Responses to Reviewer Comment #1:

We want to thank the reviewer for his/her detailed and thoughtful comments. Our point-to-point responses to these comments are provided below, with proper changes made in the revised manuscript.

1. At many places in the manuscript it is sated that a high-resolution forecast system has been developed for the LIS region. Is this an operational forecast system? Was the system operational for the entire LISTOS study? What relation does it have to the NAQFC system which is also often mentioned? Were the base and additional simulations conducted in forecast or retrospective mode? Did the WRF simulation employ data assimilation – if so, was the assimilation strategy as one would use it in forecast mode? What was the extent of the WRF modeling domain? Was WRF deployed in a nested mode – is so what was the extent of the outer domain and how did it compare with the meteorology used for the NAQFC which provided the chemical LBC for the high-resolution domain?

Response: The NAQFC is the operational air quality forecasting system operated by the National Oceanic and Atmospheric Administration (NOAA), while the LIS system presented in this study is a new research air quality prediction system under development. It is not operational. We have added this information to avoid confusion (see Lines 95 & 146 in the revised manuscript).

The LIS study has no direct relationship with NAQFC, although it may provide some useful information for NAQFC to develop a future high resolution forecasting system with emission and chemical data assimilation capabilities. NAQFC is mentioned here because it is the national operational forecasting system, hence being used as a benchmark to evaluate the model performance of the LIS system. To help understand the difference between NAQFC and LIS, we added a table to the Supplemental Information (SI) as shown below:

Table 1. Comparisons between LIS and NAQFC configurations during the LISLOS study.

Configuration	LIS	NAQFC
Horizontal resolution	3km	12km
Meteorology	WRF with Global Forecasting System (GFS) acting as ICs/BCs	North America Mesoscale Forecast System (NAM)
Lateral Boundary conditions	Various (Default/NAQFC)	climatological gaseous LBC from GEOS-chem. Dynamic aerosol LBC from GEFS-Aerosol
Initial concentration	CMAQ restart file	Previous run
Base Emission	NEI2011v2	NEI2014v2
Chemistry	CB6 Aero6	CB05 Aero6

Both the base run and adjustment simulations were conducted in forecast mode. The O_3 episodes during the LISTOS field campaign were chosen here because of the rich pool of measurements (aircrafts, lidar, etc.) that can be used for detailed model evaluation.

Regarding data assimilation for meteorology, it was not applied in the WRF model used in this study, as we primarily focused on improving the performance of the CMAQ model. The WRF model was not deployed in nested mode and its domain was one grid larger on each boundary compared to that of the CMAQ model. The WRF model run, however, was driven with initial and boundary conditions from the NOAA NCEP's Global Forecast System (GFS) simulation, which was equipped with various data assimilation techniques. We have clarified this issue in the manuscript and added the description of the configuration in Lines 151-153:

"No data assimilation was applied in the WRF simulation. The model is conducted in a single domain with 132×122 grid cells with one grid more on each boundary compared to that of the chemical transport model."

The meteorology data from NAQFC and LIS could not be compared to each other since NAQFC, in its public release, only provides CMAQ output files, not the input files, such as meteorology and emissions. Hence, such a comparison is not feasible and beyond the scope of this study. We acknowledge that many factors listed in the Table above can contribute to the difference in model performance, including meteorology inputs. As NAQFC is driven by operational weather forecasts, it is not surprising that the WRF simulation may not be as great in the NOAA operational weather forecast. Nevertheless, the new system performs better even with the current WRF input, highlighting the importance of other factors such as model resolution, emissions and chemical data assimilation, in improving prediction performance.

2. If the high-resolution forecast system has been deployed for an extended period, how does its general performance compare to that for the period of focus of this study, i.e., August 25-31, 2018? This aspect is specifically important to understand if there were conditions unique to the high ozone episode examined here or model attributes that led to the noted performance characteristics. On the other hand, if this is a limited modeling study that was conducted to examine this specific ozone episode it is okay to state that so the results can be viewed and assessed in the right context.

Response: Thank you. This is a good point. As mentioned earlier, this is a research prediction system, not operated continuously. We have clarified it in the revised manuscript (L146-147):

"The high-resolution air quality forecasting system used here is a new research prediction system deployed during the 2018 LISTOS field campaign period."

In addition, we have conducted an extended model simulation beyond the LISTOS field campaign period, and presented the model evaluation with observations from the ground monitoring network AQS. During the cold season, the simulated O_3 showed similar concentration level with the observations (see examples in Fig. 1a), as well as the spatial patterns (i.e., on Jan 15, 2020, the domain-average correlation coefficient (CORR) was 0.76 and RMSE was 6.78 ppb). In the summer season, the simulation without adjustment showed acceptable correlation (CORR=0.92) but it underestimated the concentration (Fig. 1b). With the adjustment, the model bias was reduced and the predicted concentrations were brought closer to the observed levels (Fig. 1c). Its CORR in NYC could reach up to 0.99. The model shows relatively weaker performance in the springtime, such as the average CORR of 0.45 on March 14, 2020. It means this system is able to forecast the O_3 activities over this region in most time and these adjustments could improve the performance of daily run, but it still needs further improvement.

