Response to Referee #3 (ACPD-15-C10942-2015)

The manuscript investigates the results of across variable NOx emissions adjustment in an EnKF surface ozone data assimilation on NO_2 forecasts in Beijing and surrounding areas during the 2008 Summer Olympics. The main finding is that the assimilation of ozone data improved the NO_2 estimates during night and early morning but led to a significant deterioration during daytime over some urban sites, compared to surface measurements. The authors provide a possible explanation of this mixed effect by running and analyzing an idealized data assimilation experiment in which a similar effect is a result of a strong nonlinearity in the daytime NOx-O_3 chemistry combined with the presence of bias in the assumed model emissions.

The following is my take on the potential importance of this study. The theory of data assimilation makes a number of assumptions regarding linearity (although not necessarily in the case of EnKF and probability distributions but these are not always satisfied in reality. The question is how far can we push the limits? For example, typically we assume that observations and backgrounds are unbiased while they really are and assimilation still works. In this case it is important to know how much bias is too much or to what extent the assumptions can be violated without the results breaking down. As I understand it, the present study attempts to answer this question for a particular (and very important case of air quality estimation. I really like the idealized data assimilation experiment: I think this part of the analysis is quite convincing (if lacking some minor details, although it is less clear how it relates to the real data assimilation experiment (see my general comments 2 and 3. I also like the overall logic of the presentation. However, I do have a number of critical comments and suggestions, some more serious than others. I recommend the manuscript for publication after these are addressed.

Response: We very much appreciate the reviewer's valuable comments! The reviewer's comments play a very important role in improving the manuscript. We have revised the manuscript accordingly. A point-by-point response to the review's comments is as follows.

General comments

1. The manuscript fits the criteria for a technical note. I'm not sure if it really qualifies as a research article. I would suggest publishing it as a technical note.

Response: Thanks for this comment. This manuscript highlights a potential scientific issue in linkage with emission bias, data assimilation and air quality forecast. This systematically calls for a scientific debate on bias reduction in data assimilation process and further improvement of existing method. The manuscript therefore aims at contributing to the scientific progress and publishing in the form of a research

article. Nevertheless, we also do not mind publishing it as a technical note suggested by the reviewer.

2. The study decisively attributes the mixed effects of ozone data assimilation on forecast NO_2 to nonlinearities in the model based solely on an idealized experiment done with a very different and much simplified model. I think all we can say is that the idealized experiment offers a possible explanation. Given the simplified nature of the experiment there may be other factors that influence the results of the real data assimilation run, for example transport, which is not included in the idealized case.

Response: Thanks for raising this issue. Model errors from other processes (e.g., transport) are a key issue for the DA experiment and may affect the results of the real data assimilation. Following your comments, we have conducted additional idealized experiments to investigate the influences from the errors of other processes. Because it is quite difficult to simulate the transport process in the box model, we investigated the influences from the errors of the NO_2 photolysis rates that are found to be the top five uncertainty sources of ozone modeling over Beijing and surrounding areas during the Beijing Olympic Games (Tang et al., 2010).

In order to investigate the DA performance of adjusting NOx emissions under the presence of the biases of other factors, we assumed that the NO_2 photolysis rate was overestimated by 20% in the idealized box modeling. Firstly, we were blind to the bias of the simulated NO_2 photolysis rate without any perturbations on it in the DA experiment. The NOx emission was adjusted in the same as the idealized DA experiments described by the previous manuscript. Fig. 1a displays the results of the DA experiment under the error scenario of 30% overestimation in the priori NOx emission estimation. As can be seen, the DA corrected the NOx emission too much and leaded to an underestimation of the emission. This over-correction of the NOx emission by the DA was associated with the bias of the simulated NO_2 photolysis rate.

Therefore, in the second experiment, we took the uncertainty of the simulated NO_2 photolysis rate into account and perturbed the NO_2 photolysis rate in the DA. The error scenario was the same as in the first experiment. Under this situation, the DA performed better than that of the first experiment, and the over-correction of the NOx emission was not observed. The results of the above experiments suggest that taking the model errors into account is very important for the DA and missing the model errors can lead to over-correction to the state variable. In order to deal with this issue, the simulated NO_2 photolysis rates and vertical diffusion coefficients, as the key uncertainty sources of the O_3 modeling, were perturbed to account their uncertainties in the real DA experiment.

