We thank the referee #4 for giving valuable comments. We respond to each specific comment below. The comments and questions from referee #4 are in italic font.

This paper by Ding et al. focuses on the estimation of NOx emissions during the 2014 Youth Olympic Games in Nanjing. They constrain daily NO2 column observations from OMI and simulations from the regional CHIMERE model to infer NOx emissions. I agree with one of the reviewers that the most significant results are presented in Figure 9 and 10, which I am most concerned with. I do not think this paper is suitable for publication in ACP unless substantial revisions are made.

I agree with most comments from Reviewer #2. I have few additional comments:

Although we do not know which specific comments the referee is referring to, the responses on those comments can be found in the answers to referee 2.

1) MEIC inventory as well as Zhang et al [2009] inventory suggests small monthly variation in emissions. Emission estimates from the DECSO algorithm is suggesting ~50% higher emissions in July than January. Small drop in February is explained by previous publications by Zhang et al., but the seasonal variation in Figure 10 looks unrealistic based on all existing bottom-up inventories over China. It is most likely coming from deficiency of the DECSO algorithm or the CTM they use. Satellite retrievals may have seasonal bias, yielding seasonal biases in NOx emissions. There are several factors that could lead to the biased inversion. Either exploring those factors or providing enough justification why bottom-up emission is wrong is necessary.

Unfortunately we couldn't infer from literature how the seasonal cycle in MEIC is constructed. We assume that it is similar as described in Zhang et al. (2009). According Zhang et al. (2009) the seasonal cycle is superimposed on the annual inventory. This seasonal cycle is based on monthly activities per province for industrial activities, while for residential emissions it is based on the method of Streets et al. (2003). This seasonal cycle for the residential sector is based on assumptions on the use of stove operation as function of temperature, and is derived for the whole of China.

The difference with our work is two-fold:

1) Our results are for 2013 and 2014, more than 10 years after the analysis of Streets et al. (2003), in which time a lot has changed in the Chinese society. For the city of Nanjing there is practically no coal used for heating, but on the other hand the use of air condition has increased strongly in the last 15 year.

2) The seasonal cycle was applied to the bottom-up inventory on a provincial level or even national level, while we are presenting results at a city-level (Nanjing). As shown in H. Zhang et al., (2009) electricity consumption is higher in summer than for winter for this specific city.

We would also like to add that the used NO_2 satellite data has been validated and show no significant seasonal cycle. (e.g. Boersma et al., 2011 and references therein) .

In the text we added the following discussion: "The difference with the seasonal cycle of MEIC might be attributed to the fact that our results are derived on city-level, while the seasonal cycle for bottom-up inventories are often derived on a national or provincial scale (e.g. Q. Zhang et al. , 2009)."

2) Use of OMI data: There might be some limitations in the understanding and use of OMI data. I think, DOMINO algorithm accounts for aerosol effect through not just cloud information as discussed in the paper but also surface reflectivity (OMI LER). Exclusion of scenes with high aerosols may remove polluted days since high aerosols may occur for days with high NO_2 pollution. Results from Lin et al., who use MODIS reflectivity and model aerosol, may not provide sufficient justification as the study did not examine the relationship between Kleipool LER with LER calculated from MODIS reflectivity and observed aerosol. Discussions on the application of averaging kernel are necessary since the idea here is to replace the TM4 profiles used in retrievals by DECSO profiles. Section title "Improvement of the satellite data" is misleading because this work does not improve any aspect of retrieval algorithm and satellite data product. Better title would have been "Data selection and pre-processing" or something similar. 70% cloud radiance fraction threshold is higher than many previous studies use. Since cloudy observations have larger errors, inversion is more error prone with higher threshold. Criterion for OmF is very subjective. Why choose the range [-5, 10] e15? Why not [-5, 5] e15? Why not relative value rather than absolute value? Why not percentile range? Is the selected range applicable to any region or just over China?

- Although some of our authors have a thorough understanding of the DOMINO algorithm, since they are involved in the development of DOMINO for many years, we think discussing the performance/flaws of the DOMINO algorithm and the related findings of the Lin et al. is out of scope of this paper and not relevant for our results. We only conclude, based on the data assimilation results, that a very small fraction of the DOMINO NO_2 data is unrealistic and that these observations always occur in the presence of high aerosol concentrations.