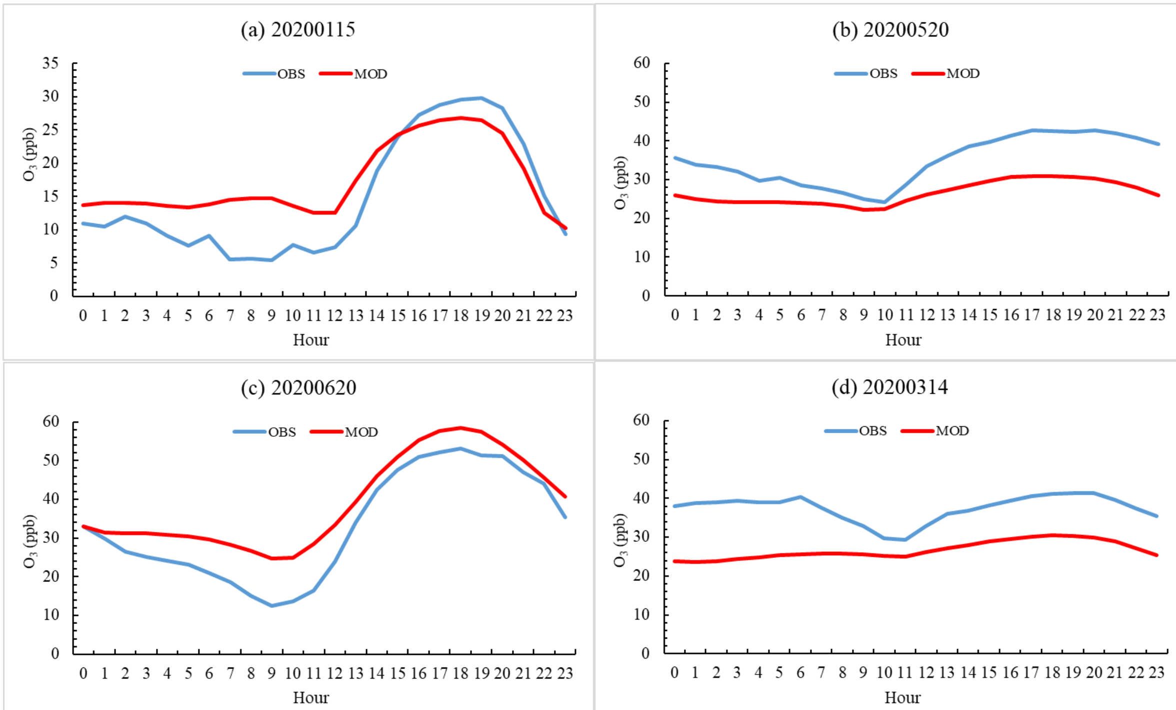

Figure 1. Diurnal variation of observed (blue line) and simulate (red line) O_3 on Jan 15 (a), May 20 (b), June 20 (c) and March 14 (d) 2020.

3. It is not terribly surprising that for the limited geographic extent of the modeling domain considered here, the specification of chemical lateral boundary conditions is found to be influential on predicted ozone (and not NO_x but I suspect also for many other species such CO, $PM_{2.5}$ etc.) distributions. Given lifetimes of these species and the typical advective time-scales this should be an expected outcome. It is not readily apparent what constitutes the default static LBC profile – what were the typical values for O_3 and other species examined? Looking at the CMAQ documentation, it appears that the default LBC profile provided represents "clean" tropospheric conditions and is recommended for use along model boundaries that are typically over remote regions devoid of significant emission forcing. From the model documentation it also appears that even those are now often substituted with conditions derived from hemispheric versions of the model. It is thus a bit surprising that for a "new" high-resolution forecast system over a high emission density region such as the northeast corridor one would consider such a profile and not a nested configuration to capture the space and time varying chemical conditions of air masses advected to the LIS. In light of this, the statement on L265 "This suggests the default profiles provided by CMAQ represent a clean environment, such as marine air layer, and are not suitable for areas with active emissions and tropospheric O_3 production" is somewhat trivial. Clearly, such impacts are well recognized as the authors do attempt to account for such by using data from the NAQFC. A clearer description of the model set up and reasons for not using a consistent one-way nested configuration (with consistent treatment of meteorology, emissions, initialization for the outer domain) to better capture the LBCs would be useful.

Response: Thank you for the suggestion. Indeed, the default static LBC profile represents clean air environment. We provide in Table 2 below the different values of O_3, NO_2, CO and sulfate (PSO4J, as an example of PM species) in the dynamic BC files from NAQFC and that from the default profile. Between the two LBCs, the O_3 concentrations increased from 37.66 ppb to 51.28

ppb, but the concentrations of NO_2 did not vary considerably during the episode. The concentrations in profile BCs are 35.0 ppb (O_3), 0.07 ppb (NO_2), 75.95 (CO) and 1.134 μg m^{-3} (PSO4J), respectively. The differences in concentration between the profile and the NAQFC dynamic BCs were higher in the summer (high O_3) season while low in cold season. It was consistent with the daily simulation results shown in Fig. 1 (see response to the comment above).

Table 1. Averaged concentrations of O_3, NO_2, CO (unit: ppb) and PSO4J (unit: μg m^{-3}) in the dynamic/static LBC files

	Date	O_3	NO_2	CO	PSO4J
	20200825	38.04	1.56	112.85	2.545
	20200826	37.66	1.42	113.49	1.944
Dynamic LBCs	20200827	44.12	1.54	133.96	3.445
	20200828	48.07	1.88	152.18	4.642
	20200829	51.28	1.65	155.05	4.721
	20200830	46.63	1.52	133.08	2.921
Static LBCs		35.0	0.07	75.95	1.134

The purpose of including the profile BCs was not to suggest using it in the final simulation, but as the control case from which we can evaluate and quantify the effectiveness of different adjustment methods in the forecasting systems. The underestimated O_3 in the control run also indicates the importance of dynamic BC during high pollution time and the influence of transported pollutants on the LIS region. In the revised manuscript, we have clarified the role of the simulation with the profile BC in the simulation design in Sect. 2.1 (L129):

"The first group (Control run) applies no adjustment, using default profile as LBCs. It serves as the reference case to allow quantifying the effectiveness of each adjustment method."

Furthermore, we removed the statement on L265 in the last version as suggested by the reviewer, and added the discussion of the influence of dynamic BCs on different season and pollutants in L304:

"As the profile BCs are static and lack spatial-temporal variations, the Control run mainly reflects the local contributions of emissions, transport and chemical processes within the domain (Tang et al., 2007). The underprediction suggests that these processes are insufficient to produce the observed O_3 levels, and that the transport of air pollutants from upwind is important to predict the high O_3 episodes. It highlights the significant influence of dynamic BCs on the simulations over this region during high pollution time. In comparison, the influence of BCs is less important during the cold season, when the simulation with the profile BCs can also result in prediction in reasonable agreement with observations (Fig. S2a, d)."

