Response to anonymous referee #3:

We thank referee #3 for constructive and helpful review comments, to which we hope to have responded appropriately. A list of comments including our response is given below.

The topic is very relevant, however I have to criticise the approach used since in the current status important open questions remain.

Let me start by asking the author, once more, to improve the language adopted in the manuscript. There are some fixed points on which the community has agreed upon since many years that simply cannot be ignored. For example the term 'validation' should be dropped for the time being in favour of 'evaluation'. This has been clearly stated in a number of important publications that cannot be neglected. Secondly, one cannot talk about 'validatio'n and than start two sections with "Intercomparison method" and "Intercomparison results". A comparison is between two or more things normally. The suffix 'inter' normally refers to a comparison of elements of the same nature, e.g. model vs model, obs vs obs. If that would not be the case, one would simply talk about "comparison", would he/she not? I think that the natures of observation and model results are already sufficiently different, to complicate further the scene and inferring, with the used of 'intercomparison', that they are not. How about "methodology for the evaluation of the ensemble" and "Results", simple straight forward, clear?

We apologize in case the terms „validation" and „intercomparison" were still used in an inappropriate manner, this was not intended. The corresponding text has been changed as suggested.

I have serious problems with reading the figures. They are excruciatingly small and the number and nature of the differences between models and models/obs is so crucial to the evaluation of the manuscript quality that I cannot precede in a conclusive way.

The most important objection resides in the ensemble treatment and the fact that the differences among the models are confined in the appendix of the paper. The differences among the models qualify the ensemble and define also the quality of your final results. Once more the figures are too small for me to say something definitive here, but from what I can judge I see small differences among models. This puts in question the necessity for an ensemble treatment especially when based on the median which by definition cuts the outliers contribution to the ensemble result and in this particular case may well make redundant the use of several models that are replicating their results. May be they are different, but this is not visible to me from the figures provided.
I do know the value of ensembles of opportunity but the opportunity should be exploited at maximum making sure that there is an added value within the use of multiple models, that the number of models is adequate, not too many not too few and that the contribution from the model results finally used is original and unbiased. This has been demonstrated in a number of works that deserve the attention of the authors.

I think the paper will benefit if the individual relationships among the ensemble members is brought to a higher degree of visibility (not only with larger figures but also conceptually) and analysis. This will increase the scientific significance of this paper which otherwise would look too much like a performance report. The later is useful indeed for the institution/s that use these results but is not at all instructive for the scientific community.

Many changes have been applied to Figures and Tables in order to increase visibility of individual model runs and to enlarge Figures shown in the main part of the manuscript::

-Figures showing non AVK-weighted tropospheric NO_2 VCDs (termed tropospheric NO_2 VCDs from method 1 in previous version) were deleted as these do not differ substantially from AVK-weighted (referred to as method 2 in previous version) values (see p 11 l 19 - p 12 l 2, revised version).

-Scatter density plots and wind directional distributions of surface partial columns have been removed as these were only used in very few sentences of the former manuscript version. Statistical values of surface partial columns which were given along with the scatter density plots in the former manuscript version are now summarized in Table 4 (see below).

-Subfigures showing means over different seasons of vertical profiles, seasonal cycles, diurnal cycles and weekly cycles were moved to the Appendix.

As less subimages are now shown in the main part of the revised manuscript version, this freed up space for remaining ones which are now larger in size and it should now be easier for the reader to concentrate on details. Note also, that the quality of all Figures is good enough to allow zooming into them. This is especially helpful for the Figures in the Appendix containing further results from individual model runs and for different seasons.

