Response to anonymous referee #1:

We thank referee #1 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 paper presents a comparison of time series of tropospheric NO2 VCDs derived from 4 European MAX-DOAS stations to an ensemble of 5 regional models. The horizontal and vertical resolution of MAX-DOAS observations fits in general well to those of the regional models. Thus such a comparison is well suited to evaluate the performance of the model simulations (and also the quality of the MAX-DOAS retrievals). In this respect, the results of this paper are of high importance, and are well suited for publication in ACP. However, I have three major concerns with respect to the evaluation and presentation of the results in the present version of the manuscript, which should be addressed before final publication:

a) One of the main advantages of MAX-DOAS observations is that profile information for the lowest layers of the atmosphere (below about 2km) can be obtained. Profile information is crucial to assess the performance of the model simulations (and to understand deviations from observations). It is a pity (and completely unclear to me), why the authors do not make explicit use of the profile information derived from MAX-DOAS. One – rather simple – way to make use of the profile information (and to compare MAX-DOAS results and model simulations) would be to determine a characteristic layer height (e.g. the layer, below 70% of the total tropospheric column resides) from both the MAX-DOAS observations and the model results.

In the manuscript, vertical information from MAX-DOAS is made use of by comparing average vertical profiles of simulations and retrievals (Figure 5 and A1 of revised manuscript) and described in the results section (p 11 l 9-18, revised version), demonstrating principle agreement between measured and retrieved profiles. We agree that comparisons of characteristic layer heights may show useful additional information on the ability of the models to reproduce the distribution of NO_2 in the vertical. However, also keeping in mind the number of Figures shown in the manuscript, we consider this as an interesting topic for future studies. The latter has been added to the summary and conclusions section on p 18 l 9-11 (revised version):

”Moreover, one could investigate the ability of the models to distribute NO_2 in the vertical in terms of characteristic layer height of NO_2, which is (in addition to other factors like vertical distribution of emissions or boundary layer schemes) expected to be affected by vertical resolution of the models.“

b) The authors compare the MAX-DOAS results to model ensembles. Although in the appendix, also the comparison results to the individual models are shown, no attempt is made to systematically asses the performance of the individual models with respect to the MAX-DOAS results. The authors should at least provide a table with some key indicators (e.g. correlation coefficient, slope, bias, etc.) for the individual model comparisons. These indicators should be provided for a) the complete time series, b) for the seasonal variation, c) the diurnal variation, and d) the weekly cycle.

A couple of changes have been applied to the text, Figures and Tables of the manuscript in order to put more weight on results of individual models in the main part of the manuscript (this was also asked for by reviewer #2). In the revised version, three Tables have been added:

-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 101^{5} 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 reproduced 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.

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).

Note also that the number of Figures and subimages has been reduced in the new version, which is both a consequence of the new Tables added and the request by reviewer #2 to increase size of the Figures:

-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.

c) The discussion of the deviations between the model simulations and the MAX-DOAS results is weak, and only rather general explanations for the disagreements are given. The paper would benefit a lot if the possible reasons for disagreement would be investigated in more depth. In particular, from the two points mentioned above, useful information could be obtained, which processes (e.g. transport, emission inventories, chemistry) might be most important reason for discrepancies for individual situations and/or model

As described in reply to point b) above, the revised manuscript contains Tables showing overall statistical values for the ensemble and individual model runs and corresponding ones for seasonal, diurnal and weekly cycles. Based on the new Tables and also as part of the response to referee #2, the contribution of seasonal, diurnal and weekly cycles to overall correlations has been investigated. This showed that overall correlations reached at all stations are mainly driven by seasonal and weekly cycles, while significantly lower and in many cases negative correlations are achieved for diurnal cycles which decreases overall correlations. An exception for the latter is Uccle, where good correlations are also found for diurnal cycles. This is now described on p 15 l 22-24 of the revised version.