And more discussion in L340:

"In addition, the improvement of simulated NO_2 using dynamic BC was much smaller compared to that of O_3. This is because the lifetime of NO_2 is relatively short (1–7 h in summertime, Lu et al., 2015), and its budget in urban areas is mainly influenced by local emissions and chemistry, and less by regional transport, indicating the effectiveness of dynamic BCs depends not only on the downwind/upwind gradients, but also on lifetimes of the concerned species."

Previous studies on regional O_3 modeling usually used nested model to generate the BCs for the inner target domain area (such as Taghavi et al., 2004; Fu et al., 2009; Yin et al., 2015). The objective of this study is to set up a forecasting system that can be deployed fast during field campaigns. The nested models will need higher computational resources and a longer execution time. So here we used the NAQFC product as dynamic BCs directly and then tested the performance of such a system. The results show that, even without the nested configuration, the results are in good agreement with observations and we can apply this for the real-time forecast run. We added the explanation why choosing NAQFC as BCs instead of using a nested modeling system in Sect. 2.3 (L185-191):

"In the previous studies of regional modeling, a nested grid approach was often applied to provide dynamic BCs for the study area (e.g., Taghavi et al., 2004; Fu et al., 2009; Yin et al., 2015). However, the nested model would need higher computational resources and a longer running time. The increasing pool of real-time national and global forecasts provides alternative BCs that be used to drive a regional forecasting system as demonstrated in this work. Here, we explore the feasibility of utilizing the products of NOAA NAQFC, which provides real-time national forecasts to prepare dynamic boundary conditions to drive the LIS3km system. The NAQFC is an operational system, operated by the National Weather Services, and the data are provided freely to the public."

Reference:
Fu, J. S., Streets, D. G., Jang, C. J., Kebin He, J. H., He, K., Wang, L., Zhang, Q. Modeling regional/urban ozone and particulate matter in Beijing, China. Journal of the Air & Waste Management Association, 59(1), 37-44, 2009.
Taghavi, M., Cautenet, S., and Foret, G.: Simulation of ozone production in a complex circulation region using nested grids, Atmos. Chem. Phys., 4, 825–838, https://doi.org/10.5194/acp-4-825-2004, 2004.
Yin, S., Zheng, J., Lu, Q., Yuan, Z., Huang, Z., Zhong, L., Lin, H. A refined 2010-based VOC emission inventory and its improvement on modeling regional ozone in the Pearl River Delta Region, China. Science of the Total Environment, 514, 426-438, 2015.

4. Perhaps one positive consequence of using the default LBC and comparisons with the run with the NAQFC LBC is demonstrating the expected influence of regional transport on O_3 levels in the LIS region, with a suggested enhancement of 10-20 ppb in peak hourly ozone on different days due to influences from outside the modeling domain, assuming that the discrepancies relative to the observations can be solely attributed to LBC and not other model processes or input? How representative is the limited set of conditions modeled here of the high ozone days in the LIS region? There were several days during summer 2018, outside the period examined here, when the LIS region witnessed high ozone levels.

Response: We appreciated this thoughtful suggestion. Indeed, the difference between the default LBCs (background levels) and the dynamic LBCs from NAQFC represents the influence of regional transport on O_3 levels. Although the reasons that cause the model bias are complicated, the pollutant differences between Control and BCON run can be generally attributed to the pollutant transport, as the only difference between these runs is the LBC input. According to the EPA (Fig. 2), high O_3 levels were observed throughout the northeast corridor during the August 29, 2018 O_3 episode, with the highest values in the LIS region. The high O_3 concentrations in LIS were contributed by both local production and regional transport, and the transported source might contribute significantly during the peak time. We made this explanation in Sect. 3.1 (L304):

"As the profile BCs are static and lack spatial-temporal variations, the Control run mainly reflects the local contributions of emissions, transport and chemical processes within the domain (Tang et al., 2007). The underprediction suggests that these processes are insufficient to produce the observed O_3 levels, and that the transport of air pollutants from upwind is important to predict the high O_3 episodes. It highlights the significant influence of dynamic BCs on the simulations over this region during high pollution time. In comparison, the influence of BCs is less important during the cold season, when the simulation with the profile BCs can also result in prediction in reasonable agreement with observations (Fig. S2a, d). This indicates the influence of dynamic BCs varies with time and it is more significant during the high pollution time."

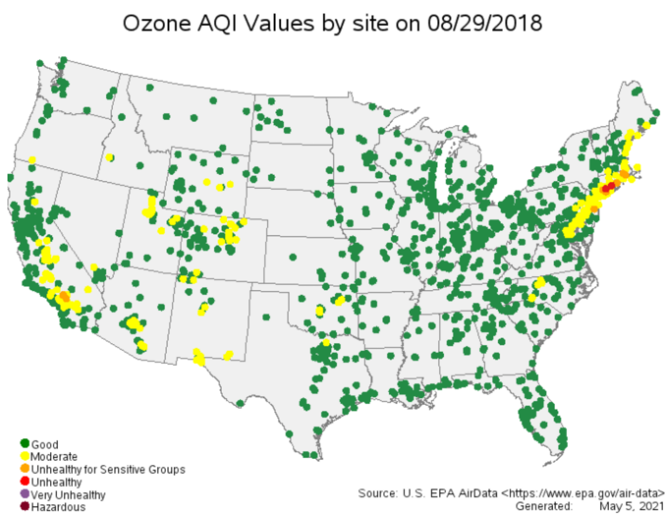

Figure 2. Ozone AQI by site on August 29, 2018 (source: https://www.epa.gov/outdoor-air-quality-data/air-data-concentration-map).

