**Responses to Reviewer #2:**

We thank the reviewer for the constructive comments, in particular the many valuable and detailed technical suggestions that have helped improve this manuscript. Please see below our point-to-point responses to these comments (*in Italic*).

Major Comments:

*1. Using static boundary conditions (BCs) is not a viable option for a small, polluted domain with well acknowledged transport impacts such as the northeastern US. Therefore, the results of applying static BCs are not interesting or useful. As expected, the model performs very poorly with static BCs. The discussion of this unviable option takes up unnecessary space and distracts the reader's attention up until the very end. The trivial conclusion that switching from static to dynamic BCs significantly improves ozone predictability undermines other (and in my view) more important improvements. I recommend removing the discussion of static boundary conditions from the main text. Including it as a supplement might be beneficial for novice readers.*

Response: Thank you for your suggestion. We agree that the clean air BCs are not a viable option for the realistic air quality forecasting over the LIS region. In the final configuration that we used to simulate the high $O_3$ event, the NAQFC BCs were used. The static BC run was presented here for two purposes: First, it is used as the reference to assess the effectiveness of each adjustment method. Second, the difference between the Control run and the NAQFC BCs (BCON) run represents the contribution of regional transport of $O_3$ and its precursors to air quality over LIS. The results show that in-domain $O_3$ production remains relatively constant during peak hours (Figure 2d), but regional transport is key to allow the LIS system to reproduce high $O_3$ events. In addition, our results reveal that the influence of BCs vary by chemical species, more significant for long-lived species such as ozone but not so for nitrogen oxides, which are more sensitive to model resolution (spatially resolved emission and chemical transformation). We believe such results provide useful information to the literature, and are an integral part of this paper to make a full story. Following the reviewer' comment, we have shortened the discussion of static BCs and explained the role of the static BC simulation in the study 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."

*2. The rationale for using different variations of optimal interpolation (OI) could be presented better. The reader may not be familiar with the OI method and its strengths/weaknesses. Therefore, there should be a short discussion of the expectations with each alternative. The fact that the control case is no initial condition adjustment should be stated upfront otherwise the reader may think that the control is still the static boundary condition. The discussion of the performance with the inverse distance weighting option is insufficient or missing.*

Response: Thank you. We have provided a general description of the OI method and the rationale with each alternative (OI_avg, OI_idw and OI_bias) in Sect. 2.3(b) L218: "Therefore, the region of influence is limited, and the adjusted fields may be discrete in spatial distribution." L221: "With this interpolation, the effect of OI will be not limited near the observational sites and most of the grid cells in the domain can be adjusted comparing to the OI_avg.", and L225: "Unlike the OI_idw which just applied the spatial interpolation to extend the OI effect, in this method the observation cells are distributed to the whole domain grids based on the spatial patterns provided by model so that it is able to better reflect the realistic fields."

We have made it clear in the revised manuscript that the control case uses the static BCs (Section 3.2):

"Here we compare the results using various OI methods with the simulations without any BC adjustment (same as the Control run)."

In addition, we modified Table 1 to clarify the model settings for each run. Finally, we added more description of the performance of each method (Lines 372-378):

"The predicted $O_3$ field with the OI_avg method shows a distribution with localized high value areas near the observation sites. As for the other two OI methods, the distribution using OI_bias has similar patterns with that of OI_idw while the concentrations over the high $O_3$ area are further elevated. Biases between observed and predicted concentrations are reduced in most of the areas. 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."

*3. I disagree with the choice of the domain-average emission adjustment based on performance. The $NO_x$ emission differences between the base year (2011) and current year (2018) are so different for the four subdomains that averaging them cannot be justified. The subdomain emission adjustment is clearly the right choice because it provides the model with the right information. The better performance with the domain average emission adjustment here is a typical case of getting the "right" answer for the wrong reason. I recommend a more detailed, site specific analysis of performance with these two adjustment methods. I expect at least the $NO_2$ performance of the subdomain adjustment to be better at the sites in and around those subdomains.*

Response: We agree that the subdomain adjustment makes more sense and has initially hypothesized that the results using subdomain emission adjustments would be better. To answer this question, we have conducted the model simulations with both full domain adjustment and subdomain adjustment (EmisAdj_sub) during the study period. We found that the performance from the model runs with different adjusting methods was quite similar. So we didn't put the results of EmisAdj_sub in the original manuscript. 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 areas where $O_3$ production is $NO_x$ saturated and 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.