The third DA experiment is quite similar to the second experiment, but we increased

the bias of the priori NOx emission estimation to 100% overestimation. The results are shown in Fig. 1c. Under the presence of the large bias in the priori NOx emission, the DA deteriorated the estimation of the NOx emission. Compared with the result of the second experiment, this result suggests that the negative DA impact still exists even taking the model errors into account during the DA.

In short, in sight of considering the influence of the model errors (influence of other factors such as transport) in the real assimilation run, the limitations of the DA method in dealing with the large bias of a highly nonlinear system are still persistent. We have incorporated the above results into the revised manuscript to investigate this issue.

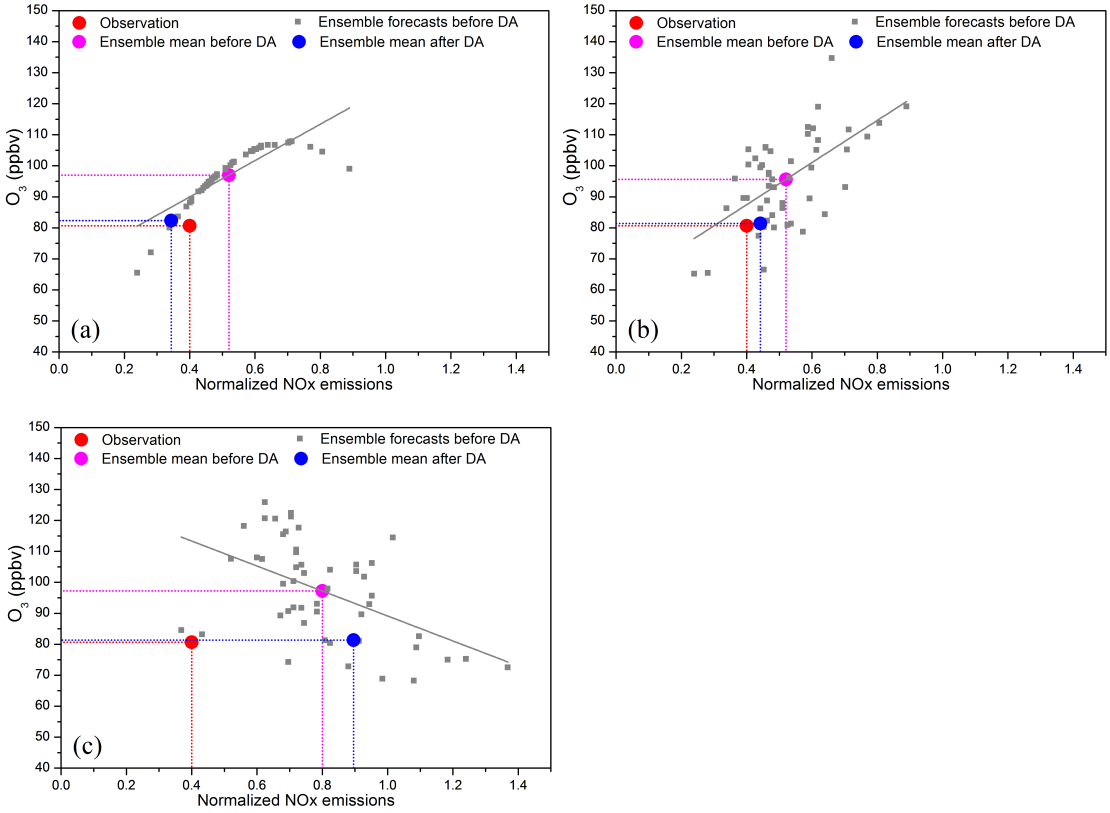
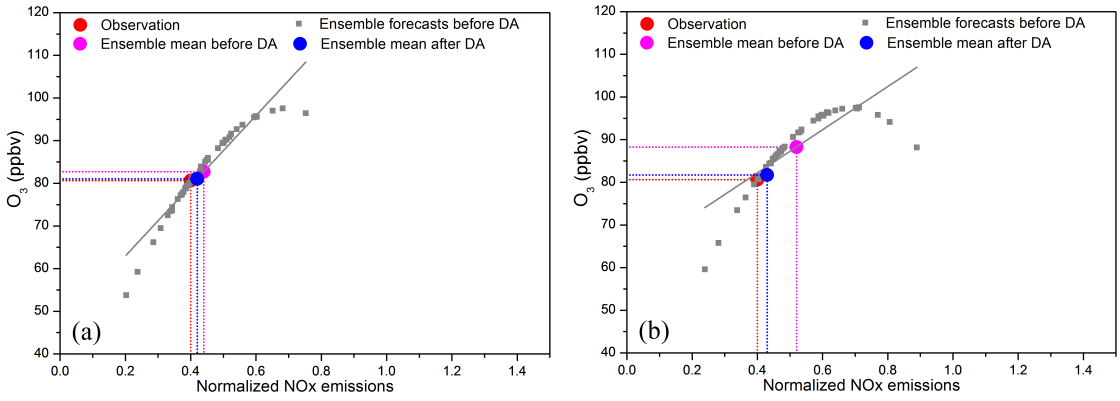


