$, plume rise height and ratio of smoke AOD and PM$_{2.5}$; hence covering both physics and chemistry aspect of fire emission modeling. The paper suggests areas which current fire emission models can improve on, which includes methodologies to represent diurnal evolution of fire emissions, improved vertical distribution of emitted pollutants and better representation of plume injection heights.

The models and methodology used are clearly described and the paper is well written. I have only two suggestion and a few minor suggestions/clarifications to make.

Major Suggestions

Line 877: It would be highly insightful to understand how the type of emission injection method (within PBL, intermediate, deep) affect the model skill in predicting plume rise heights. Certain emission injection method may be more useful for a certain kind of fire plume (fresh, aged, fire characteristics: smoldering, raging fire, etc.) and not others. If we can associate a better emission injection method with a corresponding type of fire plume, we can improve fire modelling skill. Indeed, further investigations in this regard is necessary and will definitely be a good follow up work.

Line 981: I may not be proficient enough in this aspect, so this is just some thoughts. There might be inherent problems using surface smoke PM$_{2.5}$ to smoke AOD ratio when you have different sAOD filters for different models. For models with small sAOD (denominator), the ratio will tend to be bigger and hence result in larger spread. This is consequentially seen in the large nominal mean bias. For example, ARQI and NAQFC have 0.01 sAOD threshold and consequentially have a very large NMB. HRRR smoke and WISC WRF-Chem have 0.02 sAOD threshold and also have very large NMBs. CAMS have a larger threshold, 0.05, and consequentially have smaller magnitude NMBs. AIRPACT is the
exception here. This may affect both the magnitude and spread of the ratio calculated and may lead to unfair comparisons between the models. This may affect the model evaluation.

**Minor Suggestions**

Line 36: For $2.5$ in $\text{PM}_{2.5}$, suggest to be written in subscript.

Line 85: There is one multi-model comparison done by Li, et. al., 2019. Atmos. Chem. Phys., 19, 12545–12567 for many different fire models, which may be worthy to look at.

Line 157: If the forecast system produces more than 1 cycle per day, how is the data treated? Is the data averaged?

Line 220: The style of writing for Section 2.1.6 seems to be slightly different from the rest of the paper. Consider revising.

Line 378: I would like to clarify if the models were in a spun-up condition when model forecasts were extracted to compare with observed data.

Line 446: AERONET is already defined in line 435.

Line 534: It may be insightful to suggest a reason why FRP-driven models results in higher sAOD compared to hotspot driven models.

Line 655: It may be problematic to compare point derived ground-based measurement station data against model grid predictions of smoke $\text{PM}_{2.5}$. Perhaps a small discussion about this issue will be helpful.

Line 660 and a few other places: May want to revise the use of ‘it’s’.

Line 730: Consider revising this sentence.

Figure 5: Spelling of AERONET in the figure and caption.