Regarding the representativeness of the simulated episodes, we have conducted additional simulations for an extended period (see the response to Comment #2). the system demonstrates a stable performance for the time outside the LISTOS period.

5. One interesting aspect depicted in Figure 2 is that the peak O_3 simulated in the control run is nearly the same on all days, suggesting that at least for the limited number of days examined in this study, emissions within the domain have about the same daily impact on average ozone. Assuming that the emissions within the domain are captured correctly and the WRF simulations represent the prevalent meteorology (neither of which are necessarily assessed), comparisons with the NAQFC and dynamic LBC case are suggestive of relatively large regional contributions on *these days – how representative are these regional contributions of high O_3 in the LIS region? This also links back to my earlier comment on the skill of the new forecast system over an extended time-period.*

Response: This is a good point. We have added, with thanks, the following statement in the revised manuscript to reflect this observation (L300):

"Note the peak O_3 simulated in the control run is nearly the same on all days during the simulation period. The comparisons between the peak O_3 with the default profile and dynamic LBC case indicates relatively large regional contributions on these days."

We have addressed the representativeness issue above.

6. The description of the optimal interpolation application would benefit from additional clarity. If I understand the methodology, surface observations are used to adjust the model initial state for O_3 and perhaps NO_x (please state that explicitly if that is the case). However, it is not readily apparent if this is done every 24 hours? At what frequency is the OI applied? Also based on the strong forcing at the surface from emissions and deposition, it is to be expected that the influence of the OI fade away rapidly. It appears that such methods may be more useful to aloft data (e.g., https://doi.org/10.1021/acs.est.8b02496). Did the authors consider using aloft observations from the O_3 lidar in conjunction with the OI to explore possible improvements in short-term air quality forecasts? Improved initialization of the aloft conditions may also help better represent regional transport and modulate the inferred impact of LBC specification.

Response: We have provided additional information of the OI application to address the issues raised here. In this study we use OI to adjust O_3, NO_2 and NO in the initial conditions files, namely the restart (CGRID) file in CMAQ. We added the following sentences in Sect. 2.3 (L202-204):

"In the CMAQ model, the restart file, called CGRID, is daily generated during the simulation and acts as ICs for the next day. To constrain the biases in ICs, the concentrations of ozone, NO_2 and NO in the restart file were adjusted via the OI method, which is applied every 24 hours at 0:00 Coordinated Universal Time (UTC)."

We agree with the suggestion to utilize more aloft measurements in the OI application. The O_3 lidar data from NASA LMOL are not available routinely in near-real-time (NRT), so it is not an option to be used for air quality forecasting. It was used here to evaluate CMAQ prediction of O_3 profiles. As suggested by Mathur et al. (2018), future air quality forecasting and analysis modeling can benefit from concerted efforts to provide NRT data of ozone aloft on a continuous basis to quantify its contribution to ground-level concentration. We add the following statement in the revised manuscript (see Lines 689-696 in Section 5. Summary):

"The impact of improved initial concentrations through optimal interpolation (OI) is shown to be large in urban areas initially but fades away rapidly. The influence of OI adjustment, however, lingers for a longer period in an area with low emission density where emissions and chemical reactions make a smaller contribution to the O_3 budget than that in the area with high emission density. Such method may be more useful if applied to vertical layers above the ground. Future air quality forecasting and modeling can benefit from concerted efforts to provide near real time data of O_3 aloft on a continuous basis (Mathur et al., 2018), so that improved initialization of the aloft conditions can better represent regional transport and modulate the inferred impact of LBCs on O_3 forecasting."

Reference:

Mathur, R., C. Hogrefe, A. Hakami, S. Zhao, J. Szykman, and G. Hagler.: A call for an aloft air quality monitoring network: need, feasibility, and potential value. Environ. Sci. Technol, 52 (19), 10903–10908, 2018.

7. The combined impacts of the LBC and emission adjustment simulations raise an interesting point on the representativeness of the NAQFC derived LBCs for the "forecast" year. Based on the arguments put forth, it appears that the "new" emission adjustments were applied only to the 3km resolution domain extent. By inference, the emissions utilized in the NAQFC are likely biased high since they did not benefit from these adjustments in the inferred emission trajectory since the NEI year. Conceivably, reducing regional NO_x emissions outside of the study domain will reduce the regional transported O_3 to the study region thereby possibly increasing the already low bias for the high O_3. What emissions were used in the NAQFC runs – did they benefit from similar "refresh" (reductions) relative to the NEI as the those within the LIS domain? Please clarify these aspects and any possible effects associated with a high bias in regional O_3 from NAQFC arising from possible high bias in emissions used in the NAQFC.

Response: The emission refresh was not implemented in NAQFC, which uses the 2014 NEIs for anthropogenic emissions. Without the emission reduction, the changes in NO_x emissions between the inventory and forecast years are not accounted for. Therefore, an over-prediction is expected of all other model inputs are well represented in the model. Due to the biases in other model inputs, such as meteorology and chemistry, O_3 prediction can be either over-predicted or under-prediction, depending on the actual combination of these factors (e.g., see Fig. 2c and 2d in the manuscript). On the high O_3 days, NAQFC over-predicted surface O_3 (Fig. 2c). Using the NAQFC LBCs, the BCON LIS simulation overpredicted O_3 during peak hours (Fig 2d), although it is difficult to attribute all bias to NAQFC emission bias without running a sensitivity test with the NAQFC system, which is beyond the scope of this study.

We have clarified the aspects of NAQFC emission data, and discussed the potential effect associated with the unadjusted NO_x emissions on LBCs and LIS O_3 prediction (Lines 460-465 in Section 3.3):

"Note that the emission adjustment was only implemented in the LIS system, not in NAQFC, which still uses the 2014 NEIs for anthropogenic emissions. Without the emission adjustment, the changes in NO_x emissions between the inventory and forecast years are not accounted for. On the high O_3 days, NAQFC over-predicted surface O_3 during the study period (Fig. 2c). The NAQFC LBCs are likely associated with a possible over-prediction of the regional transport, which can be partially responsible for the BCON LIS simulation overpredicted O_3 during high O_3 days (Fig 2d)."

