The referee comments are shown in blue.

The responses to the comments are shown in black.

The changes included in the revised manuscript are shown in red.

Reply to Comments from Reviewer #1

Reviewer #1 comments:

This study uses Global Forecast System with the new Finite Volume Cube-Sphere dynamic core (GFS-FV3) to drive the CMAQ v5.0.2 and evaluates the model results with observational data. The forecast system shows good agreement in meteorological variables and pollutants. This manuscript fits the scope of the journal of Geoscientific Model Development. However, more detailed and in-depth descriptions are expected in several places (see below comments).

Response:

Thank you for your positive comments. Please see below our point-by-point responses.

Specific comments

1. Line27-28: As mentioned here, NAM model is the current meteorological model. Can authors explain more in-depth why chose the GFS-FV3? Please also present some comparison results between these two models.

Response: The necessity and urgency to develop a FV3GFS-CMAQ system is owing to the retirement of the North American Mesoscale forecasting system (NAM) in the NOAA National Weather Service (NWS). The NAM model is no longer being updated (as of 2017). The current regional models for providing high-resolution forecast and guidance in NWS will be eventually replaced by a FV3-based system by 2023. The

FV3GFS-CMAQ system studied in this work is essentially a replacement of the NAQFC's offline coupled to a meteorological driver system by swapping NAM by FV3GFS.

The meteorological performance from the two systems, NAM and FV3GFS, is comparable (see, e.g., https://hmt.noaa.gov/experiments/pdf/WPC-HMT_WWE_2019_Final_Report.pdf). Huang et al. (2019) and Lee et al. (2020, https://www.cmascenter.org/conference/2020/session-presentations10.cfm) provided an air-quality-model-specific performance evaluation when a CMAQ-based Chemical Transport Model was driven by the NAM and FV3GFS meteorology. Although we did not compare performance of NAM versus FV3GFS, the FV3GFS-CMAQ interim NAQFC-β system we analyzed in this paper showed an across-board improvement in terms of the major chemical evaluation performance statistics than that by the NAM driven operational NAQFC.

To address the reviewer's comments, we have added a brief summary on the comparison of NAM and FV3 based on previous publications and reports, see below: the GFS v15 gave lower T2 prediction over CONUS comparing to NAM. While the NAM had slightly biased high T2, the GFS v15 had slightly biased low T2 during Aug 2019. The T2 underprediction was more significant in the Midwest by GFS v15 than its by NAM, especially during daytime. During the 2018-19 winter season, the FV3GFS had similar statistical scores regarding performance in snowfall prediction. While both FV3GFS and NAM gave over-forecasting in accumulated snow under most of precipitation type methods, the FV3GFS had larger over-prediction than NAM. It indicated FV3GFS had a colder forecast. The authors attributed the larger overprediction in snow depth to the consistent cold bias in GFS v15, which was identified by NWS through the intercomparison between GFS v14 and v15.

2. Line 50: "semi- or intermediate-VOCs" is mentioned as the missing sources of $PM_{2.5}$ in abstract. However, none of the S/I VOCs sources are analyzed in the rest of the manuscript. Please double check whether the analysis of this part is omitted.

Response: We have added some discussions on the impact of missing S/IVOCs and related SOA chemistry, see our response #2 to the reviewer #2's comments and also see below the revised text in the manuscript:

In CMAQ v5.0.2, the primary organic aerosol (POA) is processed as non-volatile. The emissions of semivolatile and intermediate volatility organic compounds (S/IVOCs) and their contributions to the secondary organic aerosol (SOA) are not accounted for in the aerosol module. In the recent versions of CMAQ, two approaches linked to POA sources have been implemented. One introduces semi-volatile partitioning and gas-phase oxidation of POA emissions. The other (called pcSOA) accounts for multiple missing sources of anthropogenic SOA formation, including potential missing oxidation pathways and emissions of IVOCs. These two improvements lead to increased organic carbon concentration in summer but decreased level in winter. The changes vary by season as a result of differences in volatility (as dictated by temperature and boundary layer height) and reaction rate between winter and summer. Therefore, the missing S/IVOCs and related SOA chemistry in v5.0.2 are key reasons for the OC overprediction and underprediction during cooler and warmer months, respectively.

3. Line 173-174: What is the purpose to only extracting the first 24-hour results from each 72-hour forecast? If the first 24-hour results are only needed, why still simulate the next 48 hours?

Response: We mainly focus on the first 24-h forecast for the following 3 reasons:

(1) The experimental GFSv15-CMAQv5.0.2 system is a prototype and still being developed. It is not qualified for operational application yet. Thus, the forecast results for day-2 (forecast for 25-48h) and day-3 (49-72h) were occasionally unavailable due to running issues, system fails, or archive missing, especially during the early stage in the application of this forecast system.

(2) We did the discrete statistics for day-2 performance to compare with the day-1 performance. Since there were a couple days lacking day-2 and day-3 forecast results in Jan, Feb, and Jun, the statistics are not presented for those months in case of keeping the apple-to-apple comparison. We found the day-2 performance are close to the day-1 performance. The difference in MDA8 O_3 predictions are shown in Table R1. The difference in monthly NMBs are up to

3%, mostly within 1%. Similar day-2 $PM_{2.5}$ predictions to day-1 $PM_{2.5}$ prediction could be also found in Table R2.

(3) Many of the studies for previous NAQFC forecasting performance focused on the day-1 performance (e.g., Kang et al., 2010; Lee et al., 2017). The day-1 results presented in the manuscript would be more comparable with other studies.