Following the reviewers' suggestion, 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 (L446-454):

"This demonstrates that emission adjustment alone results in limited improvement of $O_3$ prediction, due in part to the fact that the $O_3$ production in this region is $NO_x$ saturated (VOC limited) in urban areas where most AQS monitors are deployed, so the $O_3$ level is less sensitive to the change in $NO_x$ emissions. Similarly, satellite observations are weighted more toward urban plumes. In addition, 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). 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."

**Table 4: 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% |

**Table 5: Same with Table 4 but for $NO_2$**

| | EmisAdj_avg | | | | EmisAdj_sub | | | |
|---|---|---|---|---|---|---|---|---|
| Domains/Stats | CORR | RMSE | NMB | NME | CORR | RMSE | NMB | NME |
| NYC | 0.82 | 4.23 | -22% | 27% | 0.82 | 4.77 | -29% | 31% |
| PH | 0.79 | 5.69 | -36% | 41% | 0.79 | 5.53 | -33% | 40% |
| NHH | 0.49 | 7.69 | -44% | 49% | 0.49 | 7.53 | -41% | 48% |
| PP | 0.67 | 2.92 | -18% | 35% | 0.67 | 2.95 | -21% | 36% |
| OTHR | 0.69 | 2.56 | -33% | 39% | 0.69 | 2.54 | -32% | 39% |
| Average | 0.69 | 4.62 | -31% | 38% | 0.69 | 4.67 | -31% | 39% |

## 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.

*4. Section 4 should include the comparisons of $NO_2$ column and $O_3$ profile measurements during LISTOS with the model using the subdomain emission adjustment.*

Response: Thank you. Following the reviewer's suggestion, we have added the comparisons of $NO_2$ column and $O_3$ profile measurements using the subdomain emission adjustment.

In Sect. 3.4, we added the comparisons between results of domain average adjustment (EmisAdj_avg) and subdomain adjustment (EmisAdj_sub). The results show that the model performance between the two runs is similar. (L483: First, we compare two BOE simulations, one with the EmisAdj_avg emission adjustment and the other with EmisAdj_sub. The statistical metrics of BOE with EmisAdj_avg and BOE with EmisAdj_sub (Table S4, S5) are quite similar in each sub region and also have the same correlations. On average, the RMSEs of BOE (EmisAdj_avg) is slightly smaller. Therefore, in the subsequent evaluation we take BOE (EmisAdj_avg) to compare against surface and other observations.)

In Sect. 4, we added the comparison of $NO_2$ column and $O_3$ profile simulated by combined adjustment using EmisAdj_sub (L595):

"In addition, the $NO_2$ VCD from simulation with combined adjustments using EmisAdj_sub method for emission refresh shows a similar spatial pattern with that of BOE (Fig. S3) while its VCD level over the NYC area is lower, making it underestimates the hotspot but much closer to the VCD over the rest of the areas."), L601 (and unlike the results of surface $NO_2$, the $NO_2$ VCD using EmisAdj_sub has lower NMB (33%) and NME (57%) compared to that using EmisAdj_avg (40% and 61%) while their correlation is still the same (0.74).) and L668 ("The model performance of $O_3$ surface concentrations and vertical distribution using AFs from EmisAdj_sub is very close to those of using the AFs from EmisAdj_avg in the BOE case (Fig S4, Table S7)")

All results of these comparisons (Figures and Tables) have been included in the Supplementary Information.

*Specific Comments:*

*1. L31: "derived from NOAA National Air Quality Forecast Capability (NAQFC)." What is the forecasting system used here? You should give it a name.*

Response: NAQFC is the name of the NOAA operational forecasting system. We have added the information here.