Moreover, diurnal cycles based on weekdays and based on weekends only have been derived and are now presented and discussed in the revised version (see p 14 l 27 – p 15 l 10, p 16 l 20-27) and a corresponding Figure showing diurnal cycles for weekends only has been added (Figure 10, revised version). Note that results for weekdays only look similar to results based on all days of the week and are therefore not shown in the manuscript. Diurnal cycles based on weekends only in general show a rather flat shape for the urban stations. However, the shape of model simulated diurnal cycles looks very similar for weekdays compared to weekends, meaning that simulations fail to reproduce the observed changes towards the weekend. It should be checked in future studies if switching off diurnal scalings of emissions during weekends leads to an improvement in model performance compared to MAX-DOAS. A note on these results has also been added to the Abstract (p 1 l 14 – p 2 l 2, revised version).

In addition to the MAX-DOAS comparisons shown in the present study, we also carried out a comparison between the regional models and OMI satellite retrievals with similar results as Huijnen et al. (2010). A paragraph on these comparisons has been added on p 17 l 1-13 of the revised version. However, due to the generally short lifetime of NO_2, to properly relate uncertainties in the simulations over emission hotspots indicated by the OMI based comparisons to the ones derived from MAX-DOAS based comparisons would generally require investigating transport patterns of individual model runs with much higher time resolution around the MAX-DOAS sites, which is not provided by the satellite data (only one OMI orbit per day over the stations).

A Figure showing a map of OMI satellite observations and TNO/MACC-II anthropogenic NOx emissions has also been added to the manuscript (Figure 1 in revised version, corresponding text added on p 4 l 1-4). The spatial distribution of NOx emissions agrees well with pollution hotpots and cleaner areas identified by OMI. The latter shows that the spatial distribution of emissions does not seem to be a likely reason for differences between simulations and MAX-DOAS retrievals.

The impact of horizontal model resolution on the ability of the models to reproduce MAX-DOAS results is now discussed in the revised version (p 17 l 19 - p 18 l 9). One would expect that this ability increases with increasing model resolution. However, no clear relation between model resolution and performance of the models resulted from these investigations, which shows that other differ-

ences between the models such as chemistry schemes and treatment of emissions strongly impact on comparison results. (see also reply to minor point on model resolution below)

Additional comparison results described above pointed at more likely (and also less likely) reasons for differences between simulations and observations and hence provided further useful information for future studies to track down reasons of disagreement with the aim to achieve a better agreement between MAX-DOAS and model results. This would mainly involve running models with different model set-ups, emission inventories, resolution, parameterisations and chemistry schemes. The summary and conclusions section has been extended by the results described above and more ideas for future studies are now given.

Huijnen, V., Eskes, H. J., Poupkou, A., Elbern, H., Boersma, K. F., Foret, G., Sofiev, M., Valdebenito, A., Flemming, J., Stein, O., Gross, A., Robertson, L., D'Isidoro, M., Kioutsioukis, I., Friese, E., Amstrup, B., Bergstrom, R., Strunk, A., Vira, J., Zyryanov, D., Maurizi, A., Melas, D., Peuch, V.-H., and Zerefos, C.: Comparison of OMI NO_2 tropospheric columns with an ensemble of global and European regional air quality models, Atmos. Chem. Phys., 10, 3273-3296, doi:10.5194/acp-10-3273-2010, 2010.

Minor points:

Page 1, line 1: Replace NO2 by NOx

Changed to: "Tropospheric NOx ($NO+NO_2$) is hazardous to human health and can lead to tropospheric ozone formation, eutrophication of ecosystems and acid rain production."

Page 1, line 8: 'measurements are available during daylight'. To me it seems that this is not an advantage but rather a disadvantage (measurements are not available during night)

Thanks for pointing this out. More explicitly, the advantage the sentence should have referred to is, that multiple measurements are carried out during daylight, so that e.g. diurnal cycles can be derived from the retrievals. The sentence has been changed to (p 1 l 6-9, revised version):

"Compared to other observational data usually applied for regional model evaluation, MAX-DOAS data is closer to the regional model data in terms of horizontal and vertical resolution and multiple measurements are available during daylight, so that for example diurnal cycles of trace gases can be investigated."

Introduction: It should be made more clear, that the quantity of interest is NOx, but only NO2 can be measured

Added the following sentence (p 3 l 21-22, revised version):

"In contrast to NO_2, NOx cannot be retrieved from MAX-DOAS measurements directly, so that these measurements are of more interest for air quality than for atmospheric chemistry studies."