Table R1. Performance statistics of MDA8 O_3 against AIRNow dataset

Day1 performance MDA8 O_3, ppb								Day2 performance MDA8 O_3, ppb							
Period	Mean Obs.	Mean Sim.	MB	RMSE	NMB,%	NME,%	Corr	Period	Mean Obs.	Mean Sim.	MB	RMSE	NMB,%	NME,%	Corr
Jan	32.1	32.0	-0.1	7.2	-0.4	17.2	0.58	Jan	/		/		/		/
Feb	36.4	35.5	-0.9	7.8	-2.5	16.7	0.58	Feb	/		/		/		/
Mar	44.9	40.4	-4.5	8.7	-10.0	15.8	0.56	Mar	44.9	40.3	-4.6	8.9	-10.2	16.1	0.53
Apr	46.4	43.1	-3.3	7.7	-7.1	13.3	0.62	Apr	46.4	42.9	-3.5	8.1	-7.5	13.8	0.59
May	44.1	42.7	-1.4	7.8	-3.3	13.9	0.67	May	44.1	42.2	-1.9	8.3	-4.4	14.8	0.62
Jun	45.7	43.9	-1.8	10.9	-4.0	18.3	0.59	Jun	/		/		/		/
Jul	44.3	46.6	2.3	9.5	5.2	16.6	0.72	Jul	44.3	46.2	1.9	9.8	4.4	17.1	0.69
Aug	43.7	46.9	3.2	9.4	7.3	16.4	0.74	Aug	43.7	46.6	2.9	9.7	6.7	16.8	0.71
Sept	42.5	45.6	3.1	8.0	7.2	14.4	0.79	Sept	42.5	45.1	2.6	8.1	6.1	14.6	0.77
Oct	37.0	40.4	3.4	7.8	9.3	15.8	0.80	Oct	36.8	40.1	3.3	7.9	9.1	16.0	0.77
Nov	34.2	35.9	1.8	7.6	5.2	16.5	0.72	Nov	34.1	35.1	1.0	7.7	3.0	16.4	0.69
Dec	31.7	33.5	1.8	7.8	5.6	18.6	0.68	Dec	29.8	30.6	0.8	8.0	2.8	20.3	0.60

Table R2. Performance statistics of 24-h avg $PM_{2.5}$ against AIRNow dataset

Day1 performance 24-h avg $PM_{2.5}$, μg m^{-3}								Day2 performance 24-h avg $PM_{2.5}$, μg m^{-3}							
Period	Mean Obs.	Mean Sim.	MB	RMSE	NMB,%	NME,%	Corr	Period	Mean Obs.	Mean Sim.	MB	RMSE	NMB,%	NME,%	Corr
Jan	8.2	13.8	5.5	11.5	66.9	92.3	0.35	Jan	/		/		/		/
Feb	7.9	12.5	4.6	10.0	58.0	81.5	0.53	Feb	/		/		/		/
Mar	7.8	11.0	3.2	9.2	41.2	69.0	0.40	Mar	7.8	11.0	3.2	10.4	41.2	71.1	0.36
Apr	6.3	8.0	1.7	6.3	27.9	61.6	0.33	Apr	6.3	7.5	1.3	5.5	20.1	58.6	0.33
May	6.7	6.9	0.2	4.7	3.3	49.3	0.26	May	6.7	6.5	-0.2	4.6	-2.7	49.0	0.27
Jun	7.1	6.8	-0.3	5.4	-4.2	47.1	0.22	Jun	/		/		/		/
Jul	8.4	8.5	0.1	11.8	1.0	59.8	0.28	Jul	8.4	8.0	-0.4	10.5	-4.7	56.1	0.27
Aug	7.2	6.9	-0.3	4.0	-4.7	40.2	0.33	Aug	7.2	6.8	-0.4	4.1	-5.4	41.0	0.34
Sept	7.0	7.6	0.6	4.7	8.5	44.2	0.48	Sept	7.0	7.0	0.0	4.3	-0.1	43.2	0.51
Oct	6.6	9.6	3.0	9.0	44.7	73.2	0.36	Oct	6.6	8.9	2.2	7.5	33.4	67.4	0.36
Nov	8.9	13.2	4.2	9.8	47.2	72.1	0.48	Nov	8.9	12.8	3.9	9.7	43.3	70.7	0.47
Dec	8.8	13.9	5.1	10.8	57.9	82.5	0.51	Dec	8.8	13.6	4.8	10.9	54.5	82.1	0.49

4. Line 192-193: What are the criteria or references for setting this threshold (120 ppb

and 100 μg m^{-3} for O_3 and $PM_{2.5}$)? How about those abnormal low data?

Response: There are many abnormal records in the raw AIRNow data. We calculate the record numbers above certain thresholds for O_3 and $PM_{2.5}$. For O_3, records above 120, 160, 200, and 300 ppb are 0.31%, 0.17%, 0.08%, and 0.06% of the total records. For $PM_{2.5}$, records above 100, 200, 300, and 500 μg m^{-3} are 0.26%, 0.24%, 0.21%, and 0.20% of the total records. we chose thresholds of 120 ppb for O_3 and 100 μg m^{-3} for $PM_{2.5}$ as they are much higher than most observed values and provide a reasonable representation of outliers, although their selection is more or less arbitrary. We did not exclude abnormally low values.

To address reviewer's comments, we utilize the Quality Assurance/Quality Control information from the AIRNow dataset to filter the invalid records in the revised manuscript. The arbitrary thresholds are no longer used. We redo the statistics and the updated results are shown in Tables R3 and R4 below. As shown, the changes in performance statistics between the two filtering methods are minor, with slightly better results by excluding those outliners and abnormal records. Our major conclusions remain. We update the figures and the relevant sections accordingly in the revised manuscript.

Table R3. Performance statistics of MDA8 O_3 against AIRNow dataset

	MDA8 O_3, ppb in GMDD submission								MDA8 O_3, ppb with updated QC						
Period	Mean Obs.	Mean Sim.	MB	RMSE	NMB,%	NME,%	Corr	Period	Mean Obs.	Mean Sim.	MB	RMSE	NMB,%	NME,%	Corr
Jan	32.4	32.0	-0.3	7.9	-1.1	18.0	0.52	Jan	32.1	32.0	-0.1	7.2	-0.4	17.2	0.58
Feb	36.7	35.7	-1.1	8.4	-2.9	17.4	0.53	Feb	36.4	35.5	-0.9	7.8	-2.5	16.7	0.58
Mar	45.1	40.4	-4.7	8.9	-10.4	16.0	0.55	Mar	44.9	40.4	-4.5	8.7	-10.0	15.8	0.56
Apr	46.6	43.1	-3.5	8.0	-7.5	13.5	0.61	Apr	46.4	43.1	-3.3	7.7	-7.1	13.3	0.62
May	44.3	42.7	-1.6	7.9	-3.7	14.0	0.66	May	44.1	42.7	-1.4	7.8	-3.3	13.9	0.67
Jun	45.9	43.9	-2.0	11.2	-4.4	18.5	0.58	Jun	45.7	43.9	-1.8	10.9	-4.0	18.3	0.59
Jul	44.5	46.6	2.1	9.7	4.7	16.7	0.70	Jul	44.3	46.6	2.3	9.5	5.2	16.6	0.72
Aug	43.9	46.9	3.0	9.5	6.8	16.3	0.73	Aug	43.7	46.9	3.2	9.4	7.3	16.4	0.74
Sept	42.7	45.6	2.9	8.1	6.8	14.5	0.78	Sept	42.5	45.6	3.1	8.0	7.2	14.4	0.79
Oct	37.2	40.2	3.1	8.0	8.3	15.8	0.77	Oct	37.0	40.4	3.4	7.8	9.3	15.8	0.80
Nov	34.3	34.8	0.5	8.4	1.6	16.9	0.64	Nov	34.2	35.9	1.8	7.6	5.2	16.5	0.72
Dec	30.7	31.2	0.5	9.0	1.6	20.5	0.49	Dec	31.7	33.5	1.8	7.8	5.6	18.6	0.68
O_3-season	44.3	45.1	0.9	9.4	2.0	16.0	0.67	O_3-season	44.1	45.1	1.0	9.2	2.5	16.0	0.69
Non O_3-season	38.2	37.4	-0.9	8.4	-2.3	16.4	0.68	Non O_3-season	37.7	37.5	-0.2	7.8	-0.4	16.0	0.72
Annual	41.1	41.0	-0.1	8.9	-0.1	16.2	0.70	Annual	40.5	40.9	0.4	8.5	1.0	16.0	0.73