*2. L73: Tong and Tang, 2018. There are certainly older references to cite here.*

Response: Thank you. Two older references were added:

Eder, B., Kang, D., Rao, S. T., Mathur, R., Yu, S. C., Otte, T., Schere, K., Wayland, R., Jackson, S., Davidson, P., and McQueen, J.: A demonstration of the use of national air quality forecast guidance for developing local air quality index forecasts, B. Am. Meteorol. Soc., 91, 313–326, doi:10.1175/2009BAMS2734.1, 2010.

Oliveri Conti, G., Heibati, B., Kloog, I., Fiore, M., Ferrante, M. A review of Air QModels and their applications for forecasting the air pollution health outcomes.Environ. Sci. Pollut. Res.http://dx.doi.org/10.1007/s11356-016-8180-1, 2017.

*3. L113: (NAQFC). You need a reference here.*

Response: A reference has been added:

Davidson, P., Schere, K., Draxler, R., Kondragunta, S., Wayland, R. A., Meagher, J. F., and Mathur, R.: Toward a US National Air Quality Forecast Capability: Current and Planned Capabilities, in: Air Pollution Modeling and Its Application XIX, edited by: Borrego, C. and Miranda, A., pp. 226–234, Springer, Dordrecht, The Netherlands, 2008.

*4. L119: NAQFC. State the resolution (12 km)*

Response: Changed to "… NOAA NAQFC with a horizontal resolution of 12 km were applied …"

*5. L182: $X_b$. Undefined*

Response: We have added the definition here:

"where $x^a$ and $x^b$ are the analyzed and background fields, respectively."

*6. L190: "Experiments were also performed with two different interpolation methods." Are these the same OI methods mentioned above. This paragraph is confusing.*

Response: Thank you for pointing out this. Yes, these are the same OI method mentioned above and the interpolation methods here indicate the method is used for preparing observational data for OI. We remove the first sentence in this paragraph as it was repetitive with L190, and edit this as "Besides this method, experiments were also performed with two different interpolation methods for preparing the observational data."

*7. L213: " $f_S$ is set to 1 and $f_G$ to 100 to avoid dominance by either data source." Sounds like surface data dominance.*

As the monthly OMI data is calculated from the daily files, its temporal resolution is relatively lower. In comparison, the ground-based (AQS) $NO_2$ is measured more frequently (hourly) and can better reflect the local emission situation. Therefore, the AQS data is considered more reliable for calculating the emission changes. Additionally, in the formula, $N_S$ (number of satellite data) and $N_G$ (number of AQS data) depend on the number of grid points in the OMI file and number of observational sites, respectively. So $N_S$ is much larger than $N_G$. We set $f_G$ to 100 to avoid the dominance by OMI data.

*8. L216: AFs from May to September. Why do you call it "rapid" refresh?*

Response: The rapid refresh means the emission input can be adjusted quickly using the observations, compared to the typical time lag of several years in the model emission input files from national emission inventories. This approach has the potential to update emission as soon as the observations are made available (within a day from the forecast time).

*9. L225: obtained from the AQS network. Repetition*

Response: We have removed "obtained from the AQS network and"

*10. L236: Ozone Monitoring Instrument. Defined earlier*

<u>Response:</u> We have removed the "Ozone Monitoring Instrument" here.

*11. L237: aboard the Aura satellite. Repetition*

<u>Response:</u> Removed. Thank you.

*12. L242: Multiyear OMI $NO_2$ data were further aggregated to calculate state-level emission adjustment factors using a mass conservation approach (Tong et al., 2015). This does not say much. Why mass conservation is mentioned here?*

<u>Response:</u> Unlike the concentrations, the fluxes and emissions are related to the area and time. So it is necessary to keep the total emission amount constant during the data processing. The "using a mass conservation approach" indicates the method will keep the processed $NO_2$ emission consistent with the original one.

*13. L245: GeoCAPE Airborne Simulator. Previously defined.*

<u>Response:</u> Removed.

*14. L252: Langley Mobile Ozone Lidar. Previously defined.*

<u>Response:</u> Removed.

*15. L253: LMOL is part of a NASA-sponsored ozone lidar network called the tropospheric ozone lidar network (TOLNet; Sullivan et al., 2019). Why is this relevant?*

<u>Response:</u> Here we provide a short introduction of the lidar data and relevant reference for the readers who may be interested in learning more about LMOL.