Reference:
Tang, Y., Bian, H., Tao, Z., Oman, L. D., Tong, D., Lee, P., Campbell, P. C., Baker, B., Lu, C.-H., Pan, L., Wang, J., McQueen, J., and Stajner, I.: Comparison of chemical lateral boundary conditions for air quality predictions over the contiguous United States during pollutant intrusion events, Atmos. Chem. Phys., 21, 2527–2550, https://doi.org/10.5194/acp-21-2527-2021, 2021.

8. Even though the study promotes high resolution modeling, much of the analysis focuses on aggregate metrics (averaged over model-station pairs). Thus, the relative advantages of using the 3km resolution are not readily apparent. Some discussion of the gains realized from higher resolution (say even relative to the 12km NAQFC) would be useful.

Response: Thank you. We have added additional discussion of the gains realized from higher resolution in the revised manuscript:

In Section 3.1 (L322-327):

"Compared to that in the NAQFC prediction, the O_3 from the 3 km BCON run demonstrated more detailed spatial distributions in the predicted O_3 fields. For instance, the O_3 concentration over the Long Island Sound is lower than its surroundings and the 3 km simulation could reproduce this pattern while the O_3 from the 12 km NAQFC showed a relatively coarser pattern of the concentration gradient. The O_3 distribution along the coastal area also agrees better with the observations than the 12 km NAQFC prediction. This proves the high-resolution simulation can better reproduce the pollutant variability over this coastal urban area."

In Sect. 4.1 (NO_2 VCD distribution), L581-586:

"Moreover, the VCD prediction from the BOE run presents a northeastward pattern and it was lower over water area of LIS than that over surrounding lands. In comparison, the VCD from NAQFC shows a high NO_2 plume over the land and the water around LIS. This spatial distribution from BOE is more consistent with that of GCAS compared to that from NAQFC. Similarly, this situation is also similar for the prediction of surface NO_2 distributions (Fig. 3), indicating the high-resolution system can outperform NAQFC through resolving the fine-scale processes."

In addition, we compared the hourly variations of observed and simulated O_3 and NO_2 at individual monitoring sites. We found that the results from the 3 km system showed better performance, especially in predicting the peak values and timing of O_3 concentrations during the exceedance the NAAQS. This may be partly attributed to the high resolution of the simulation. So we added the analyses in L637:

"As the emissions and meteorological inputs play important roles in determining the magnitude and timing of high peaks (Pan et al., 2017), the emissions and meteorological data at the 3 km resolution could improve the simulation of peak values and timing, especially over urban areas."

9. L76-77: "complex urban areas" is a vague – please elaborate on the specific challenges for air quality forecasts.

Response: We have added additional information to elaborate on the specific challenges (L75): "Prior studies have also revealed that air quality models face additional challenges in predicting surface O_3 concentrations at coastal locations or over complex urban areas, including uncertainties in vertical mixing, deposition processes, spatial-temporal allocation of emissions to the air quality models (Hogrefe et al., 2007; Tong et al., 2006)."

10. L82: perhaps should say six different sources or six different representations of lateral boundary conditions.

Response: Thank you. We have changed it to "examined the impact of six different sources of lateral BCs" (L82)

11. L93: Is "modeling techniques" the correct terminology for the experiments that largely investigate different forms of input data to the CTM? I wouldn't necessarily characterize changes in input data (initial conditions, boundary conditions, and emissions) as modeling techniques, especially since the methodology (at least the emission adjustment and OI) appear to be based on previously investigated methods.

Response: These experiments include both the input data, and the underlying tools/techniques to prepare these data. Although this study focuses on improving input data to the model, as the reviewer pointed out, each input was prepared by a tool or an algorithm, which was commonly referred to as modeling techniques, such as Tang et al. (2017) that called OI used to improve initial conditions as a chemical data assimilation technique. Admittedly, these techniques are not applied to alter the CTM, the core of the air quality forecasting system, but to alter the inputs, which are also parts of the overall modeling system in a broad sense.

Reference:
Tang, Y.H., Pagowski, M., Chai, T.F., Pan, L., Lee, P., Baker, B., Kumar, R., Delle Monache, L., Tong, D., Kim, H.-C., A case study of aerosol data assimilation with the Community Multi-scale Air Quality Model over the contiguous United States using 3D-Var and optimal interpolation methods, Geosci. Model Dev., 10, 4743–4758, https://doi.org/10.5194/gmd-10-4743-2017, 2017.

12. L111: It's not clear to me what does domain size has to do with fine-scale processes - I would have thought grid resolution would be more influential in that regard – perhaps this sentence would benefit from some restructuring.

Response: Thank you. We have rephrased this sentence to avoid confusion (L113):

"While this model domain is large enough to capture key physical/chemical processes within the LIS area, such as sea breeze circulation and photochemistry, the influence of regional transport outside this domain cannot be adequately represented."

13. L181: it was not apparent to me what the 11 ×11 grid cell block refers to and what its relevance is? If it's an area of influence in the OI, please provide some rationale for its choice?

Response: Yes, the 11×11 grid cell block refers to the area of influence in the OI. This area of influence choice is based on correlation length analysis. According to the calculations of correlation length shown in Chai et al. (2017), it showed a correlation length scale of ~84 km, which is the separation distance where the corresponding correlation coefficient falls to e^{-1} when using the National Meteorological Center (NMC) approach, while the Hollingsworth-Lönnberg method has a longer correlation length scale (~160 km). Chai et al. (2017) chose 84 km as the OI influence range for the CONUS domain. Moreover, this influence length scale also varies from region to region. Over remote regions, the length scale may be longer and it should be shorter over polluted areas as the correlation reduces more rapidly. Considering the high emission density and high model resolution over the LIS area, we chose a shorter influence length (33km) for a higher correlation threshold (r >= 0.5) for the LIS area. We have provided this explanation in the revised manuscript (L204).