Table R4. Performance statistics of 24-h avg $PM_{2.5}$ against AIRNow dataset

24-h avg $PM_{2.5}$, µg m^{-3} in GMDD submission								24-h avg $PM_{2.5}$, µg m^{-3} with updated QC							
Period	Mean Obs.	Mean Sim.	MB	RMSE	NMB,%	NME,%	Corr	Period	Mean Obs.	Mean Sim.	MB	RMSE	NMB,%	NME,%	Corr
Jan	8.3	13.8	5.5	11.4	66.4	92.4	0.34	Jan	8.2	13.8	5.5	11.5	66.9	92.3	0.35
Feb	8.0	12.5	4.5	10.0	55.9	81.0	0.51	Feb	7.9	12.5	4.6	10.0	58.0	81.5	0.53
Mar	7.9	11.0	3.1	9.4	39.6	68.9	0.38	Mar	7.8	11.0	3.2	9.2	41.2	69.0	0.40
Apr	6.3	8.0	1.7	6.6	26.5	62.0	0.30	Apr	6.3	8.0	1.7	6.3	27.9	61.6	0.33
May	6.8	6.9	0.2	5.0	2.3	49.8	0.23	May	6.7	6.9	0.2	4.7	3.3	49.3	0.26
Jun	7.2	6.8	-0.4	5.6	-5.1	47.4	0.20	Jun	7.1	6.8	-0.3	5.4	-4.2	47.1	0.22
Jul	8.3	8.5	0.1	11.7	1.7	59.9	0.30	Jul	8.4	8.5	0.1	11.8	1.0	59.8	0.28
Aug	7.3	6.9	-0.4	4.1	-5.2	40.4	0.33	Aug	7.2	6.9	-0.3	4.0	-4.7	40.2	0.33
Sept	7.0	7.6	0.5	4.7	7.6	44.4	0.48	Sept	7.0	7.6	0.6	4.7	8.5	44.2	0.48
Oct	6.7	9.5	2.8	8.6	41.7	71.9	0.35	Oct	6.6	9.6	3.0	9.0	44.7	73.2	0.36
Nov	9.0	13.2	4.2	9.8	46.7	72.0	0.48	Nov	8.9	13.2	4.2	9.8	47.2	72.1	0.48
Dec	8.8	13.8	5.0	11.0	56.6	82.9	0.49	Dec	8.8	13.9	5.1	10.8	57.9	82.5	0.51
DJF	8.4	13.4	5.0	10.8	59.7	85.6	0.45	DJF	8.3	13.4	5.1	10.8	61.0	85.5	0.46
MAM	7.0	8.6	1.6	7.2	23.5	60.6	0.33	MAM	6.9	8.6	1.7	7.0	24.8	60.4	0.36
JJA	7.6	7.4	-0.2	7.9	-2.6	49.7	0.26	JJA	7.6	7.4	-0.2	7.8	-2.5	49.5	0.27
SON	7.5	10.0	2.5	8.0	33.0	63.4	0.45	SON	7.5	10.1	2.6	8.1	34.4	63.8	0.46
Annual	7.6	9.8	2.2	8.6	29.0	65.3	0.40	Annual	7.6	9.9	2.3	8.5	30.0	65.2	0.41

5. Line 259-263: What is this “artificial temporal allocation algorithm”? Please introduce more details about this algorithm.

Response: The “artificial temporal allocation algorithm” means the calculation in preprocessing from accumulated precipitation, which is recorded originally in GFS v15 outputs, to hourly precipitation, which will be used by CMAQ model. To address reviewer’s comments, we provide a more detailed description in the revised manuscript as follows:

The precipitation from the original FV3 outputs are recorded as 6-h accumulated precipitations. Artificial errors were introduced to the forecast by an issue in precipitation preprocessing during the early stage development of the GFSv15-CMAQv5.0.2 system. The precipitation at first hour of the 6-h cycle would be dropped occasionally. We corrected this issue and the hourly precipitation still shows large underprediction against surface monitoring networks. It indicates the difficulty for the forecast system in capturing the temporal precipitation, especially during summer (Figure S4). During the summer season, the discrepancy in capturing the short-term heavy rainfall worsens the model performance in predicting hourly precipitation. Besides, we use the threshold of 0.1 mm hr^{-1} to filter the valid records. If the model predicts precipitation that did not occur, the record will be excluded into the statistics calculation. However, all the predicted precipitation is counted in the spatial evaluation against the ensemble datasets of GPCP and CCPA. Therefore, the spatial performance of monthly accumulated precipitation shows better agreement than its of hourly statistics.

In general, the relatively poor performance of the forecast system in capturing the precipitation at the same hours with the observation is the major cause for the large underprediction in hourly statistics.

6. Line 335: What is “Higher predicted $PM_{2.5}$, typically SOA, in California is expected in the future using GFS-FV3-CMAQv5.3.” means? Does it mean that GFS-FV3-CMAQv5.3 would predict higher concentrations than GFSv15-CMAQv5.0.2 for $PM_{2.5}$? If so, what leads to these higher concentrations in GFS-FV3-CMAQv5.3? An updated mechanism or some updated PM sources? Which one is more important for the $PM_{2.5}$ prediction?