Figure 1 (a-c) O_3 concentrations (ppbv) and NOx emissions (no unit, normalized by the true NOx emission) before and after data assimilation (DA) and their ensemble samples before DA at 12:00 LT on August 12, 2008 in the three ideal DA experiments. The NO_2 photolysis rate is assumed to being overestimated by 20%. (a) The prior NOx emission is overestimated by 30% and adjusted by the DA. The uncertainty of the NO_2 photolysis rate is missed (without perturbations on it) in the DA. (b) The same as the DA in (a), but the uncertainty of the NO_2 photolysis rate is taken into account through perturbing it. (c) The same as the DA in (b), but the bias in the prior NOx emission increase to 100%. The magenta dot represents the ensemble mean of the O_3 concentrations and NOx emissions before DA, and the gray squares denote the ensemble forecasts of O_3 concentrations corresponding to

the perturbations of the NOx emissions. The gray line represents a linear relationship calculated from the ensemble samples of O_3 concentrations and NOx emissions. The red dot represents the true state of NOx emission and the observed O_3 concentration. The ensemble mean of the O_3 concentration and NOx emission after DA are denoted by the blue dot.

3. I don't understand why all three idealized simulations are run with error scenarios in which the NOx emissions are underestimated compared to the truth. Is it expected to be the case for the real data assimilation experiment? Since the latter uses INTEX-B 2006 emissions I would rather expect them to be higher relative to the period of assimilation as, presumably, the air was less polluted during the Olympics than it was in 2006 (e.g. Wang etal. 2009, there maybe more suitable references. Possibly, I've misunderstood something.

Response: Thanks for raising this issue. In the real case for the free run of model, the NO_2 concentrations were overestimated at most of the urban stations but were underestimated at some of the urban stations. In the previous manuscript, we mainly considered the error scenarios for the underestimations of the NOx emissions in the three idealized simulations. However, following your comment, we have performed four new idealized DA experiments in which the NOx emission was assumed to being overestimated by 10%, 30%, 50% and 100% respectively. The results were shown in Fig. 2(a-d). In the first three experiments with 10%, 30% and 50% overestimations in the priori estimation of the NOx emission, the DA worked well and significantly reduced the biases of the emission estimation. In the fourth experiment with the largest bias in the priori emission estimation, the DA increased the bias of the emission estimation during the daytime. These mixed DA effects under different biases of the priori emission estimation were similar to those observed in the previous idealized experiments conducted under underestimate scenarios. In clear, both underestimate and overestimate scenarios confirm the mixed effects of the DA. The present new experiments are added into the revised manuscript.