*16. L262: "As a reference." Why? NAQFC has 12-km resolution.*

<u>Response:</u> Here we introduce the NAQFC product as a reference/benchmark to compare and evaluate the new LIS 3 km system, aiming to evaluate and analyze the performance of the new high-resolution forecasting.

*17. 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." Well, obviously. There is nothing new or original here. Nobody uses default BCs for their region.*

Response: We moved this statement and further discussed the influence of dynamic BC in Sect. 3.1 (Major comment 1).

*18. L275: "The performance of the two high-resolution simulations was next compared to that by the NAQFC." Again, why? You should do this when you put in all the upgrades.*

Response: The comparison here is for the simulations with each adjustment method, the single method adjusted results in Sect. 3.1~3.3 can compare with the NAQFC results and present the effectiveness of each adjustment method. It also showed the differences of high-resolution simulations with NAQFC on spatial patterns and metrics to the reader at first sight. In Sect. 3.4 and Sect. 4, we did further inter-comparisons for the combined adjustment results.

*19. L286. Are there any tables or figures to support these statements?*

Response: Thank you. We have added Table S2. in the Supplementary Information to support these statements.

*20. L294. I recommend removing the static boundary condition. This section gives the impression that static boundary condition is better than coarse resolution. This may be the case for $NO_2$ but overall, when $O_3$ and $NO_2$ are considered together, static boundary conditions are worse.*

Response: Thank you. We have responded to this comment earlier (see Response to Major Comment #1).

*21. L299. Is the control run still the static BC run? Or, is it the original OI?*

Response: Yes, the Control run uses static BCs, not with the OI adjustment. We have explicitly stated it in the revised manuscript (L356) and further clarified the settings for each run in the revised Table 1.

*22. L302: Fig. 4a. Do you want us to compare 4a and 4b?*

Response: Thank you for pointing this out. It has been revised:

"… regions within five model grid cells in each direction of the observations compared to the Control run (Fig. 4a, b)".

*23. L303: (Fig. 4b, c). Which panels do you want us to compare?*

Response: Thank you. It has been changed to 4c,d: “The $O_3$ fields adjusted by OI_idw (Fig. 4c) and OI_bias (Fig. 4d) show similar horizontal distributions.”

*24. L308: “that.” than*

Response: We have revised this sentence to “The adjusted $O_3$ fields show different patterns compared to that in the Control run with no IC adjustment.”

*25. L308: “the Control run with no IC adjustment.” You should define the control run earlier.*

Response: The Control run here is the same as in the previous sections. We have added the definition of each run Table1 of the revised manuscript.

*26. L310: “The RMSEs for $O_3$ are reduced from 14.97 ppbv to 13.72 ppbv in the OI_bias run and to 14.30 in the OI_avg run.” What about idw?*

Response: Here we added the description for metrics of OI_idw: “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.” (L378)

*27. L314: “Generally, the improvement of the simulated results due to OI data assimilation over the study domain is smaller than that from the dynamic BCs for this particular domain.” You are comparing apples with oranges.*

Response: All LBCs and OI runs are compared to the same Control run to assess the effectiveness of different adjustment methods.

*28. L316: “may yield different relative changes between BCs and ICs adjustments.” Of course.*

Response: Thank you. We have removed this sentence.

*29. L319: “the duration of OI influence.” Perhaps you should remind that you run each day with new ICs.*

Response: Thank you. We added the information in Sect. 2.3 (L203):

“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).”

Again in L397:

“The ICs for each day were adjusted by OI using real time observations.”

*30. L340. Delete “emissions”.*

Response: Removed.

*31. L341: Add “in ground concentrations”*

Response: Thank you. Added “in ground concentrations” to the text.

*32. L343: “The average adjustment factor (AF)”. AF already defined, just refer back to Equation 2.*

Response: Revised this into “The average AF for …”.

*33. L347: “input” should be inputs.*

Response: Revised this into “inputs”. Thank you.

*34. L352: delete “as”*

Response: Removed “as”.