In the previous manuscript version, standard deviations calculated based on results from individual ensemble members were used as an indicator of how much individual ensemble members differ from each other and shown along with vertical profiles as well as seasonal, diurnal and weekly cycle Figures (Figure 4, 7, 8 ,9, 10, 11 of the previous manuscript version). In the revised version, standard deviations have been removed from text and Figures which now show individual model runs in addition to the ensemble median instead (see Figure 5, 8, 9, 10, 11 of revised version). Moreover (in response to comments by reviewer #1), three Tables have been added to the main part of the manuscript (further increasing visibility of individual model results in the main part of the manuscript):

-Table 3 shows statistical values of AVK-weighted tropospheric NO_2 VCDs for the four stations for the ensemble and individual model runs

-Table 4 shows the same as Table 3, but for surface partial columns of NO_2

-Table 5 shows the same as Table 3, but for seasonal, diurnal and weekly cycles of AVK-weighted tropospheric NO_2 VCDs

More text on individual model results has been added in several parts of the manuscript, which also points at differences among ensemble members including:

-(p 11 l 14-16, revised version) ”For example, SILAM largely overestimates NO_2 partial columns up to 1.5 km altitude at OHP, while MOCAGE (apart from the lowest observation layer) overestimates values up to about 1 km altitude at Uccle.”

-(p 12 l 14-22, revised version) ” The largest rms and bias (10.5 and 5 x 10^{15} molec cm^{-2}, respectively) are found for LOTOS-EUROS at De Bilt. Considering that values for OHP are generally

smaller than for the three urban sides, SILAM also shows a considerably high rms and bias (2.6 and 1.2 x 10^{15} molec cm^{-2}, respectively) at this station. Vertical profile comparisons described above show that the overestimation mainly occurs at altitudes up to about 1.5 km. Our findings agree with Vira and Sofiev (2015) who found that SILAM tends to overestimate NO_2 at rural sites based on in-situ data and concluded that this is due to an overestimation of the lifetime of NO_2, which is also consistent with findings by Huijnen et al. (2010). For surface partial columns, biases are negligibly small for OHP and Bremen for the ensemble and most of the individual models, while the ensemble is negatively biased by about 1 x 10^{15} molec cm^{-2} at Uccle. The largest rms and bias in surface partial columns are found for EMEP at Uccle (3.3 and -1.8 x 10^{15} molec cm^{-2}, respectively). ”

-(p 13 l 21-26 on seasonal cycles shown by Fig. 8, revised version) “In the present study, the spread between individual models is quite large for OHP indicating that some of the models perform better than others. Looking at the spread between individual models also shows that seasonal cycles are generally more pronounced compared to the other model runs and retrievals for LOTOS-EUROS and MOCAGE. Especially LOTOS-EUROS largely overestimates the observed seasonal cycle at OHP. Low to moderate correlations in seasonal cycles are found for De Bilt, followed by moderate ones for Bremen. All models perform well in terms of correlation at Uccle and OHP (values around 0.8).”

-(p 13 l 27-34, revised version) “Figure 9 shows comparisons of diurnal cycles for the whole time series. Overall, the model ensemble fails to reproduce diurnal cycles for all stations, reflected by generally low correlations (Table 5) for all models at De Bilt, Bremen and OHP. All models show negative correlations at De Bilt, while some of the models only reach negative correlations at Bremen as well. MAX-DOAS retrieved values increase from the morning towards the afternoon, while simulated values in general decrease from the morning towards the afternoon. At Uccle however, high or at least moderate correlations are achieved. CHIMERE performs best in terms of correlation at Uccle and OHP (0.92 and 0.6, respectively). For this model, diurnal scaling factors of traffic emissions have been developed by analyzing measurements of NO_2 in European countries (Menut et al., 2013; Marécal et al., 2015).”