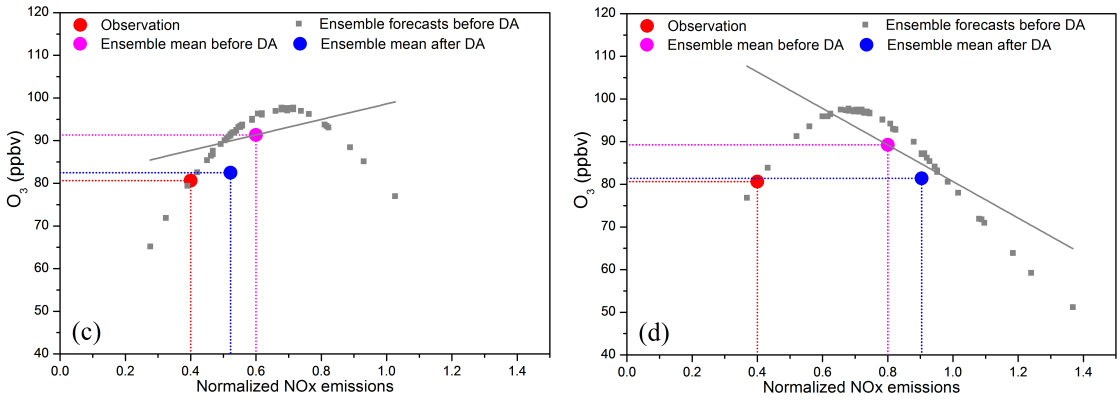


Figure 2 (a-d) O_3 concentrations (ppbv) and NOx emissions (no unit, normalized by the true NOx emission) before and after data assimilation (DA) and their ensemble samples before DA at 12:00 LT on August 12, 2008 in the four idealized DA experiments. (a) DA experiment with 10% overestimation in the priori NOx emission estimation; (b) DA experiment with 30% overestimation in the priori NOx emission estimation; (c) DA experiment with 50% overestimation in the priori NOx emission estimation; (d) DA experiment with 100% overestimation in the priori NOx emission estimation. The magenta dot, the gray squares, the gray line, the red dot and the blue dot represent the same as in Fig. 1.

4. The authors focus on nonlinearity as the sole cause of the mixed results but the idealized experiment simply that it is the presence of a bias in the NOx emissions which leads to problems in a strongly nonlinear model. So it seems that the main culprit here is there action of the nonlinear system to the bias, not the nonlinearity by itself. Isn't EnKF supposed to work well with highly nonlinear systems? This point is important for conclusions and recommendations stemming from the study: in the real world cases, where nonlinearity may be hard to avoid, bias correction is essential.

Response: Thanks for this important comment. We agree with you. Your suggestions are very good for summarizing the main results of this study. Therefore, we have revised the abstract, the conclusions and the other related contents in the revised manuscript.

Revisions in the abstract: "… ***The mixed effects observed in the cross-variable DA, i.e., positive DA impacts on NO_2 forecast over some urban sites, negative DA impacts over the other urban sites and weak DA impacts over suburban sites, were found to be strongly associated with the limitations of the EnKF in a strong nonlinear system. When the uncertainties of the daytime ozone were strongly nonlinearly related to those of the NOx emissions and large biases existed in the priori estimation of the NOx emission, the EnKF may bring out an inefficient or a wrong adjustment to the NOx emissions during the daytime. The results of this study highlight the DA problem under***

***the presence of a large bias in a highly nonlinear system. It implies that bias correction is essential in the application of the EnKF for dealing with the DA problem in a strong nonlinear system.*"**

Revisions in the conclusions: "… ***The results of the DA experiments highlighted a mixed effect of the cross-variable DA using the EnKF. The DA worked properly in improving the NO_2 forecasts and optimizing NOx emissions during the night and the morning when the uncertainties of the O_3 modeling are almost linearly related to those of the NOx emissions. During daytime, the DA resulted in positive impacts on NO_2 forecasts over some urban sites, negative impacts over the other urban sites and weak impacts over suburban sites. Through the idealized DA experiments, the mixed effect was found to be strongly associated with the difficulty in dealing with the highly nonlinear DA problem especially under the presence of large model biases. The results revealed a critical limitation of the EnKF in the chemical DA despite its strong performance for improving ozone forecasts (e.g., Tang et al., 2011).***

***The results suggest that bias correction should be very important in the application of the EnKF for the highly nonlinear chemical DA problem. Alternatively, avoiding the cross-variable DA between two strong-nonlinearly related variables such as the NOx emissions and O_3 is also a possible way to overcome this problem. For example, assimilating the NO_2 observations directly in optimizing the NOx emissions might be better than assimilating the O_3 observations in improving the NO_2 forecasts and the NOx emission estimations. Nevertheless, how to deal with the strong nonlinear problem remains a challenge in the chemical DA. Assimilation approaches that enable dealing with high nonlinear problems in both model evolution and analysis step are needed. Particle filters as a nonlinear filter method (e.g., Moral et al., 1996; van Leeuwen, 2009; 2010) might have potential in this field if its limitation in application for high dimensional system (Stordal et al., 2011) can be overcome.*"**

5. The use of English could use some polishing but I'm not going to focus on this aspect.

Response：Thanks. We have improved the English writing in the revised manuscript.