*35. L362: “Considering both $O_3$ and $NO_2$ performance, the domain-average approach (EmisAdj_avg) is selected for subsequent multi-adjustment simulations.” This is counterintuitive. Theoretically, the subdomain approach should be better because it gets you closer to reality, does it not? There is much to be analyzed here. Just because performance is better with the domain-average approach does not mean that is how one should proceed. An analysis of how individual monitors behave is recommended. You are getting a better result for the wrong reason. You should proceed with the better approach.*

Response: We analyzed the reasons why the results from domain-average and subdomain AFs are so similar. We also provided the statistical metrics in each subregion and added the simulation results using the subdomain AF in the relevant sections. See details in the responses to Major Comments #3 and #4, *36. L377: "centered root-mean-square difference (RMSD)." Explain how to read this on the diagram.*

<u>Response:</u> We added the instruction to understand RMSD on the diagram (L489):

"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."

*37. L381: Should be Fig. 8e.*

<u>Response:</u> Thank you so much. We have changed to Fig. 8e

*38. L382: "The three adjusted runs (#2, 3, 4)". Is this the same as BOE?*

<u>Response:</u> These runs are three simulations with different adjustment settings: dynamic BCs alone (BCON), dynamic BCs and OI (BO), and all three methods together (BOE). We revised the text (L499):

"The three adjusted runs, namely BCON (#2), BO (#3) and BOE (#4) run in diagrams, have well reproduced surface $O_3$ concentrations over the NYC region."

*39. L396: "All of the high-resolution simulations, including the Control run, perform better than the NAQFC run (Fig. 9)". For $NO_2$, right?*

<u>Response:</u> Yes, this sentence has been revised:

"All of the high-resolution simulations, including the Control run, perform better for $NO_2$ prediction than the NAQFC run …"

*40. L398: "Since the boundary conditions used by the high-resolution CMAQ runs are derived from NAQFC, the large negative $NO_2$ bias from NAQFC also contributes to the overall bias to the high-resolution runs." This contradicts what you said earlier about $NO_2$ when you were comparing static versus dynamic BCs. You said BCs did not matter for $NO_2$.*

Response: The NO$_x$ is primarily influenced by local emissions because of its short lifetime, the transported source is not as significant as that of O$_3$, although the effect is not zero. We have removed this statement.

*41. L403: "Such a contrast suggests either an underestimate of emission sources in Connecticut, or flawed model chemistry and transport, or a combination of both." What were the results for subdomain emissions corrections?*

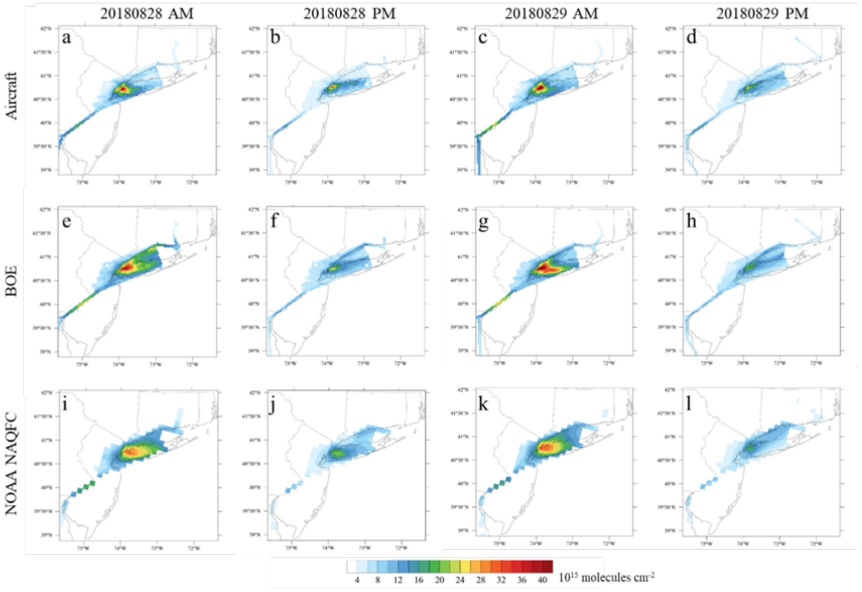
Response: The results for subdomain emissions are analyzed in L446 (as discussed in the response to General Comment #4) of revised manuscript and the metrics are put in Table S4 and Table S5.