Reference:
Chai, T., Kim, H. C., Pan, L., Lee, P., and Tong, D.: Impact of Moderate Resolution Imaging Spectroradiometer aerosol optical depth and AirNow PM2.5 assimilation on Community Multi-scale Air Quality (CMAQ) aerosol predictions over the contiguous United States, J. Geophys. Res.-Atmos., 122, 5399–5415, https://doi.org/10.1002/2016JD026295, 2017.

14. Was the emission "refresh" applied only to NO_x emissions or were emissions of other species also modulated relative to the base NEI. One would imaging that VOC emissions would have also changed between 2011-2018. Please also clarify why the 2011 NEI (L200) is used when *conceivably updated (and closer to the forecast year) versions of the NEI (2014, 2017) may have been available?*

Response: The emission refresh was applied only to NO_x emissions, not to the emissions of other species. VOC emissions may have changed as well, but the emission sources are different for VOCs and NO_x. Previous studies revealed that decrease in NO_x emissions has shifted the transition zone of NO_x-saturated to NO_x-limited regimes closer to urban centers and it showed increasingly NO_x-limited ozone chemistry in warm seasons in the NYC area (Jin et al., 2017; Jin et al., 2020). This means the regional O_3 production is more controlled by NO_x emissions than VOCs. Finally, unlike NO_x, biogenic sources contribute a large portion of the VOCs emissions during summertime. In this study, the biogenic VOC emissions are calculated inline, so the biogenic VOC emissions are updated using real-time meteorology.

When we started to prepare this forecasting system, the 2014NEI had not been finalized and well tested. Therefore, we chose the well vetted 2011NEI version 2 to provide the emission input file. For this study, the outdated NEIs also provided the opportunity to apply and evaluate the emission adjustment approach to using the observed trends from NEI years to the forecast year.

Reference

Jin, X. M., Fiore, A. M., Murray, L. T., Valin, L. C., Lamsal, L. N., Duncan, B., Boersma, K. F., De Smedt, I., Abad, G. G., Chance, K., and Tonnesen, G. S.: Evaluating a space-based Indicator of surface ozone-NOx-VOC sensitivity over midlatitude source regions and application to decadal trends, J. Geophys. Res.-Atmos., 122, 10231–10253, 2017.

Jin, X., Fiore, A., Boersma, K. F., Smedt, I. D. and Valin, L., Inferring Changes in Summertime Surface Ozone–NOx–VOC Chemistry over US Urban Areas from Two Decades of Satellite and Ground-Based Observations. Environ. Sci. Technol., 54(11), 6518-6529, https://doi.org/10.1021/acs.est.9b07785, 2020.

15. L150: Were model estimates of isoprene concentrations compared with observations from the LISTOS study? What may be the possible role of uncertainties in isoprene emissions within the LIS region on model predicted O_3 and its discrepancies relative to observations?

Response: We have added new comparisons of the predicted and observed isoprene concentrations (see the figure below). There were limited measurements of isoprene in the LIS region, and the field campaign for VOC sampling occurred outside the study period. Instead, we compared the simulated hourly isoprene with AQS observations from a monitoring site in Bronx, NYC (Fig. 4). The predicted isoprene concentration agrees well with the observations regarding both levels and diurnal patterns, except underpredicting the peak values. The correlation coefficient is 0.93, higher than that of NO_2. While the evaluation is limited, the results indicate that the role of VOC uncertainties may not be a major concern, although future study is guaranteed to look into this issue with more details. We added the figure below and relevant discussion in the Supplemental Information (see Figure S5).

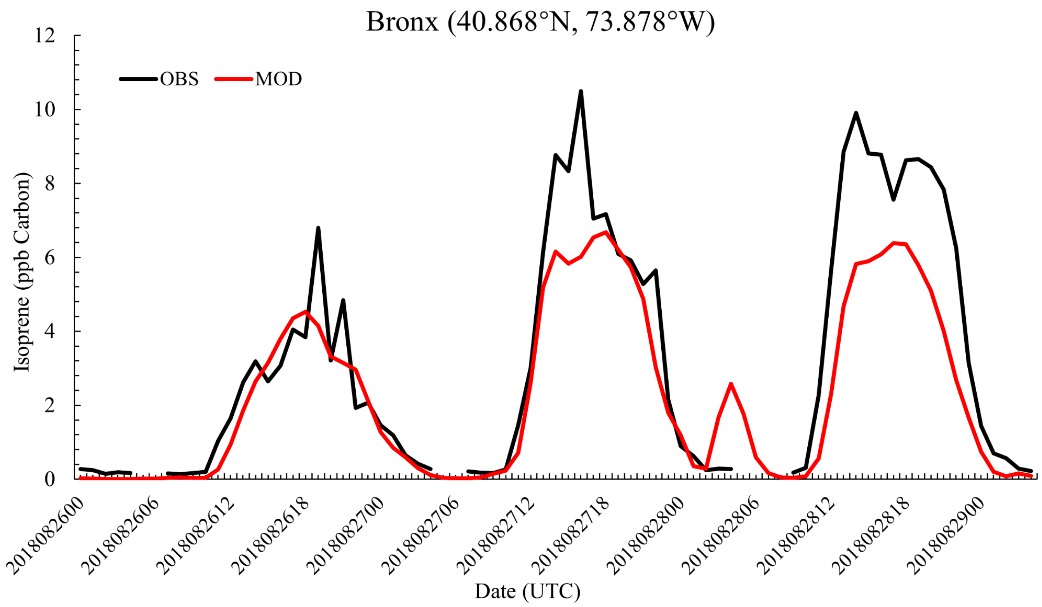

Figure 4. Observed and simulated hourly isoprene concentrations at Bronx, New York (40.868°N, 73.878°W) during the episode.