Specific comments & technical corrections

P35696 L11 ‘indicates gaps’-indicate that gaps

Response: We have revised this as suggested in the revised manuscript.

P35696 L13 ‘calls’-call

Response: We have revised this as suggested in the revised manuscript.

P35698 L8.‘The simplicity in...’ I’m not sure what this sentence means

Response: We have revised this sentence in the revised manuscript. “*EnKF can directly calculate the background error covariance from the ensemble forecasts of the full non-linear model, which is very suitable for data assimilation in complex high-dimensional models (Carmichael et al., 2008).*”

P35698 L10.‘Its implementation is very simple...’ This sentence needs to be edited for grammar

Response: We have revised this sentence in the revised manuscript. “*Its implementation is very simple and does not need an adjoint model which is a very cumbersome task for complex high-dimensional model.* ”

P35699 L21. 60 sounds like a lot! I would like to see a more quantitative justification for that number. Also, ‘the changes of emissions mover Beijing (...) during the (...) Olympic Games ’ are likely to be systematic, i.e. the assumed INTEX-B estimates are probably biased (high) compared to the situation in 2008.

Response: Thanks for this comment. We have added new reference information to justify the estimation of the NOx emission uncertainties in the revised manuscript. “*The uncertainty estimation of NOx emissions used for the modeling during the Beijing Olympic Games is a hard task. The INTEX-B Asia inventory used is estimated to contain 31% uncertainty in NOx emission estimation. But the base year of this inventory is 2006. Another key factor affecting the emission uncertainty is the temporary air pollution control measures during the Beijing Olympic Games. The control measures are estimated to reduce the NOx emissions by 36% to 47% in the studies of Wang et al. (2009; 2010). This would induce large biases into the emission inventory and lead to significant increase of the uncertainties of the emission inventory. Therefore, we estimated the uncertainty of the NOx emissions to be 60 % of the first guess emission rates, about twice the uncertainty in the INTEX-B Asia inventory.*”

P35700 top of the page. Do the perturbations have zero mean?

Response: We have clarified this in the revised manuscript. “*Based on the method suggested by Evensen (1994), the perturbations of the variables in three dimensions are implemented through adding a pseudo smooth random field. The random samples are*

***Gaussian distributed with zero mean.*"**

P35701 Eq(7). Shouldn't U be U', consistent with the notation used in Eqs. (4) and(5) ?
Response: We have revised this in the revised manuscript.

$$"\boldsymbol{U^a(i) = U'(i)}+\mathbf{K}(\boldsymbol{y'(i) - HU'(i)}), \boldsymbol{i} = \mathbf{1, 2, \dots, N} \qquad (7)"$$

P35701 L20. I assume the ensemble mean ($U^a(i)$ averaged over i=1,..., N) is then used as the output analysis state for comparisons (e.g. the blue dots in Figures 4 and 5). Can you clarify this?
Response: Thanks. We have clarified this in the revised manuscript. "*The ensemble mean of $U^a(i)$ is taken as the best estimation after assimilating observations and used as the output analysis state for comparisons (e.g. the blue dots in Figures 4 and 5).*"

P35702 L7. So surface ozone observations are assimilated every hour, correct?
Response:
Thanks. We have clarified this in the revised manuscript. "*The network provides observations of O_3 and NO_2 at the same temporal resolution as the model, i.e., 1 h. The O_3 observations are hourly assimilated into the model to adjust NOx emissions.*"

P35703 L5. Here, 'forecast' is the mean of the ensemble of forecasts, correct?
Response: Yes. We have clarified this in the revised manuscript. "*Figure 2 compares the root mean square errors (RMSEs) of the 1 h ensemble mean forecast of NO_2 at the 17 stations in the RDA experiment with the RMSEs in the NonDA experiment.*"