*42. L408: "(August 28-29, 2018)". What is different from the week of August 25-31 stimulated before?*

Response: It's the same simulation covered in this study. The O$_3$ concentrations were very high (exceeding the O$_3$ air quality standards or NAAQS) on August 28-29, 2018 and there are LISTOS observations during that time. So here we focus on these two days which have more observational dataset (aircraft and lidar). We change this to "During the high O$_3$ pollution days (August 28-29, 2018) in this episode".

*43. L446: "(Fig. 11e, 11i)". Labels are missing hence I am unable to follow this discussion.*

Response: It has been corrected now. See the revised Figure 11 and caption.



*44. L466: "false alarms". Do you mean a predicted ozone exceedance that does not realize?*

<u>Response:</u> Yes, "false alarms" are used to refer to a predicted exceedance that is not observed. Here the overestimation on August 29, 2018 was not large enough to trigger a false alarm. We have removed this statement to avoid confusion.

*45. L481: "$O_3$ concentrations in the free troposphere are more controlled by regional $O_3$ production and transport than in the PBL." Another reason why static BCs should not be used.*

<u>Response:</u> Yes, and the effect of BCs on local $O_3$ also depends on the simulated seasons. We had a discussion about this in the previous section.

*46. L498: "although the system needs to be further refined to reduce bias." The subdomain emission refresh should be explored more in this simulation.*

<u>Response:</u> The results of BOE using the subdomain emission refresh have been added.

*47. L507: "reproduce" to "simulate"*

<u>Response:</u> Changed.

*48. L520: "One possible direction to explore is to use atmospheric observations to constrain emissions through coupled data assimilation approaches in future efforts." Unclear*

<u>Response:</u> We have revised this sentence to add more information. Since we only applied a simple emission adjustment approach here, there are other ways to constrain emissions. This sentence has been changed to (Lines 699-702):

"One possible direction to explore is to apply other methods to constrain emissions that use both variational (e.g. Elbern et al., 2007; Vira and Sofiev, 2012) and ensemble-based (e.g. Miyazaki et al., 2012, 2017) solutions to analyze the 3D chemical tracers as well as their respective precursor emissions simultaneously."

Reference:

Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O.: Emission rate and chemical state estimation by 4-dimensional variational inversion, Atmos. Chem. Phys., 7, 3749–3769, https://doi.org/10.5194/acp-7-3749-2007, 2007.

Miyazaki, K., Eskes, H. J., and Sudo, K.: Global $NO_x$ emission estimates derived from an assimilation of OMI tropospheric $NO_2$ columns, Atmos. Chem. Phys., 12, 2263–2288, https://doi.org/10.5194/acp-12-2263-2012, 2012.

Miyazaki, K., Eskes, H., Sudo, K., Boersma, K. F., Bowman, K., and Kanaya, Y.: Decadal changes in global surface $NO_x$ emissions from multi-constituent satellite data assimilation, Atmos. Chem. Phys., 17, 807–837, https://doi.org/10.5194/acp-17-807-2017, 2017.

Vira, J., & Sofiev, M. On variational data assimilation for estimating the model initial conditions and emission fluxes for short–term forecasting of $SO_x$ concentrations. Atmospheric Environment, 46, 318–328. https://doi.org/10.1016/j.atmosenv.2011.09.066, 2012

49. L526: GCAS. Are you redefining all the acronyms in the summary?

Response: We removed the repeat definition.

