This manuscript presents a comparison of predictions from the CMAQ model driven by meteorological fields derived from the FV3-GFSv16 (or GFS) and WRF models for periods coinciding the 2019 FIREX-AQ field campaign. The GFS-CMAQ system comprising of CMAQv5.3.1 and the FV3-GFSv16 now constitute the operational National Air Quality Forecast Capability. Consequently, a study assessing the performance of the system relative to both predictions from another modeling system (in this case the widely used WRF-CMAQ configuration) as well as multi-species measurements from a field campaign would be of interest for establishing the efficacy of the recently updated NAQFC system. While the manuscript presents a significant amount of work in terms of simulations with the two modeling systems and comparisons with airborne measurements taken during the FIREX-AQ campaign, and for most part shows similar performance for the GFS-CMAQ and WRF-CMAQ models, the execution of the analysis in the current form in my assessment did not present in a compelling manner either the impacts of the different meteorological drivers on the predicted air quality or the ability of the NAQFC improvements in capturing the effects of fire emissions on air quality. Some of this perhaps results from the study design trying to cover too much ground. For instance, uncertainties in estimating fire emissions (magnitude, composition, placement (vertically and burn area)) confound isolating meteorological differences from those resulting from emission uncertainties. Comparisons with high time (and space) resolution measurements are confounded by limitation in model grid to reasonably isolate meteorological difference impacts. Similarly, use of data assimilation (strong nudging) also likely reduces the differences between the GFS and WRF fields making it difficult to identify impacts associated with either physics (PBL, cloud representation), or computational differences (Interpolation vs. native grid). Perhaps a clearer articulation of the study objectives and closer linkage of the analysis with the objectives would help improve the overall usefulness of the manuscript. Also, many of the inferences conveyed by the analysis could benefit from additional substantiation. The following suggestions are offered that may help address these
shortcomings:

- I find the use of the terminology WRF-CMAQ downscaling (in the title and manuscript text) to be somewhat confusing in context of what is presented. In my view “downscaling” is methodology typically used to translate coarse grid information to finer resolutions with a dynamical model but without any other observational assimilation. Why not just refer to it as WRF-CMAQ configuration?
- It would be useful to explain in a little more detail the differences in data assimilation approaches used in the FV3-GFS and WRF simulations. Since this is a retrospective analysis, it does appear from Table 1 that WRF did utilize data nudging. How often were WRF runs initialized? Does the 4D-IAU utilized in GFS also include the observations that were incorporated in the WRF nudging? How often were the GFS runs initialized? Some discussion of the impacts assimilation (nudging) approaches in the two models on meteorological fields and any differences would be useful.
- Page 4, L11-12: Since the NACC uses bilinear interpolation, why is the interpolation of the scalar variables more straightforward and relative to what?
- Page 5, L4-10: Why were the WRF model top and vertical grid structures chosen to be so different from those of the GFS? Does that not also introduce another source of differences in the predicted meteorological fields and confound interpretation of differences arising due to different model structures vs. interpolation to a coarse vertical layer structure? What is the vertical structure of the CMAQ model deployed in the NAQFC?
- P5, L23: what does a “novel” inline dust model imply? As written, it is not readily apparent to the reader what novelty the FENGSHA approach provides for dust emissions relative to other approaches. Please expand briefly to better convey the novel aspects of the system, especially those that pertain to forecasting.
- P5, L31: What does “total organic aerosol” imply especially in context of emissions? Did the authors imply primary organic aerosols?
- P5, L36-40: The description of scaling of wildfire emissions could be improved. Why are only fires west of 110W assumed to be long lasting? How is the model grid cell fraction (am assuming of burned area) determined?
- Section 3.1: Did the authors compute performance statistics for meteorological fields from both the native GFSv16 as well as the fields interpolated to the CMAQ grid? Were these comparable? It may be useful to state so.
- P7, L8: It was not apparent to me what “produce synoptic scale body forces” implies. Please clarify.
- P7, L13-20: I found it difficult to quantitatively infer the PBL differences from the color scales depicted in Figure 2. It may be useful to also include histograms of the differences. Are the monthly means based on all hourly data or only daytime values? Are monthly mean differences of the order of 500-1000m not too large? Would these not result in more pronounced differences in predicted ambient concentrations?
- Figure 3: Significant differences in WRF wind fields in the West and Northeast are shown for the period after 8/21. Why do these arise if winds were nudged to observations? The discussion on page 8 suggests that “some WRF setting” leads to large bias for storm weather, but it is not clear what these settings are and what causes the error? It would be useful to expand the discussion on page 8.
- P8: should “northwest” on L28 be “west”? L30: perhaps “shallower PBL” is better description than “thinner PBL”
- P8, L32-33: At night, is the stable surface layer not decoupled from the residual layer,
so that there is not much exchange anyway?

- P9: demonstrating the impacts of systematic differences in PBL heights and wind fields between GFS and WRF on the eventual PM2.5 distributions is very useful. However, I am not sure I fully understand the discussion. The GFS nighttime PBL heights appear to be lower (Fig 2b), but so do the PM2.5 (Fig 2c). Would both shallower PBL and weaker winds not result in higher PM2.5 within the stable boundary layer in GFS-CMAQ? While it is possible that lack of transport, results in lower downwind PM2.5 concentrations, would the lower PBL depicted in Fig2 not result in higher PM2.5 at least in the source regions at night?