16. Figs 2d and 3d - please state the time zone on the x-axis label - looks like UTC?

<u>Response:</u> The time zone (UTC) has been added to the x-axis.

17. L310-312: Please restate here and indicate in Table 2 caption the length over which these metrics are computed - are these for hourly paired model and observations.

<u>Response:</u> The averaging length is hourly for the period of August 26 to 31, 2018. We have added this information to Table 2 and revised the text (L373):

Table 2 caption: "Regional mean statistical metrics between hourly observed and simulated O_3 from August 26 to 31, 2018 over the Long Island Sound region".

New L376: "The statistical metrics calculated from hourly simulated and observed data from August 26 to 31, 2018 were reported in Table 2, the RMSEs for O_3 are reduced from 14.97 ppbv to 13.72 ppbv in the OI_bias run, to 13.79 ppbv in the OI_idw run and to 14.30 in the OI_avg run."

18. L323-324: This is somewhat of a misleading statement - the spike is just an indication of the model error at a specific time and location - not necessarily the impact of OI in large metropolitan areas. Please reword this sentence.

<u>Response:</u> Thank you. We have reworded the sentences as follows (L400):

"In large metropolitan areas, OI adjustments result in spikes in large metropolitan areas indicate the model errors at the time of OI adjustment at the monitor sites, with the mean errors being up to 14 ppbv in surface hourly O_3 concentrations over NYC and 16 ppbv over Philadelphia, respectively."

19. L325-327: I find the suggestion that the magnitude of the OI adjustment is related to the emission strength/density to be speculative and not substantiated by any presented analysis. Restructuring the discussion would be useful.

Response: Thank you. We revised the sentence and have deleted the speculative part of the statement (L404):
"The New Haven–Hartford region sees a smaller change of O_3 concentration compared to between that in large cities."

20. L327: The OI effect in large cities – this is to be expected since emissions at surface are the dominant forcing and thus one would expect the OI signal to get swamped out more rapidly in locations with higher emissions.

Response: This discussion here is meant to show the effectiveness and limitations of the OI method to adjust initial concentrations, and to provide quantitative results of the OI influence time in different cities in the LIS region. We added the discussion of using aloft measurements in the future when such data become available routinely.

21. L330: please clarify if these durations of the difference imply a corresponding improvement also in model skill?

Response: The different durations indicate the influence time of OI adjusted ICs, not necessarily the improvement in model skill, which is determined by both initial concentrations and other processes (chemical production and transport, etc). The improvement using the OI adjustment is similar in different subdomains (Table S3). We added this discussion to the revised manuscript (L410).

22. L347-348: Consider qualifying this statement for the specific use in a forecast system. The NEI's are updated for this very reason - to capture changes in emissions. Perhaps they are not available in time for the forecast application and thus the need to project the NEI's to the forecast year. As written the sentence is somewhat open to misinterpretation – note that a specific year is attached to each NEI to indicate the period of its representativeness.

Response: Thank you for this suggestion. It has been changed to

"This trend highlights the importance of updating the emissions to the model year, in order to reduce the bias in the emission inputs for model simulations, especially for time-sensitive applications such as air quality forecasting." (L434)

23. L350: Please elaborate why the uniform and spatially varying emission adjustments result in similar predictions? Do the likely differences get averaged out in the aggregate comparisons?

Response: There are several reasons causing the similar performance using the uniform and spatially varying emission adjustments. First, regional transport of air pollution results in dispersion of emitted NO_x and its byproducts/reservoirs. The observations from satellite or ground monitors, based on which the emissions were adjusted, may not accurately capture the temporal evolution of the emission sources. A large geographical range may better reflect the overall changes of NO_x emissions in the LIS region. Previous studies either use a coarse model resolution (e.g., 1 degree in Lamsal et al., 2011, or state-level adjustment in Tong et al., 2016). Second, the AQS sites in the city regions are usually located near high emission density areas. Similarly, satellite observations are weighted more toward urban plumes. The O_3 production in these places is NO_x saturated, where the O_3 formation is less sensitive to changes in NO_x emissions. As a result, the simulated concentrations using different methods were very close and the limited difference can also get averaged out when calculating the averaged statistical metrics. To illustrate the small difference, Table 4 and Table 5 have been added in the revised manuscript showing the metrics in different subdomains. The tables and the discussion above have been added to the manuscript (L445).

Table 2: Statistical metrics of O_3 simulations after NO_x emission adjustment in different sub-regions from August 26 to 31, 2018

	EmisAdj_avg				EmisAdj_sub			
Domains\Stats	CORR	RMSE	NMB	NME	CORR	RMSE	NMB	NME
NYC	0.78	15.54	-34%	36%	0.78	14.93	-32%	35%
PH	0.78	15.29	-30%	35%	0.78	15.38	-31%	35%
NHH	0.85	13.24	-25%	31%	0.85	13.24	-25%	31%
PP	0.81	17.26	-31%	35%	0.81	17.06	-30%	34%
OTHR	0.84	12.24	-24%	29%	0.84	12.17	-24%	29%
Average	0.81	14.71	-29%	33%	0.81	14.55	-28%	33%

Reference
Lamsal, L. N., Martin, R. V., Padmanabhan, A., Van Donkelaar, A., Zhang, Q., Sioris, C. E., K. Chance, T. P. Kurosu, Newchurch, M. J. Application of satellite observations for timely updates to global anthropogenic NOx emission inventories. Geophysical Research Letters, 38(5), https://doi.org/10.1029/2010GL046476, 2011.
Tong, D., Pan, L., Chen, W., Lamsal, L., Lee, P., Tang, Y., Kim, H., Kondragunta, S. and Stajner, I.: Impact of the 2008 Global Recession on air quality over the United States: Implications for surface ozone levels from changes in NOx emissions, Geophys. Res. Lett., 43(17), 9280–9288, https://doi.org/10.1002/2016GL069885, 2016.