*50. L531: Data Availability. What about the other data?*

Response: In the revised manuscript we have provide the sources of the other data used in this study:

"CMAQ and SMOKE source code is available on the Community Modeling and Analysis System (CMAS) Center of University of North Carolina, Chapel Hill: https://www.cmascenter.org/ (last access: July 31, 2021). WRF is an open-source community model. The source code is available at http://www2.mmm.ucar.edu/wrf/users/download/get_source.html (last access: July 31, 2021). The AirNOW hourly data of $O_3$ and $NO_x$ is available at https://files.airnowtech.org/?prefix=airnow/ (last access: May 2021) and the hourly $NO_x$ data from US EPA Air Quality System (AQS) surface network is available at https://aqs.epa.gov/aqsweb/airdata/download_files.html (last access: May 2021). The GCAS $NO_2$ vertical column density and the LMOL $O_3$ vertical profile data are available at https://www-air.larc.nasa.gov/missions/listos/index.html (last access: May 2021). The monthly product of $NO_2$ vertical column density from OMI is available at https://avdc.gsfc.nasa.gov/pub/data/satellite/Aura/OMI (last access: July 31, 2021)"

*51. Table1. Is the control the same in all the figures? The control in Figure 2 is much different than the control in Figures 4 and 5.*

Response: Yes, this Control run is the same in all figures. In Figure 2, the spatial distribution is daily concentration, the values are averaged, while Figures 4 and 5 show the hourly distributions.

*52. Figure2. The label on the panel is "default".*

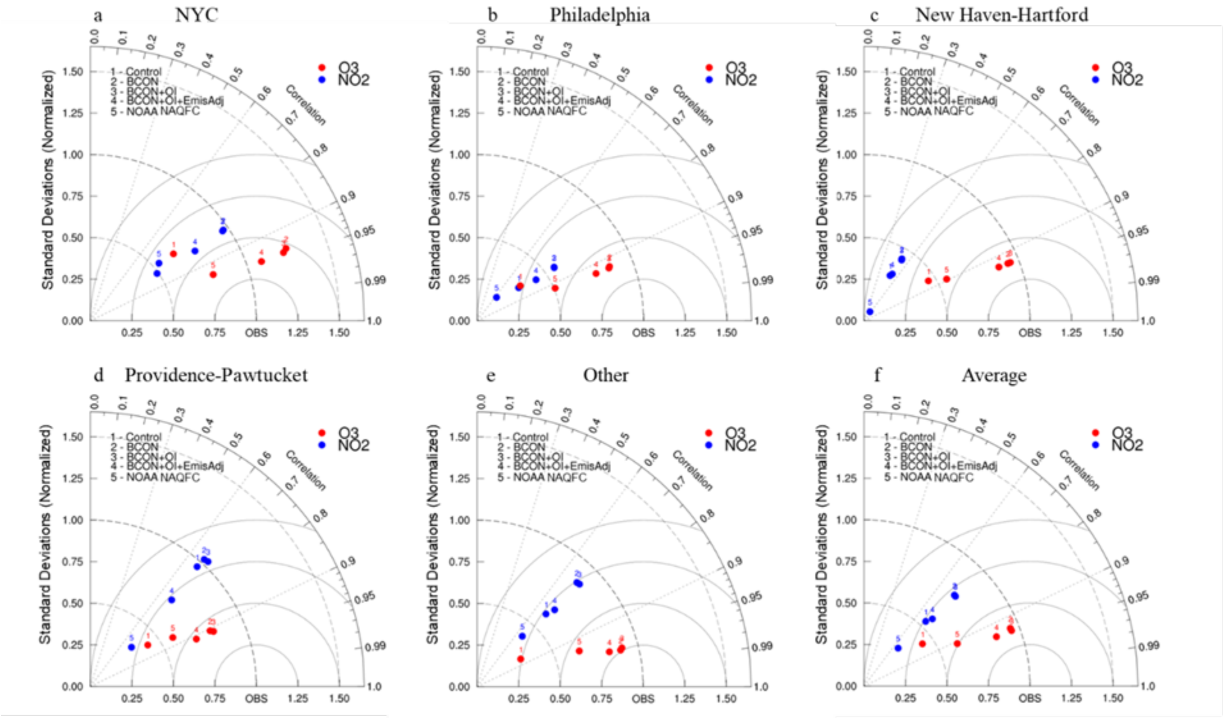
Response: We have changed it to the Control run, which uses the default BC profile in CMAQ.

*53. Figure 6. Better call it OI*

Response: Thank you for your suggestions. Revised this to "OI adjusted initial concentrations".