Response: As discussed in the comment #2, the GFS-CMAQv5.3 system is expected to give higher predicted SOA in California during summer compared to the current GFSv15-CMAQv5.0.2 system. The primary PM emissions generally decrease from previous NEI to the more recent NEI. Some previews and intercomparison could be seen at http://views.cira.colostate.edu/wiki/wiki/10202/inventory-collaborative-2016v1-emissions-modeling-platform. However, the updated chemical mechanism also includes enhanced SOA formation from anthropogenic and biogenic sources, and is one of the key factors in improving the underestimation of organic aerosol in CA during summer 2010. Therefore, the updated mechanism would be more important for the $PM_{2.5}$ underprediction in those areas.

7. Line 347-359: It’s better to move the method introduction to the section 2.

Response: The method for categorical evaluation has been moved to section 2 now.

8. Line 383-385: As mentioned above, GFS-FV3-CMAQv5.3 will have higher $PM_{2.5}$ concentrations. Since the significant overprediction of $PM_{2.5}$ leads the poor performance in capturing the category of “Unhealthy for Sensitive Groups” in cooler months mentioned here, whether the updated system GFS-FV3-CMAQv5.3 would have worse prediction? Can authors provide any suggestion to avoid this?

Response: The combined effect of semivolatile POA and pcSOA tends to decrease organic aerosol in winter. In addition to the semivolatile POA and pcSOA mentioned above, monoterpene SOA was also updated in CMAQv5.3. The impact of updated monoterpene SOA chemistry is more significant during summer because the BVOC emissions are much more reactive in summer than other months in southeastern US (Pye et al., 2018, 2019). Therefore, the POA and SOA updates in v5.3 are likely to lead to improvements at all times of year. The revised discussion is added in the main text. Please refer to the text in red in the response #2 in the reviewer #2 comments.

9. Section 3.3: What is the difference of the meteorological prediction among regions? Please introduce it. It would be helpful to explain the pollutant prediction bias in different region.

Response: The regional performance of meteorological prediction and its relationship with the chemical prediction are added in the revised text:

We further quantify the meteorology-chemistry relationships by conducting the region-specific evaluation of the meteorological variables. The regional performance for the major variables is shown in Figure S9. The regional biases in T2 predictions show high consistency with the regional biases in MDA8 O_3. It indicates that the cold biases in the Midwest (including region 5) and the warm biases near the Gulf coast (including regions of 4 and 6) are important factors for the O_3 underprediction and overprediction in those regions, respectively. The ozone temperature relationship was found (S. Sillman and Samson, 1995; Sillman, 1999). O_3 is expected to increase with increasing temperature within specific range of temperature (Bloomer et al., 2009; Shen et al., 2016). The surface MDA8 O_3-temperature relationship was found at approximately 3-6 ppb K^{-1} in the eastern US (Rasmussen et al., 2012). According to such relationships, the biases in T2 predictions could explain large portion of the O_3 biases. Heavy convective precipitation and tropical cyclones occur more often in the southeastern US, where are mainly regions of 4 and 6. Therefore, the performance in precipitation predictions are lower in those two regions comparing to other regions as we have shown the model has relatively poor performance in capturing short-term heavy rains during summer seasons in section 3.1. Meanwhile, the performance in wind predictions in regions 4 and 6 is relatively poor. Such performance in the meteorological predictions is consistent with the mixed performance in $PM_{2.5}$ prediction in regions 4 and 6. The discrepancy in meteorological inputs, mainly in precipitations and wind, can be attributes to the low temporal agreement shown as correlations of predicted $PM_{2.5}$ in those two regions.

Technical corrections

1. Line 1: “GFSv15-FV3-CMAQv5.0.2” should be “GFSv15-CMAQv5.0.2” to be consistent with the expression in other part of the manuscript.

Response: The FV3 dynamical core was firstly implemented in the operational GFS starting at v15. To include the complete information of the model versions in the manuscript title per the requirement of submission on Geoscientific Model Development, we incorporate the abbreviation of “GFSv15-FV3-CMAQv5.0.2” for the air quality forecasting system: GFS v15 with FV3 dynamical core offline coupling with CMAQ v5.0.2.

2. Line 214:215: the term “ozone season” should be rewrite as “O_3-season” to be consistent with the expression in other part of the manuscript.

Response: The sentence is reworded as “O_3-season”.

3. Line 419: the term “overpredicted” should be “underpredicted”.

Response: The sentence is corrected.

4. Line 539: “nemsio” should be “NEMSIO” to be consistent with the expression in other part of the manuscript.

Response: The “nemsio” is reworded as “NEMSIO”.

5. Figure 2, Figure 8b and 8d: Some labels and lines are overlap. Please modify these pictures and make it clearer.

Response: The labels in these figures are adjusted to be shown more clearly.

6. Figure 8: The serial number of the figure ((a), (b), (c), (d)) should be in front of the title.

Response: The serial numbers are adjusted to be in front of the title.

7. Figure 8a, 8c: The term “CONUS” should be “Overall”.

Response: The labels for the term “CONUS” in these two figures are reworded as

“Overall”.

8. Figure 9b, 9d: Part of the labels of the figure is missing. Please modify it.

Response: Figure 9 is adjusted to show the labels completely.

Reply to Comments from Reviewer #2

Reviewer #2 comments:

This contribution details a region-specific, time-specific, and categorical evaluation of the meteorological and chemical forecasts from the offline-coupled GFSv15-CMAQv5.0.2 for the year 2019. This manuscript fits the scope of the journal of Geoscientific Model Development. However, the manuscript has important limitations. Although the paper is well organised and detailed, the results shown in the paper are, in my opinion, not sufficiently original and new to merit publication in this specific journal.

Response:

Thank you for your constructive comments. Please see below point-by-point responses.

Major comments

1. My main objection focuses on the fact that it is not clearly stated which elements of the study are genuinely new and original and which echo the findings of previous studies. I cannot clearly state either new methodology developed or any new results, apart from applying state-of-the-art validation methods to a new set of simulations with the GFSv15-CMAQv5.0.2.

Response:

Compared to previous publications in the literature, there are several new and unique aspects of our work.