P35703 L5. How many observation forecast differences went into each RMSE? I'm getting ~14 *24 = 336 observations per location. Please provide these numbers here and in the caption of Figure 2. Would the result be different if, say, only the second week of assimilation was used in the RMSE computations, allowing assimilation to spin up? Are the reported differences between the RMSEs at different stations statistically significant?
Response: Thanks for this comment. The observations used for the RMSE's calculation were a little different at different stations, because some observations were removed due to the quality control process for the data. We have listed the number of the observations used for each station in the revised manuscript. "*The RMSE of each site is calculated based on the hourly differences between NO_2 observation and the ensemble mean forecast of NO_2 from 00:00 LT 9 August to 00:00 LT 23 August in 2008. The number of the valid observations used for each station is also listed in Figure 2.*"
In order to investigate the sensitivity of the result to the period of the calculation, we

did similar comparisons as in Figure 2 of the previous manuscript but focused on the first week and the second week independently. Figure 3a displays the result for the first week and Fig. 3b shows the results of the second week. Although the values of the RMSEs at the stations during the first week are different from those during the second week, the main result is similar during the two periods. The DA increased the RMSEs of the NO_2 forecast over the stations of TJ, BY, IAP, YF and CP, while it reduced the RMSEs over the stations of TS, PEK, SJZ, QHD and CZ. This result is also very similar to that shown in Figure 2 of the previous manuscript. Therefore, the figures for the two periods were skipped in the revised manuscript and a sentence was added into the revised manuscript to clarify this issue. "*The DA impacts during the first week (from 00:00 LT 9 August to 00:00 LT 14 August in 2008) is quite similar to those during the second week (from 00:00 LT 15 August to 00:00 LT 23 August in 2008), which suggests the mixed DA effects is relatively stable during the Beijing Olympic Games.*"

We have checked the significance of the differences between the RMSEs at different stations and incorporated the information into the revised manuscript. "*The differences of the RMSEs before and after DA are statistically significant over 11 (TJ, BY, YF, IAP, CP, XH, CZ, PEK, QHD, SJZ and TS) of 17 stations at the 95% level of the t-test, while there are no statistically significant differences of the RMSEs before and after DA over 6 (XL, YuF, YJ, YLD, LF and BD) of 17 stations. For different types of the stations, the DA impacts on the NO_2 forecast vary substantially from the suburban to the urban stations. Over the suburban sites, the DA shows minor influence on NO_2 forecasts and has no statistically significant impacts on the RMSEs over 5 of the 6 suburban sites.*"

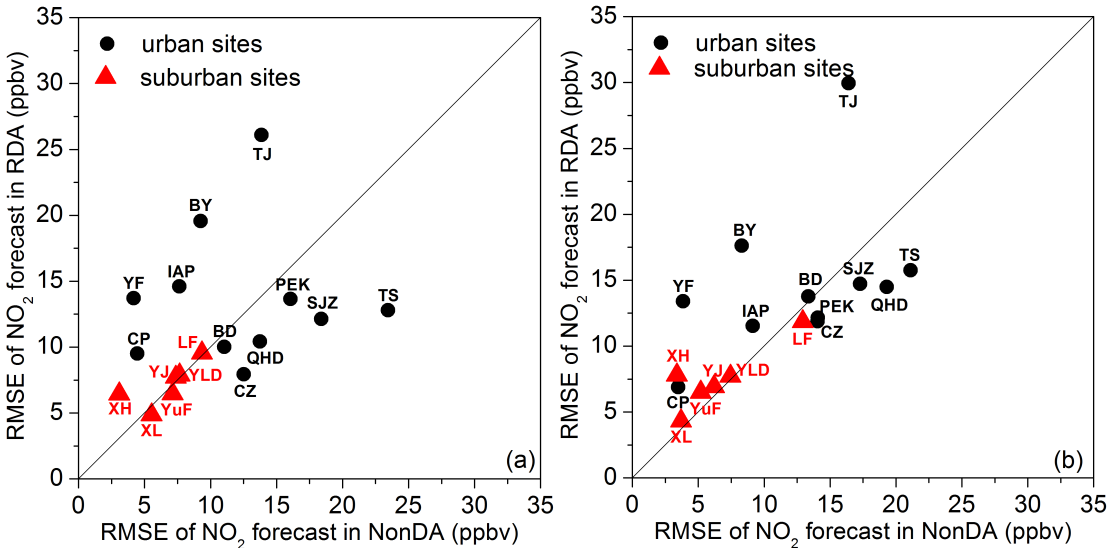


Figure 3. Comparison of the root mean square errors (RMSEs) (ppbv) of 1 h NO2 forecasts at the 17 stations of Beijing and its surrounding areas in the real data

assimilation (RDA) experiments and those in the reference (NonDA) experiment with a free run of the model (a) during the period of 00:00 LT 9 August to 00:00 LT 23 August in 2008 and (b) during the period of 00:00 LT 9 August to 00:00 LT 23 August in 2008. The comparisons at urban sites are denoted by the dots and those over suburban stations are represented by the triangles. The abbreviations of the station names are displayed close to the marks.