-(p 14 l 8-14, revised version) “The peak at 8 am for Bremen is most pronounced for EMEP-MACCEVA, MOCAGE and LOTOS-EUROS. Individual model runs show the same shape of the diurnal cycle for Bremen, while the shape of diurnal cycles differs for OHP. Moreover, large differences regarding the magnitude of simulated values occur for both stations. As described in Section 2.1, all models use the same emission inventory as a basis, except the EMEP run. There is a strong difference between the magnitude of the values simulated by EMEP and EMEP-MACCEVA specifically for the diurnal cycle at Bremen (while the shape of the cycles is similar), which could be either related to the difference in resolution or different emission inventories incorporated in both of the two runs. ”

-(p 16 l 27-34, revised version) “The largest differences to MAX-DOAS retrieved seasonal and diurnal cycles generally occurred for LOTOS-EUROS and MOCAGE at Bremen and De Bilt and also for EMEP-MACCEVA at Bremen. LOTOS-EUROS and SILAM showed the largest differences to retrieved diurnal and seasonal cycles for the background station OHP. However, weekly cycles are better represented by the model ensemble, which indicates that applied scalings of emissions on a daily basis are at least more appropriate than hourly ones. However, the models generally underestimate the decrease in tropospheric NO_2 VCDs towards the weekend. This decrease was repro-

duced much better by SILAM compared to the other models. The comparisons to MAX-DOAS also showed that this model overestimates values at the background station OHP, in agreement with a study by Vira and Sofiev (2015) who related this to an overestimation of the lifetime of NO_2."

Note also that the abstract has been reformulated in order to reflect the performance of individual models in general.

As results of individual models were moved to the main part of the manuscript, the wording has been changed in some parts of the manuscript in order to be able to differentiate if it is referred to the ensemble or individual model results. Moreover, as standard deviations have been removed in the revised version, it is now referred to "the spread between individual models" instead, e.g.:

-(p 13 l 21-22, revised version) "In the present study, the spread between individual models is quite large for OHP indicating that some of the models perform better than others."

Regarding the use of the model ensemble median, the following text has been added in the revised version (p 9 l 21-30, revised version): "While the calculation of an ensemble median is a common approach to reduce individual model outliers, it is mainly used here for the sake of simplicity and presentation purposes, allowing easier overall evaluation of how the models compare to MAX-DOAS retrievals. The model ensemble is based on five of the seven models (though with partly different set-ups) which constitute the CAMS regional model ensemble (http://www.regional.atmosphere.copernicus.eu/) for which Marécal et al. (2015) have shown that at least for ozone, the ensemble median performs on average best in terms of statistical indicators compared to the seven individual models and that the ensemble is also robust against reducing the ensemble size by one member. Statistical indicators for NO_2 (see Table 3 to 5) show that the ensemble median of the present study performs best in terms of overall correlation to individual MAX-DOAS measurements at each station. Compared to individual models for other statistical indicators and also comparisons for seasonal, diurnal and weekly cycles, reasonable results are achieved by the ensemble median."

What I find contradicting a bit in this paper is also the fact that data are used to validate an ensemble, use nature is obscure, and the main message that this brought forward is indirectly that this exercise demonstrates that Max-Doas data are suitable to validate models. So what is validating what and how?

In the revised version, the corresponding text stating that that this study focuses on evaluating the usefulness of using MAX-DOAS data to improve model performance has been deleted (p 3 l 29-30 former version), as it was partly misleading. Moreover, the term 'validation' has been removed as suggested above. MAX-DOAS retrievals do not constitute direct measurements of NO_2 conditions but base on measurements of light intensity in specific wavelength windows. In this sense, they are closer to NO_2 conditions than simulations. This should be accounted for by a conservative overall uncertainty of MAX-DOAS retrievals of 30 % which is assumed for all stations within this manuscript and given along with the data plots, where appropriate (p 9 l 5-8 of former version, p 9 l 31- p 10 l 3 of revised version).

In the present status the manuscript can not, in my view be published in ACP. GMD would be more suitable, but provided that more insight is given into the ensemble workings.

This work was initially submitted to GMD, where it was regarded as out of the journal‘s scope with prompt recommendation to submit to ACP instead. We believe that results of the present MAX-DOAS based comparison study and differences found between simulations and retrievals are of interest to both modelling and measurement community (therefore fit to the scope of ACP) and hope that this work stimulates future studies on improving model performance.