General responses from the authors

The authors thank both referees for their time and effort in reviewing this manuscript. Their suggestions were very helpful in improving the paper.

We address individual referee's comments below, in blue font, following the original comments indicated by italics.

Please note that all table and figure numbers referenced in our responses are based on those in the *original* manuscript and Supplementary Material (SM). However, we have made a number of revisions to both documents (attached). Aside from textual changes, we also moved Figure S5 from the SM to the end of Section 3.1.5 of the main text as Figure 11, and we have added four new tables, Tables S2, S4, S5, and S10, to the SM.

Anonymous Referee #2:

- *The paper presents a number of updates to the Canadian operational biomass burning (BB) emissions model and its verification for North America (NA) for 2017 fire season. Several improvements have been made to the model to improve the parameterizations of the BB emissions, fire plume rise and behavior. These updates have resulted in improvement of the simulated O3 and PM2.5 concentrations.*

  The development of new capabilities for smoke forecasting is very important. As recent years showed the wildfires in the US and Canada can cause severe air pollution episodes affecting millions of people. Accurate and timely air quality forecasting plays a critical role for stakeholders and public to mitigate the effects of the adverse air pollution from wildfires.

  The paper is well organized. This study deserves to be published in GMD.

Thank you for your positive comments.

- *My major comment is that while the verification of the ground level PM2.5, O3 and NO2 are important, it is uncertain how accurately the model simulates concentrations of the chemical species aloft. Smoke aerosols in the atmosphere affect radiation, thus affecting weather and climate. The authors demonstrate that the new plume rise algorithm injects fire emissions at higher altitudes compared to the previous version of the model. This change leads to the reduction of the high bias in the ground level PM2.5 concentrations forecast by the older model. To verify the PM2.5 concentration simulations within entire atmospheric column, it would be helpful to compare the model predicted AOD fields. Figure 10 illustrates the model's ability in capturing the wide smoke plume over NA. However, this is a qualitative comparison. I realize that a full quantitative verification of the model versus the satellite AOD is beyond scope of the paper. Therefore, I suggest comparing the model with satellite measured AOD over NA for 1-2 episodes at least, so a reader can get an idea how realistic is the forecast total aerosol burden from fires.*

We agree that a more systematic assessment of model evaluation is desirable, especially for species concentrations aloft where changes to the plume-injection-height parameterization will have the largest impacts. However, the main goal of the operational FireWork system is to provide numerical guidance on air quality conditions to regional forecasters and emergency first-responders, for whom pollution episode arrival time and surface-level concentrations are the most important forecast quantities. Surface-level model performance metrics are thus the focus of this work.

Additional research is currently underway to evaluate the model's upper-air performance and plume-injection height parameterization using a research version of the GEM-MACH model with a 12-bin aerosol size representation. Model results are being analyzed against satellite plume-height retrievals from the MISR sensor and aircraft measurements of a wildfire from a measurement campaign that took place in Alberta in July 2018. This work is alluded to in Section 4 (2nd, 5th, and 6th paragraphs). In addition, AOD estimates made from the 12-bin GEM-MACH model are better suited for these analyses than AOD estimates from the 2-bin GEM-MACH model used in the operational forecast evaluation experiments presented here, where operational bin 1 covers the diameter size range from 0 to 2.5 µm.

Given that the focus of the paper is on the first-step "operational evaluation" of the new AQ forecast system where surface level results are critical, and that the paper is already lengthy, we would prefer to save a more detailed, upper-air analysis for a next, follow-up study. As you noted, though, we have provided a subjective VCD comparison with satellite imagery as a proxy subjective evaluation for total column $PM_{2.5}$ in Sec. 3.1.5, as this imagery is part of the forecast product suite from the operational FireWork system. In addition, we have moved Figure S5, which shows a comparison of both satellite imagery and surface measurements at individual stations with predictions from the two FireWork versions, from the SM to the end of this section (new Figure 11).

- *Page 18. If you discuss these SI figures here, then move to the main text.*

Thank you for the suggestion. We have chosen not to implement this change because we were concerned that the additional tables and figures for the 2017 O_3 and NO_2 evaluations (Tables S3-S6 and Figures S1-S4) might distract from the $PM_{2.5}$ model forecast results, which we think are more important given wildfire $PM_{2.5}$ impacts on human health and visibility. Furthermore, our analysis showed that most significant changes are in $PM_{2.5}$ predictions as compared to O_3 and NO_2 predictions.

- *Table 7. Are these daily concentrations? Specify.*

Thank you for pointing this out. Table 7 shows categorical scores (POD, FAR, CIS) for $PM_{2.5}$/O_3/NO_2 over regions of interest for the 3 months in 2017. These scores are calculated based on hourly modelled and measurement values paired by grid location and time. These categorical scores were defined in Section 3, and we have modified the description in the paper at the beginning of Section 3 to specify clearly that hourly values were used in these calculations:

> "… Model hourly results for the near-surface concentrations at measurement site locations were extracted, paired by time, and evaluated using common model evaluation statistics as well as three operational, forecast-oriented categorical scores (Jolliffe and Stephenson, 2012). The categorical scores**,** calculated from hourly values**,** were probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI), where:…"

We have also modified the Table 7 caption to indicate that hourly values were used.

- *I suggest merging section 4 and 5.*

Sections 4 and 5 cover different topics and are naturally separate. In the current layout, Section 4 highlights outstanding development issues not currently considered in FireWork-CFFEPS system, and emphasis future evaluations that may better attribute changes of forecast results to model developments. Section 5 provides an overall summary and conclusions of the work and can be useful for readers who want a quick synopsis of the study.

As the paper is already lengthy, with many figures and tables, we believe that keeping these sections separate improves the overall readability of the paper.