Firstly, the historical and current NAQFC are based on the NAM-CMAQ system. However, the NAM has been no longer updated since March 2017, as NOAA Environmental Modeling Center (EMC) has transitioned to devote its full resources towards the development of an ensemble model based on the FV3 dynamical core. The FV3-based system will eventually replace all current NOAA National Centers for Environmental Prediction (NCEP) mesoscale models used for forecasting. The next generation NAQFC will be based on the FV3. The NOAA National Weather Service (NWS) is currently coordinating an effort to inline a regional scale meteorological model basing on the same dynamic core as that in FV3GFS to be coupled with an atmospheric chemistry model partially based on CMAQ. The implementation of this inline system NAQFC is expected a few years in the future. Therefore, the system in this study is an important interim operational air quality forecasting capability the nation will have before the future inline system matures to be operational. The forecast skill of the interim FV3GFS-CMAQ system we analyzed in this work could also provide as a benchmark for the future inline system.

Secondly, while the FV3-based GFS v15 showed similar and even better performance compared to the previous GFS v14 (https://www.emc.ncep.noaa.gov/users/meg/fv3gfs/) and the NAM (https://hmt.noaa.gov/experiments/pdf/WPC-HMT_WWE_2019_Final_Report.pdf), air quality forecasting driven by FV3-based meteorology has not been fully evaluated. This is the first evaluation paper for the FV3GFS-CMAQ forecasting system, and demonstrates the ability of FV3-based GFSv15 to drive the air quality forecast.

Thirdly, an updated study is needed because previous journal publications for NAQFC were based on older versions of CMAQ (v4.6 or v4.7) and historical 2005 and 2011 NEIs (Kang et al., 2010a, 2010b; Garner et al., 2015; Bray et al., 2017; Huang et al., 2017; Lee et al., 2017; Pan et al., 2020). In our work, we implement 2014 NEI emission, GBBEPx fire emission, initial and lateral boundary conditions, and a CMAQ version closer to those in the current operational NAQFC thus allowing for the results and findings to be more directly transferable to the current NAQFC and inform further development of the future FV3-based NAQFC.

To further address the reviewer comments, we reorganize our manuscript in the revision. A section is added for detailed discussion of the processes and causes attributing to the biases. The analysis summarized in the evaluation section of the previous manuscript are incorporated and expanded for detailed analysis in the discussion section. Here we describe the major items added in the discussion section. We focus more on the impact of meteorology-chemistry relationship and the chemical processes on the model biases. Firstly, the meteorology-chemistry relationship is further quantified by adding the region-specific analysis of meteorological biases. Secondly, the discussion for major biases in O_3 and $PM_{2.5}$ prediction is enhanced.

For O_3: The underpredicted O_3 during non-O_3 season was neglected in our previous manuscript. We include the temperature-O_3 relationship to quantify the impact of cold biases on the O_3 underprediction in the regions of 1, 3, 5, and 10, and on the overprediction in regions of 4 and 6 near the Gulf coast. While the cold biases could explain certain portion of the MDA8 O_3 underprediction, the dry deposition algorithm of O_3 to snow cover in CMAQ v5.0.2 is discussed for further contribution to the underprediction in the northern regions during winter and early spring. In addition to the temperature-O_3 relationships, the significant O_3 overprediction near the Gulf coast would be associated with the missing halogen chemistry and the overestimated emissions from oil-gas sources.

For $PM_{2.5}$: By this point, more and more observation datasets have been available. For example, most of the daily $PM_{2.5}$ composition observations from Air Quality System (AQS) dataset are available for the entire 2019 now. We include them into further analysis and discussion for the manuscript. The contributions of the OC, sulfate, and dust compositions to the $PM_{2.5}$ biases are quantified. Our findings are better supported by the additional evidence. These additional datasets and analysis help us to dig deeper into understanding the $PM_{2.5}$ biases and their causes instead of referring previous studies to support our analysis in the original manuscript.

Our study supports the future development of the FV3GFS-CMAQ system and even the operational NAQFC, which has similar model settings and inputs with the system evaluated in this study. For example, by introducing the evaluation of $PM_{2.5}$ compositions, we found the overestimation in dust compositions is one of the major sources for the $PM_{2.5}$ overprediction in cooler months. We further conduct a sensitivity simulation by implementing an adjustment for suppressing the fugitive dust by snow cover, which was implemented in the operational NAQFC of the NAM-CMAQv5.0.2 system in 2020. We found that the suppression and adjustment could improve the model performance during cooler months. It further indicates the need of further development and improvement for the NAQFC.

2. Furthermore, objective (3) states that one of the aims of the manuscript is to "investigate underlying causes for the biases to provide a scientific basis for improving the model representations of chemical processes and developing science-based bias correction methods for O_3 and $PM_{2.5}$ forecasts.". However, after reading the manuscript, I don’t find any specific section of the manuscript devoted to understanding the physicschemical processes causing over- or underproduction of air quality parameters (apart from some general comments). The authors should deepen into the processes leading the levels of air pollution so that objective (3) can really be achieved.

Response:

A discussion section is added for deepening the analysis of the underlying causes in three aspects: (1) Meteorology-chemistry relationships, (2) Causes for major biases in O_3 predictions, and (3) Causes for major biases in $PM_{2.5}$ predictions. The revised text is listed below:

We further quantify the meteorology-chemistry relationships by conducting the region-specific evaluation of the meteorological variables. The regional performance for major variables is shown in Figure S9. The regional biases in T2 predictions show high correlation with the regional biases in MDA8 O_3. It indicates that the cold biases in the Midwest (including region 5) and the warm biases near the Gulf coast (including regions of 4 and 6) are important factors for the O_3 underprediction and overprediction in those regions, respectively. The O_3-temperature relationship was found (S. Sillman and Samson, 1995; Sillman, 1999). O_3 is expected to increase with increasing temperature within specific range of temperature (Bloomer et al., 2009; Shen et al., 2016). The surface MDA8 O_3-temperature relationship was found at approximately 3-6 ppb K^{-1} in the eastern US (Rasmussen et al., 2012). According to such relationships, the biases in T2 predictions could explain large portion of the O_3 biases. Heavy convective precipitation and tropical cyclones occur more often in the southeastern US, which covers mainly regions 4 and 6. Therefore, the performance in precipitation predictions is worse in those two regions comparing to other regions as we have shown the model has relatively poor performance in capturing short-term heavy rains during summer seasons in section 3.1. Meanwhile, the performance in wind predictions in regions 4 and 6 is relatively poor. Such performance in the meteorological predictions is consistent with the mixed performance in $PM_{2.5}$ prediction in regions 4 and 6. The discrepancy between simulated and observed meteorological variables, mainly in precipitations and wind, can be attributed to the poor temporal agreement shown as correlations of predicted $PM_{2.5}$ in those two regions.