24. L355-357: Please explain what "O_3 production is NO_x saturated" implies - is ozone essentially titrated at the monitors examined? If so, shouldn't there be a low bias in the case where emissions were not adjusted? What LBCs did the emission adjustment runs use?

Response: In the NO_x saturated chemical regime, the ozone production efficiency is low, compared to that in a NO_x sensitive chemical regime. In this case, O_3 level is not very sensitive to NO_x change. Therefore, the difference in O_3 prediction between the two adjustment methods is small.

The default LBCs (same as in the Control run) were used here, since this section focused on the effect of a single adjustment (emission).

25. L374-375: Please state what time average values these metrics are computed for and over what time-period.

Response: This information has been added in the revised text (L487) and the caption of Figure 8:

"Figure 8 compares the predicted hourly O_3 and NO_2 concentrations against in-situ observations from August 26 to 31, 2018 in five subdomains and the overall domain with Taylor diagrams (Taylor, 2001)"

26. L382: Please clarify what are the 3 adjusted runs? L370-374 indicate two adjusted runs? what does the reader associate runs #2,3,4 with?

Response: We have clarified in the revised manuscript that there are three adjusted runs: BCON (dynamic BC), BO (dynamic BCs+OI) and BOE (dynamic BCs+OI+EmisAdj). We have changed the statement in L382 (L493 in revised manuscript):

"The three adjusted runs, namely BCON (#2), BO (#3) and BOE (#4) in diagrams, have …" to replace the "#2,3,4".

27. L384-385: It was not clear to me how one could infer concentration levels and over/underestimation from the Taylor plots in Fig 8 – please elaborate.

Response: Thank you. We added more information to help understand the diagram in L488: "In the Taylor diagram, the relative skill of each forecasting system to reproduce the O_3 and NO_2 variability is represented using three statistical metrics: correlation (R) with values on arc of the right angled sector, normalized standard deviation (SD) with values on y-axis, and centered root-mean-square difference (RMSD) with values on x-axis. The normalized SD is shown as the dashed line concentric circles while RMSD is shown as line concentric circles with the observation point acting as center (OBS on the x-axis). Their values higher (lower) than 1 indicate biased high (low) of the simulations. In general, the forecasting skill is measured by the distance to the OBS point on these diagrams."

28. L390: Please reword this sentence - ICs impact the initial state, BCs do not.

Response: This sentence has been rephrased (L506):

"Concurrent improvements of boundary conditions and initial concentrations allow a more realistic initial state and boundary conditions to demonstrate the effectiveness of the emission adjustment in improving O_3 forecasting (Fig. 9)."

29. Fig 10 and associated discussion: Presumably, the NAQFC which used older emissions overestimated the NO_x emissions relative to the BOE - why then does it consistently predict lower NO_2 relative to the BOE? Is it due to resolution or representativeness of emissions to the "forecast" year?

Response: Yes, it is due to the spatial distribution of the emission. The NAQFC emission input is based on the 2011 National Emission Inventories (NEI), while the emission in the LIS system is updated to the forecast year (2018). The larger bias of NO_2 prediction is primarily attributed to the model resolution, with the higher resolution runs are more capable of describing the spatial distribution of emission sources.

30. L459-460: Are the meteorological drivers also not different between NAQFC and BOE simulations? Could this also not influence the comparisons of the chemical constituents?

Response: The BOE uses different meteorological fields from NAQFC, therefore the comparisons are certainly influenced by the difference in the meteorology. While a quantitative attribution to meteorology and emission differences is difficult to obtain, the magnitude of O_3 improvement from the base run to the BOE run is compared to that of reduced O_3 bias. We added the following statement to emphasize the limitation of this comparison (L616):

"Compared to NAQFC, BOE demonstrates enhanced prediction skills at all sites (Fig. 12). Note the comparisons may be attributed to the differences in meteorology, emission and other factors. Although it is difficult to attribute the improvement quantitatively to each factor, the magnitude of O_3 improvement from the base run to the BOE run is compared to that of the overall reduced O_3 bias, suggesting a significant contribution from these improvement techniques."

31. L466: If the peaks are underestimated, then should it not be expected that the false alarms will also not increase? Not sure what aspect of the BOE the authors are attempting to highlight here? Also, the sample size (number of days, and sites) does not appear large enough to make a conclusive statement on FAR.

Response: We have removed this statement.

32. L470: Could the difference in the timing of the peak also result from differences in the meteorological fields?

Response: Thank you for pointing out this. Research shows that emissions and meteorological fields are important in determining the magnitude and timing of high peaks of O_3 and NO_x, in particular wind direction is critical in determining the timing and location (Pan et al., 2017). Improving the meteorological fields will help improve the simulations of peak O_3 timing and concentrations (Bei et al., 2008). We also added this discussion in the text (L635):

"This may be in part attributed to the high resolution of the LIS3km system, which can better resolve meteorology and emission variations. As the emissions and meteorological fields play important role in determining the magnitude and timing of high peaks (Pan et al., 2017), emissions and meteorological data with 3 km resolution could improve the simulation of peak value and timing, especially over urban areas."

Reference

Bei, N., de Foy, B., Lei, W., Zavala, M., and Molina, L. T.: Using 3DVAR data assimilation system to improve ozone simulations in the Mexico City basin, Atmos. Chem. Phys., 8, 7353–7366, https://doi.org/10.5194/acp-8-7353-2008, 2008.

Pan, S., Choi, Y., Jeon, W., Roy, A., Westenbarger, D. A., and Kim, H. C.: Impact of high-resolution sea surface temperature, emission spikes and wind on simulated surface ozone in Houston, Texas during a high ozone episode, Atmos. Environ., 152, 362–376, https://doi.org/10.1016/j.atmosenv.2016.12.030, 2017.