This manuscript reports the extensive comparison of the REAM chemical transport model (CTM) simulations with the NOx and NOy observations acquired during the DISCOVER-AQ 2011 over the Baltimore and Washington-DC area. The observations include the data from surface monitors, PANDORA, P3 aircraft, ACAM, and satellites OMI and GOME-2. The model results with two spatial resolutions, 36 km and 4 km are compared in order to elucidate the impact of the resolutions on the model NOx and NOy simulations. Differences between the model and observations are discussed in details and causes for the discrepancies are suggested.

The manuscript reflects the extensive works dealing with almost all available data sets to evaluate NOx measurements and CTM results over the Baltimore and Washington-DC area for July 2011. I appreciate the efforts the authors made for this study. The manuscript will be more valuable if quality of presentation and interpretation of the results are enhanced.

The main focus of the paper seems to be the comparison of the model simulations with the 36 km and 4 km resolution and advocate the use of 36 km in the end. I think the authors should focus more on the analysis of 4 km resolution results and causes for the similarities and discrepancies with various observations. The emissions at 36 km resolution are simply accumulations of the emissions at 4 km. It is not important to compare the emissions at the two resolutions and judge which one is better. The authors have the best spatial resolution of emission inventory data and the model simulations at the comparable scale (4 km). If the model overestimates the NOx, NOy observations at one height or vertically column integrated, that simply means the model emissions are overestimated. For the pollution hot spots in the domain, the model values are higher than the observations (judging from the ACAM data). This may be about the spatial location error in the NEI as the authors jumped to the conclusions, but it is more probable that the uncertainties in the emission factors (or activities) over populated urban or roads as represented as MOVES caused the problem. Section 3.7 should be deleted or rewritten. This section is confusing and misleading.

Except for Figure 10, the model results and observations were not analyzed at the measurement sites. The plots are all averages for large domains and many sites. As the ACAM data demonstrate, there are heterogenous distributions of NOx at fine scales. This is important. The model results should be compared at each site of PANDORA and P3 spiral locations.
Figure 2 to 5 (and Figure 9) are about the surface monitor data and interpretation. There are many other interesting, important data sets from the DISCOVER-AQ campaign, which is discussed in short compared to the surface routine monitors. The interpretation of nighttime PBL height (or PBLH as in the model output name) may be right, or may be wrong. As authors mentioned in the manuscript, this may be due to overestimation of nighttime emissions, which can not be ruled out. People do not know much about nighttime PBL height and nighttime emissions. PBLH in the model output is not simply PBL height during nighttime. The nighttime PBL height from YSU scheme is sometimes recalculated based on many other nocturnal PBL height definition. Thus Figure 4 may need to be carefully revised or explain limitation of this analysis. Figure 7 is too busy. It is difficult to see the details. More expansion of analysis in Figure 7 or a summary in Table would be helpful. A plot comparing satellite NO$_2$ spatial distributions and more presentation and discussions on the GMI, TM4, REAM as a priori for the retrieval would be useful. Authors frequently use the figures in Supplementary Material. It is difficult to read the manuscript with many supplementary figures. Because there are important plots in the supplementary, I suggest to move some of the plots in the supplementary to the main manuscript. For example, Figure S1, S2, S17, S21, S22, S23 and discussions about them would be useful. Differences between the model and ACAM NO2, differences between two ACAM NO2 retrievals, and differences between weekdays and weekends are interesting and can have important implications for emissions and model assessments. Figure S12, S13, and S19 or one of them can be also shown in the main text. For S19, one-to-one comparison of the model simulations at 36 km and 4km resolution would be more useful. It is difficult to understand the purpose of Figure 8.

Regarding WRF model options, I am wondering why Single Moment 3 Class microphysics scheme is used. This scheme is for warm clouds. There are Grell ensemble or other Grell cumulus parameterization options that were widely tested in CTM groups. Because the REAM model is an offline model, the performance of model at 4 km resolution may not be caused by original WRF physics, but by the integrator between the WRF and REAM. The performance of model at 12 km was not discussed, but on/off of cumulus parameterization option at this resolution may be another factor to be tested. For the analysis like Figure 6, the model temperature (potential temperature), moisture (specific humidity), U, and V also need to be analyzed with observations, particularly for afternoon.

One minor point is frequent use of red/green combinations in the plots, which is not ideal.

If the manuscript is revised following the comments above, I think publication can be reconsidered.