In the original submission, we focus on the significant O_3 overprediction near the Gulf coast during O_3-season. In the revised manuscript, we analyze likely causes for additional model biases, e.g., underpredictions in the Northeast, Mid-Atlantic, Midwest, Mountainous states, and the Northwest (mainly corresponding to the regions 1, 3, 5, 8, and 9) during non-O_3 season as follows:

The O_3 concentration is underforecasted for the Northeast, Mid-Atlantic, Midwest, Mountainous states, and the Northwest (mainly corresponding to the regions 1, 3, 5, 8, and 9) during non-O_3 season. Large difference in dry deposition algorithms between CMAQ v5.0.2 and other common parameterizations was reported (Park et al., 2014; Wu et al., 2018). Large discrepancy between modeled dry deposition velocity of O_3 by CMAQ v5.0.2 and the observation during winter was shown and attributed to the deposition to snow surface. Improvement was indicated in revising the treatment of deposition to snow, vegetation, and bare ground in CMAQ v5.0.2. Lower deposition to snow was found to improve the consistency between the O_3 deposition modeled by CMAQ v5.0.2 and the observations. Therefore, the dry deposition module in v5.0.2 needs to be updated and improved for more accurate representation of low-moderate O_3 mixing ratios (Appel et al., 2020). For the cases in this study, the snow cover for the months of Jan and Apr in winter and spring is shown in Figure 7a and 7b. The underpredicted O_3 during non-O_3 season may be caused by the overestimated O_3 deposition to snow in the northern regions, corresponding to the previous regions 1, 3, 5, 8, and 9. The mixed effects of the temperature-O_3 relationship discussed above and the large deposition to snow contribute to the moderate O_3 underpredictions.

In the original manuscript, we conducted part of the analyses for the $PM_{2.5}$ biases and the causes based on referring to some previous studies and several speculations. By introducing the additional AQS dataset into our evaluation, we can gain further insights into the specific reasons for the $PM_{2.5}$ biases:

The variation in predicted $PM_{2.5}$ composition between cooler and warmer months indicates that major seasonal biases are caused by multiple factors. We introduce the AQS dataset for evaluation of daily $PM_{2.5}$ composition to provide additional insights into the specific reasons. Figure 9 shows the biases of the key $PM_{2.5}$ composition for the cooler month of Jan and warmer month of Jul. While the overall mean biases of $PM_{2.5}$ composition, including elemental carbon (EC), ammonium (NH_4^+), and nitrate (NO_3^-) are within ±0.5 μg m^{-3} for all months of the year, the major biases in $PM_{2.5}$ predictions are mostly contributed by organic carbon (OC), soil components (SOIL), and sulfate (SO_4^{2-}). The soil components are estimated using the Interagency Monitoring of Protected Visual Environments (IMPROVE) equation and specific constituents (Appel et al., 2013). During a cooler month, the significant overprediction in $PM_{2.5}$ is mainly attributed to the overprediction in OC and SOIL.

During warmer months, the overprediction of SOIL and sulfate compensate for the overall underprediction in OC in v5.0.2, leading to the moderate $PM_{2.5}$ underprediction in the Southeast but slight overprediction in the Midwest, Mid-Atlantic, and the Northeast. These high $PM_{2.5}$ SOIL concentrations are consistent in spatial characteristics with large emissions of anthropogenic primary $PM_{2.5}$, and primary coarse PM in the Midwest, Northeast, and the Northwest (Fig. S6). The underprediction in $PM_{2.5}$ OC during summer compensate the overestimation in dust during cooler months, resulting in the overall biases with an annual NMB of 30%.

The large emissions of anthropogenic primary coarse PM, as well as the wind-blown dust are the major sources for predicted $PM_{2.5}$ SOIL components. Appel et al. (2013) indicated CMAQ overpredicted soil components in the eastern United States partially due to the anthropogenic fugitive dust and wind-blown dust emissions. The overprediction in $PM_{2.5}$ soil compositions by our forecast system could be mainly attributed to the overestimation of the anthropogenic fugitive dust emission because the meteorological conditions were not included in processing the anthropogenic fugitive dust sector. The dust-related components of aluminum, calcium, iron, titanium, silicon, and coarse mode particles are overestimated in the regions with snow and precipitation, especially during winter, early spring, and late autumn with snow cover in the north, which contributes to the $PM_{2.5}$ overprediction, with more significant temporal-spatial pattern in the northern U.S. during cooler months.

In CMAQ v5.0.2, the primary organic aerosol (POA) is processed as non-volatile. The emissions of semivolatile and intermediate volatility organic compounds (S/IVOCs) and their contributions to the secondary organic aerosol (SOA) are not accounted for in the aerosol module. In the recent versions of CMAQ, two approaches linked to POA sources have been implemented. One introduces semi-volatile partitioning and gas-phase oxidation of POA emissions. The other (called pcSOA) accounts for multiple missing sources of anthropogenic SOA formation, including potential missing oxidation pathways and emissions of IVOCs. These two improvements lead to increased organic carbon concentrations in summer but decreased level in winter. The changes vary by season as a result of differences in volatility (as dictated by temperature and boundary layer height) and reaction rate between winter and summer. Therefore, the missing S/IVOCs and related SOA chemistry in v5.0.2 are key reasons for the OC overprediction and underprediction during cooler and warmer months, respectively.

The detailed analysis for the underlying cause by the uncertainty and biases in emissions of anthropogenic fugitive dust will be discussed further in the discussion section and in the response to the “other comment #2 by Reviewer 2”.

In general, we reorganize the evaluation and the discussion sections. The discussion of relationships between the meteorological factors and the chemical performance are further enhanced. The physical drivers are linked with the biases in air pollutant predictions more deeply. The discussion of the model chemistry is expanded in more detail.

3. Despite the large number of statistical figures presented, I have the very personal opinion that the authors do not take the advantage of the compiled information to point to the specific causes for model biases.

