Referee report on the manuscript *Limitations of ozone data assimilation with adjustment of NO_x emissions: mixed effects on NO_2 forecast over Beijing and surrounding areas*

The manuscript investigates the results of a cross-variable NO_x 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 NO_x-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.

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.
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.
3. I don't understand why all three idealized simulations are run with error scenarios in which the NO_x 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 et al. 2009, there may be more suitable references). Possibly, I've misunderstood something.

4. The authors focus on nonlinearity as the sole cause of the mixed results but the idealized experiments imply that it is the presence of a bias in the NO_x emissions which leads to problems in a strongly nonlinear model. So it seems that the main culprit here is the reaction 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.
5. The use of English could use some polishing but I'm not going to focus on this aspect.

Specific comments & technical corrections

P35696 L11 'indicates gaps' → indicate that gaps

P35696 L13 'calls' → call

P35698 L8. 'The simplicity in…' I'm not sure what this sentence means

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

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.

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

P35701 Eq (7). Shouldn't U be U', consistent with the notation used in Eqs. (4) and (5)?

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?

P35702 L7. So 17-3 = 14 surface ozone observations are assimilated every hour, correct?

P35703 L5. Here, 'forecast' is the mean of the ensemble of forecasts, correct?

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?

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

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.

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 1 hour or longer? Please, clarify.

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

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.

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

References used in this report

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. 2009, 9, 5237– 5251