- We agree that the title is not entirely correct and we changed the title of the section into "Quality control of satellite data"

- Although a limit of 70 % for the cloud fraction on average increases the error of the measurements by less than 20%, this is compensated by the 37% increase in the number of observations.

- Although the exact limit to filter outliers in OmF values is always arbitrary, we give a detailed description why we choose these limit values (page 6349, line 13 - page 6350, line 18). The asymmetric limits are chosen because the OmF distribution is asymmetric in its tails, caused by the fact that there are no negative NO_2 observations, but there is virtually no limit to the maximum NO_2. To use a relative factor will filter a lot of data for the low NO_2 concentrations where we have found no problems with the observations. In the end, we have been deliberately on the cautious side with our choice of this criterion and less than 3% of the data is filtered over China. The limit is also applicable for other regions.

In our discussion we added on page 6349,

On line 11: "….new power plants. <u>Not losing sensitivity to new emission sources is also the reason we do not choose a relative filter criterion</u>. We select…."

On line 12: "We select a OmF filter criterion in the range of [-5, 10] x 10^{15} molecules cm^{-2} <u>based on our analysis discussed below</u>."

On line 14: "….. September 2014is Gaussian <u>except for its tails and 97% of ...."</u>

3) Use of surface data: I do not understand the logic of using surface data. There is a big unknown about the quality of the surface NO_2 data they use for validation of the model results. How the (comparison) exercise is going to be insightful if the accuracy of the data used in the analysis is unknown? In addition, the comparison of NO_2 at a surface site with model results at 0.25x0.25 is not really helpful.

This is exactly what we argue in the manuscript: the validation by comparison with surface data is of very limited value. Therefore, we use only hourly values averaged over a long time period of 8 months for the comparison. Since the surface data is available, we did not want to ignore this source of information for our model validation. For the sake of completeness we feel obliged to present our conclusions on the comparison with ground data.

4) Data analysis: Based on information presented in Table 2, it is more logical to focus the analysis for May-September period examining how each regulation was effective in reducing pollution level. From Figure 9, it is difficult to link the changes observed in the YOG period to regulations in place as the results are similar for 2013 and 2014. Authors state in introduction that derived emission is better to study the effectiveness of the air quality measures, but it is unclear to me how satellite-derived emission is better than satellite observations themselves as the model is not providing any additional information regarding regulations. In fact, one might introduce model errors in the inferred emissions. For the nature of the work presented in the paper, I do not see much advantage of the chosen approach.

As argued here and in previous papers (e.g. Mijling et al., 2009) this analysis based on satellite concentrations only is not sufficient, since meteorological conditions and transport of polluted air are strongly affecting the concentrations. This means that there is no linear relation between air quality regulations and NO_2 concentrations. On the other hand the relation between regulations and emissions is very direct and linear, which not only justifies our approach but it is also an <u>improvement</u> on the analysis of measured NO_2 concentrations.

The emission estimates use not only satellite observations in the location of the YOG but use all observations over China that are transported from and to Nanjing. Besides taking transport into account the meteorological effect on the lifetime of NO_2 is taken into account.

Indeed, this is, as pointed out by the referee, shown by the fact that Figure 9 is less convincing that the results in Figure 10. Figure 9 shows only concentrations, while Figure 10 is showing the emissions taken into account transport and meteorological conditions leading to actually <u>smaller</u> errors on the results.

At the end of line 5 page 6352, we add:

"The emission estimates use not only satellite observations in the location of the YOG but use all observations over China that are transported from and to Nanjing. Besides taking transport into account the meteorological effect on the lifetime of NO_2 is taken into account."

5) Several statements in the "Model improvement" section require citations. Please, use NO2 columns consistently instead of NO2 concentrations and NOx emissions instead of NO2 emissions.

The Chimere model is described in section 2.1 with references to Schmidt, 2001; Bessagnet et al., 2004; and Menut et al., 2013.

We have made the use of “NO_2 columns” and “NO_x emissions” more consistent throughout the text.