Response:

We agree that some of the presented figures and materials were not well interpreted to informative messages to the readers. For example, the evaluation of monthly accumulated precipitation is shown in three figures (Figures S2 to S4 in the original submission). But the main purpose of those three figures in the original manuscript was only to indicate that the GFSv15-CMAQv5.0.2 system has better agreement in the spatial characteristics than the temporal variations. We revise the presentation of the evaluation of accumulated precipitation as Figure S3. The performance and the spatial patterns of the monthly accumulated precipitation during four seasons are more straightforward. Meanwhile, the findings from the evaluation section 3 are further discussed in section 4 in details. We closely link the meteorological drivers to the analysis for chemical biases as we indicated in the response of comment #2. The cold biases shown in Figure S2 are further discussed in section 4.1. Additional information and analysis are added in section 4.2 to address the major biases in O_3 predictions shown in Figures 2 and 6. Furthermore, the information and description for Figure 7 in the original manuscript are revised and enhanced in section 4.3 by introducing the evaluation of $PM_{2.5}$ against AQS dataset. The original Figure 9 is further supported by the additional analysis of seasonal variation of $PM_{2.5}$ composition and the diurnal emissions as indicated in the response to major comment #2 and other comment #2.

In general, we reorganize some figures to make them more concise. We deepen the analysis by providing additional supporting material for several speculations in the revised discussion section.

Other comments

1. The authors should compare the skills of the model (categorical evaluation) with other published model studies, in order to have a flavor of the behavior of the model when compared to other forecasting systems worldwide.

Response:

Major RT-AQF systems over the world were comprehensively reviewed in (Zhang et al., 2012a, 2012b). Here we include a comparison with the more recent air quality forecasting studies from Canada (Moran et al., 2018; Russell et al., 2019), Europe (Struzewska et al., 2016; D'Allura et al., 2018; Podrascanin, 2019; Stortini et al., 2020), East Asia (Lyu et al., 2017; Zhou et al., 2017; Peng et al., 2018; Ha et al., 2020), and CONUS (Kang et al., 2010; Zhang et al., 2016; Lee et al., 2017). We summarize the performance in these studies in a table in the supplementary material. As for the categorical performance, the air quality standards vary in different regions (Oliveri Conti et al., 2017). For example, National Ambient Air Quality Standards (NAAQSs), the Ambient Air Quality and Cleaner Air for Europe (CAFE) Directive (2008/50/EC), and the national ambient air quality standard (GB 3095-2012) are set up by U.S., Europe, and China, respectively. Therefore, the definition of the categorical metrics may vary between regions even with the same metric name. Their categorical performance are discussed specifically in the revised text:

Table S3 summarizes air quality forecasting skills reported in the literature along with that from this work. For those studies with data assimilation in air quality forecasting, the performance from the raw results without data assimilation are presented. The performance in predicting O_3 and PM varies largely between model systems. The discrete and categorical performance in O_3 prediction is not significantly better than that in PM prediction. O_3 tends to be slightly overpredicted in an annual base or for the warmer months. The annual NMB and Corr for O_3 over the North America are 1.4% and 0.76 for 2010 in Moran et al. (2018), while they are 1.0% and 0.73 in this study. However, the performance in $PM_{2.5}$ prediction varies largely from our study. The $PM_{2.5}$ for warmer months were moderately overpredicted in Russel et al.

(2019), with the MBs ranging from 3.2 to 5.5 μg m^{-3}. The categorical performance of GFSv15-CMAQv5.0.2 in predicting MDA8 O_3 is similar with that of the previous NAQFC (Kang et al., 2010), in which the FAR and H are ~68 % and ~31% for "Unhealthy for Sensitive Groups", and the H is ~47% for "Moderate" category, respectively. The H for $PM_{2.5}$ also decreased largely from ~46% for "Moderate" to ~21% for "Unhealthy for Sensitive Groups" category, and the FAR was over 90% for the "Unhealthy for Sensitive Groups" category in Kang et al. (2010). The overpredicted $PM_{2.5}$ was also found when using the historical 2005 NEI in forecast for Jan 2015 (Lee et al., 2017). The performance was improved by updates of 2011 NEI and real-time dust and wildfire emissions. It indicates the needs of improving our emission inventory. As for the categorical performance in regions other than CONUS, the air quality standards vary (Oliveri Conti et al., 2017). For example, National Ambient Air Quality Standards (NAAQSs), the Ambient Air Quality and Cleaner Air for Europe (CAFE) Directive (2008/50/EC), and the national ambient air quality standard (GB 3095-2012) are set up by U.S., Europe, and China, respectively. Metrics also vary between studies. The primary forecasting products are O_3 and PM_{10} from some forecasting systems instead of O_3 and $PM_{2.5}$ in this study. The threshold for categorical evaluation of O_3 used in D'Allura et al (2018) was 83.0 μg m^{-3}. The applied metrics of the False Alarm Ratio and Probability of Detection (POD) were defined the same as the FAR and H used in our study. The FAR and POD were 36.14% and 71.16%, respectively. The categorical evaluation of $PM_{2.5}$ in Ha et al. (2020) was applied for four categories: (1) 0-15 μg m^{-3}, (2) 16-50 μg m^{-3}, (3) 51-100 μg m^{-3}, and (4) >100 μg m^{-3}. The overall FAR and Detection Rate for four categories are 59.0% and 36.1%, respectively. Although the metrics of FAR and Detection Rate were defined for four categories, rather than every single category as for this study, the categorical performance is comparable with our results. In general, the discrete and categorical performance of O_3 forecast in this study is comparable that of the air quality forecasting systems in many regions of the world. However, the PM forecasts vary largely between studies. While our GFSv15-CMAQv5.0.2 system shows consistent performance with the systems covering CONUS, the high FAR and low H for "Unhealthy for Sensitive Groups" category with higher thresholds indicate that the categorical performance could be further improved by addressing the significant overprediction during cooler months in this study.

2. Emissions are really important in forecasting system; however, this manuscript lacks information about the emissions used (time series, spatial patterns, seasonal behavior, etc). The authors should explain in a higher degree of detail how emissions are considered and implemented in the forecasting system.