- Would the impact of these large differences on predicted air quality perhaps be better captured through comparisons with more widespread surface observations (of ozone and PM2.5 and constituents) than just limited flight segments, especially given the uncertainties in the estimated fire emissions? Hourly surface measurements (say from the AQS network) may also help better delineate the impacts of any systematic diurnal differences in WRF and GFS predictions on the eventual atmospheric composition predictions.

- Figure 7: How well did the native WRF and GFS vertical wind velocities compare with the measured ones? Also, it appears that the re-diagnosed W in GFS-CMAQ show greater variability than those in WRF-CMAQ, especially at higher altitudes – would the authors know what may cause that? If it is due to interpolation, does it suggest an issue with the interpolation scheme employed for the GFS fields? Some additional discussion would be useful in context of the variability in the observed W fields.

- P11, L16-17: The sentence should be reworded a bit – the relationship between underestimation of CO and NOz is not apparent to me.

- P12, Figure 10 discussion. The model CO values do not appear to deviate too much from their “baseline”. Is this because of underestimation of fire emissions or possibly due to the model fire plume being displaced relative to the flight path?

- P12, L17-19: does “background” represent regional average values or the amount of ozone from outside the domain?

- P12, L21-23: I am curious why modeled NO2 as opposed to CO is used as an indicator of fire location – would CO not be a better tracer? How were the column averaged NO2 and ozone observations (filled circles) shown in Fig S3 computed?

- P12, L31: flight path missing the locations of the modeled peak values is a somewhat awkward way to suggest that the model and observed plumes were misplaced. Consider rewording this sentence.

- The axis labels in the sub-figures of Fig 11 are extremely small and illegible.

- Tables 2 and 3: Please explain how the data was segregated into fire and no-fire bins used for these statistics.

- P14, L19-21: Please explain why accounting for warm bias during non-fire impacts would be influenced by the lack of using wildfire heat content in the energy balance equations of the meteorological models. Would the warm bias not get exacerbated?

- P14, L38-39: I was surprised that the SOx bias was attributed to the SO2 point source inventory? Do the emission estimates used in these simulations not use the continuous emissions monitoring data for point sources which provide accurate constraints for SO2 emissions from point sources?

- Tables 2 and 3: I was curious why the sum of the observed NOx+NOz is not equal to the observed Noy values shown in these tables? It may be due to differences in the averaging but would be useful to verify and explain.

- P15, L1-4: While PAN is underestimated, HNO3 is not – do the PAN underestimates fully explain the NOz underestimation? What other species constitute NOz in the analysis presented?

- P15, L11: would be useful to provide a reference for the “our other analysis”.

- P15, L14: it is not clear what VOC “speciation issue” in NEIC2016v1 is implied here - please elaborate.

- P15, L18: what size range does the “submicron ammonium” refer to and how were the corresponding model values estimated?
- P15, L22: I was curious what chemical pathway in the model converts HNO3 to organic nitrates? Does the gas-phase chemical mechanisms employed include such conversions?
- P16, L2-5: One way to isolate the effects of resolution vs. emission errors could be to examine ratios of model and observed species associated with fire emissions, since the ratios will not be impacted by systematic artificial dilution effects of grid resolution.
- P16, L6-7: Given that there is not much difference between observed O3 values of 58 and 56 ppb, I am not sure one can discern systematic titration effects conveyed in this discussion. Similarly, the O3 enhancements suggested in downstream areas (L11) are not apparent in the comparisons presented. Some additional discussion and substantiation of what is being conveyed would be useful.
- P16, paragraph starting L30: The discussion contains many sentences that are somewhat vague and should be rewritten. As examples: (i) why only refer to advection as horizontal? Similarly, does the vertical transport imply only cloud mixing? (ii) What is the issue mentioned by Qian (2020)? The physical significance of the issue should be mentioned. (iii) How does one quantitatively ascertain that the differences between the GFS and WRF physics have larger impact on the meteorological fields than the “meteorology driver” methodology, since no interpolation was used in recasting the WRF fields for CMAQ? Some elaboration would be useful to support the statement.
- P17, L37-40: Given that the FV3-GFS and CMAQ grid structures are different, what option other than interpolation is available? I am not sure I see what specific advantage of the NACC is being conveyed here. It is also not clear what is implied by “it avoids running another model as WRF” since the choice of the meteorological driver appears to be user driven. It seems the NACC was created for a specific purpose of translating the GFS meteorological fields and interfacing with CMAQ and as such does not preclude use of other models such as WRF but just provides an option. It is also not clear what is meant by the statement that the NACC is “faster and more consistent with the original meteorological driver”? Relative to what? What does consistency imply in this context? Also, as indicated earlier, I feel that the “downscaler” terminology for WRF unnecessarily introduces additional confusion.
- P18, L12-13: “enough variables to drive CMAQ with other supplied data” is awkward and should be reworded.
- P18, L15-25: I found this discussion to be somewhat wishy-washy with no clear conclusion or indication of next steps.