P35703.Was the RMSE dominated by a bias or random error? If it's a bias then is it low or high?

Response: Thanks for this comment. The RMSEs of the NO_2 forecasts in the free run of model are dominated by the biases which account for 55~90% (Bias/RMSE) of the RMSEs. The simulations over the urban sites suffered relatively larger biases than those over the suburban sites. The free run of model overestimated the NO_2 concentrations at most of the urban stations, while it underestimated the NO_2 concentrations over most of the suburban sites. We have added the above information into the revised manuscript. "… ***The RMSEs of the NO_2 forecasts in the free run of model are dominated by the biases which account for 55~90% (Bias/RMSE) of the RMSEs. The simulations over the urban sites suffered relatively larger biases than those over the suburban sites. The free run of model overestimated the NO_2 concentrations at most of the urban stations, while it underestimated the NO_2 concentrations over most of the suburban sites.*** **"**

P35705 L2. I wouldn't call it 'in-depth analysis'. The expression suggests analyzing every detail of the problem. What is really done here is one possible explanation of the results using a much idealized experiment.

Response: We have revised this in the revised manuscript. "*Ideal experiment with known true state provides a simply way to investigate the potential consequences of some key inspected factors in a very complex system. In order to investigate the cause of the mixed effects observed in the RDA experiment, this study employed a simplified box model to performed several ideal data assimilation (IDA) experiments in which the true state of ozone concentrations and NOx emissions are assumed to be as known.*"

P35705. Do I understand correctly that the IDA experiment is just a single analysis step with a single ozone observation? Was the box model forecast run for 1hour or longer? Please, clarify.

Response: We have clarified this in the revised manuscript. "*The ensemble runs of the box model were initialized by the ensemble forecasts of the chemical species of NAQPMS at 19:00 LT on 11 August 2008, and the NOx emissions were perturbed to*

***provide ensemble samples of emissions during the following ensemble runs of the model. At 12:00 LT on 12 August 2008, the artificial O_3 observation was assimilated into the box model to adjust the NOx emissions.*"**

Figure 4. Is the magenta dot the result of averaging the grey dots? Is 'before DA' the same as 'forecast'?

Response: We have clarified this in the revised manuscript. "*The grey squares denote the ensemble forecast O_3 concentrations (ensemble forecasts before DA) corresponding to the perturbations of the NOx emissions, and the magenta dot represents the result of the ensemble mean of the grey squares (ensemble mean forecast before DA).* "

P35709 L7. '...due to the needs of linearization at the analysis step, the assimilation should avoid the linearization...'. If DA requires linearization how can it avoid it? I think what the authors mean is that one should avoid problems in which very strong nonlinearities exist (as explained a few lines below. But then how does it jibe with the usual wisdom that the EnKF methodology works well for nonlinear problems? This sentence should be rephrased or dropped.

Response: We agree with you. We have removed this sentence in the revised manuscript

Conclusions. Based on this analysis is it seems that the problem is the presence of a large bias in a highly nonlinear system.

Response: Thanks. We agree with you. Please see our response to the general comment 4.

References

Tang, X., Wang, Z. F., Zhu, J., Wu, Q. Z., and Gbaguidi, A.: Preliminary application of Monte Carlo uncertainty analysis in ozone simulation, Clim. Environ. Res., 15, 541–550, 2010 (in Chinese).

Wang, S.X., Zhao, M., Xing, J., Wu, Y., Zhou, Y., Lei, Y., He, K.B., Fu, L.X., Hao, J.M.: Quantifying the Air Pollutants Emission Reduction during the 2008 Olympic Games in Beijing, Environ. Sci. Technol., 44 (7), 2490–2496, 2010.

Wang, Y., Hao, J., McElroy, M. B., Munger, J. W., Ma, H., Chen, D., Nielsen, C. P.: Ozone air quality during the 2008 Beijing Olympics: effectiveness of emission restrictions, Atmos. Chem. Phys., 9, 5237–5251, 2009.