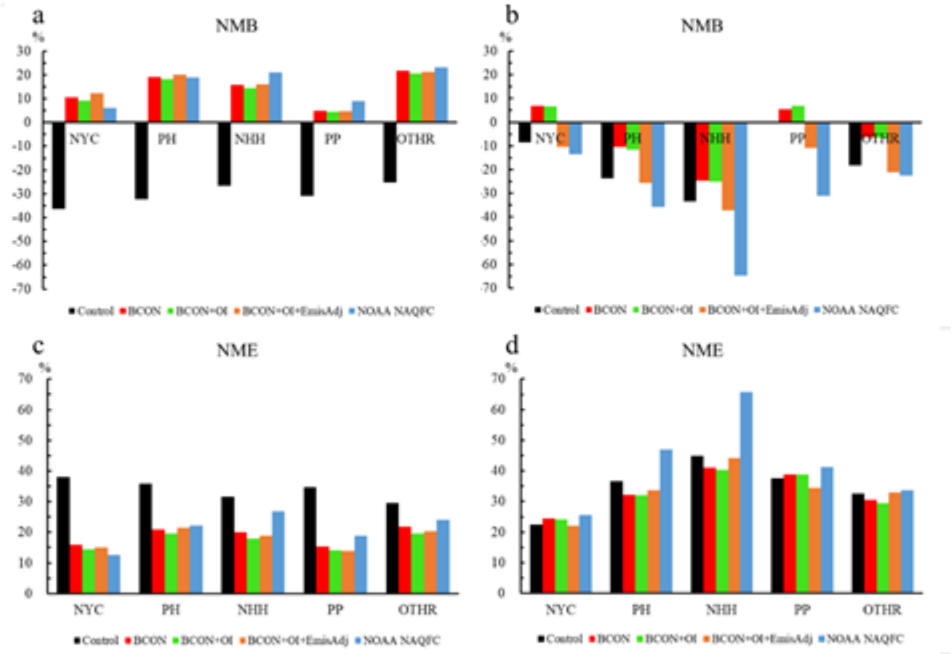
*54. Figure 8. Change standardized to standard on the y axis.*

Response: Thank you. It has been changed in the revised Figure 8.



*57. Figure 9. The magnitude cannot be % in a and b.*

Response: This was a typo and has been corrected. Thank you!



*58. Figure 10. Label these sites on Figure 1.*

<u>Response:</u> Thank you. The labels have now been added to Figure 1:

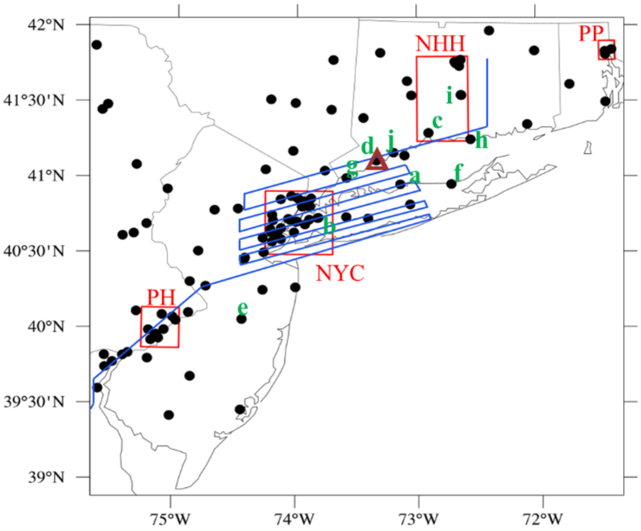


Figure 1: Study area over the Long Island Sound and the surrounding region. Red boxes depict four subdomains: New York City (NYC), Philadelphia (PH), New Haven-Hartford region (NHH), and Providence-Pawtucket region (PP). Black circles indicate the locations of EPA air quality system (AQS) ground monitors, the brown triangle indicates the TOLNet $O_3$ site located in Westport, CT, and the blue lines present an example flight path conducted by the NASA B200 aircraft on August 28~29, 2018. Letters a–j indicate the monitoring sites: a) Flax Pond, b) Queens

College, c) New Haven, d) Westport, e) Colliers Mills, f) Riverhead, g) Greenwich, h) Madison-Beach Road, i) Middletown-CVH-Shed and Stratford.