Response:

We expand the introduction of how we prepare and implement the emission into the GFSv15-CMAQv5.0.2 system in the methodology section 2:

The anthropogenic emissions from area, mobile, and point sources in National Emissions Inventory of year 2014 version 2 (NEI 2014v2) are processed by the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system. The onroad mobile sources include all emissions from motor vehicles that operate on roadways such as passenger cars, motorcycles, minivans, sport-utility vehicles, light-duty trucks, heavy-duty trucks, and buses. Onroad mobile source emissions are processed using emission factors output from the Motor Vehicle Emissions Simulator (MOVES). SMOKE uses a combination of vehicle activity data, emission factors from MOVES, meteorology data, and temporal allocation information to estimate hourly, gridded onroad emissions. The nonroad, agriculture, anthropogenic fugitive dust, non-elevated oil-gas, residential wood combustion, and other sectors are included in the area sources. The sectors of airports, commercial marine vessel (CMV), electric generating units (pt_egu), point sources related to oil and gas production (pt_oilgas), point sources that are not EGUs nor related to oil and gas (ptnonipm), and point sources outside US (pt_other) are included in the point sources. The sulfur dioxide (SO_2) and nitrogen oxide (NO_X) from point sources in NEI 2005 are projected to year 2019 following the methods used in Tang et al. (2015, 2017). The biomass burning emission inventory from the Blended Global Biomass Burning Emissions Product system (GBBEPx) (Zhang et al., 2019b) is impletemented for the forecast of forest fires. The GBBEPx fire emission is treated as one type of point source. Its heat flux is derived from satellite retrieved fire radiative power (FRP) to drive fire plume rise. The GBBEPx is a near real time fire dataset. The fire emission implemented in the current forecast cycle comes from the historical fire observation, typically 1-2 day behind. In this system, we use landuse information to classify fires into forest fire and other burning such as agriculture burning. We assume only forest fire can last longer than 24 hours. We assume the forest fire emission will continue on day 2 and beyond. Other types of fires will be dropped as we assume few of them could continue beyond day 2. The plume rise of the point source is driven by the meteorology and allocated to the 35 elevated layers in GFSv15-CMAQv5.0.2 system by the PREMAQ preprocessing system. Biogenic emissions are calculated inline by Biogenic Emission Inventory System (BEIS) version 3.14 (Schwede et al., 2005). Sea-salt emission is parameterized within CMAQ v5.0.2. While the deposition velocities are calculated inline, the fertilizer ammonia bi-directional flux for in-line emissions and deposition velocities is turned off.

The impact of the emissions on the biases in O_3 prediction is added in the revised text:

In addition to the impact of meteorological biases and missing halogen chemistry on the O_3 overprediction near Gulf coast, the overestimated VOC emission could increase O_3 biases. The anthropogenic VOCs emissions continuously decrease from historical NEIs to 2016 NEI (http://views.cira.colostate.edu/wiki/wiki/10202/inventory-collaborative-2016v1-emissions-modeling-platform). We compare the VOCs emissions between 2016 NEI and the emissions used in this study. Figure S10 shows the difference in the elevated source of pt_oilgas. The Gulf coast is impacted by the oil and gas sector due to the oil and gas fields, and the exploration activity near it. By comparing the 2016 NEI to the current emissions we used in the system, we found that the overestimation of the VOCs emissions could be one aspect to the O_3 overprediction near the Gulf Coast. We only project the SO_2 and NO_x from 2005 NEI to 2019 and we do not project the VOCs for the elevated sources. The monthly VOCs emissions from pt_oilgas sector for July in regions 4 and 6 are 2876.0 tons month^{-1}, while they are 2497.0 tons month^{-1} in 2016 NEI. The reduction mainly locates along the coastline, where the significant overprediction takes place. It indicates the complicated effect of meteorological biases, missing gas-phase chemistry, and the overestimation of emissions on the O_3 prediction in such area.

While the diurnal characteristics of the emissions with the revised Figure S11 are added in the revised manuscript to understand the diurnal $PM_{2.5}$ biases, additional analyses are conducted for specific issues below:

During cooler months, the significantly overpredicted $PM_{2.5}$ is mainly attributed to the emission of anthropogenic fugitive dust. In reality, the meteorological conditions could largely impact the amount and characteristics of anthropogenic fugitive dust. For example, the snow cover and the soil moisture are important factors in calculating the dust emissions in SMOKE. However, the anthropogenic fugitive dust implemented in this GFSv15-CMAQv5.0.2 system was not adjusted by the precipitation and snow cover. The large emissions of anthropogenic primary coarse PM, as well as the wind-blown dust are the major sources for predicted $PM_{2.5}$ SOIL components. Appel et al. (2013) indicated CMAQ overpredicted soil components in the eastern United States partially due to the anthropogenic fugitive dust and wind-blown dust emissions. The overprediction in $PM_{2.5}$ soil compositions by our forecast system could be mainly attributed to the overestimation of the anthropogenic fugitive dust emission because the meteorological conditions were not included in processing the anthropogenic fugitive dust sector. The dust-related components of aluminum, calcium, iron, titanium, silicon, and coarse mode particles are overestimated in the regions with snow and precipitation, especially during winter, early spring, and late autumn with snow cover in the north. Thus, it contributes to the $PM_{2.5}$ overprediction, with more significantly temporal-spatial pattern in the north U.S. during cooler months.

An adjustment of precipitation and snow cover for fugitive dust was implemented in the operational NAQFC. The dust-related PM emissions will be clean up using a factor of 0.01 when the snow cover is higher than 25% or the hourly precipitation is higher than 0.1 mm hr^{-1} before they are used as input for CMAQ v5.0.2 forecast. We conduct a sensitivity simulation for Jan 2019 using the GFSv15-CMAQv5.0.2 system with the adjustment implemented in the operational NAQFC. Figure 7c shows the $PM_{2.5}$ overprediction in the northern regions 1, 2, 5, and 10 during Jan is largely improved corresponding to the spatial-temporal characteristics of snow cover. The monthly MB and NMB for Jan improves from 5.5 μg m^{-3} and 66.9% to 2.1 μg m^{-3} and 24.0%, respectively. The improvement is mainly attributed to the decrease in overpredictions in $PM_{2.5}$ soil components, with MBs decreased from 3.3 μg m^{-3} to 1.2 μg m^{-3} for Jan (Fig. 7d). The overprediction in the Northeast and Northwest during spring is expected to be improved by the suppression of the fugitive dust by the snow during early spring. This indicates the importance of including the meteorological forecast in processing the emission of anthropogenic fugitive dust. It should be calculated inline or be adjusted by the meteorological forecast.