Page 2, line 30: The statement 'using zenith measurements as intensity of incident radiation' is unclear to me. Do you mean incident solar irradiation? Then I would disagree. Please clarify.

This sentence was misleading and has been rephrased to (p 3 l 1-3, revised version):

" Therefore, using observations in low elevation angles as measurement intensity and zenith measurements as reference intensity, the total amount of molecules of a certain species along the light

path difference (zenith subtracted from non-zenith measurement), so called differential slant column densities, can be determined using Lambert Beer's law."

Section 2.1: What is the spatial resolution of the models? How does it compare to the horizontal sensitivity ranges of the MAX-DOAS results?

In response to this question, the following text has been added to p 17 l 19 - p 18 l 9 of the revised manuscript (this is combined with a response to referee #2 who also asked about the impact of model resolution on comparison results):

"The horizontal grid spacing (Table 1) differs for the 6 model runs evaluated in the present study, with a resolution of approximately 9x7 km^2 for the highest resolution run (LOTOS-EUROS) and 50x50 km^2 for the coarsest one (EMEP). The resolution of the remaining model runs is approximately 20x20 km^2. As described in Section 2.2, the horizontal averaging volume of MAX-DOAS retrievals strongly depends on aerosol loading, viewing direction and wavelength (Richter et al., 2013). As a rough estimate, it ranges from 5 to 10 km for the stations used in the present study. Therefore, the horizontal averaging volume is (apart from the coarsest resolution run) expected to be either on the same spatial scale as the horizontal model resolution or by a factor of 1 to 4 smaller. From the latter (i.e. horizontal averaging volume of MAX-DOAS smaller than model resolution) one would expect an underestimation of enhancements in tropospheric columns observed by MAX-DOAS in case of horizontal changes in tropospheric NO_2 columns below the model resolution and, similarly, an overestimation of local minima in tropospheric NO_2 columns. However, in reality, the comparison between horizontal averaging volume of MAX-DOAS and horizontal resolution of the models is much more complicated, as MAX-DOAS instruments usually measure in one azimuthal pointing direction meaning that measurements are performed only on a specific line of sight whereas model simulations are performed for three dimensional grid boxes. This could for example mean that a pollution plume with a horizontal extent on the order of the model resolution and hence showing up in the simulations is missed by the line of sight of the MAX-DOAS instrument. It would therefore be desirable to perform multiple MAX-DOAS measurements over a range of different azimuthal angles for each station and use these in future model to MAX-DOAS comparison studies.

A pollution plume and related increase in the time series of tropospheric NO_2 VCDs observed by MAX-DOAS would be expected to be reproduced better by model runs with higher horizontal resolution compared to lower resolution runs. The lifetime of NO_2 is also expected to increase with model resolution. However, in the present study, the LOTOS-EUROS run with significantly higher horizontal resolution than the other runs in general did not perform better than lower resolution runs which can probably be explained by its low number of vertical layers. Similarly, the EMEP run with significantly lower horizontal resolution did not perform worse than higher resolution runs, which shows that other differences between the models such as chemistry schemes and treatment of emissions strongly impact on comparison results. It would be interesting to investigate the ability of the models to predict the scales of NO_2 spatial variations derived from time scales of NO_2 variations and wind speeds in the context of model resolution in a future study. "

Richter, A., Godin, S., Gomez, L., Hendrick, F., Hocke, K., Langerock, B., van Roozendael, M., Wagner, T.: Spatial Representativeness of NORS observations, NORS project deliverable, available online at: http://nors.aeronomie.be/projectdir/PDF/D4.4_NORS_SR.pdf, 2013.

Section 2.2: The retrievals are described in an inconsistent and partly incomplete way. For example, for KNMI the retrieval procedure is completely unclear. Was a profile inversion performed or not? This section should be harmonised and completed. The effect of the different inversion procedures on the NO2 results should be briefly discussed.

This section has been harmonized. In the first paragraph, a brief general description of how NO_2 profiles/columns are derived from the measurements is given. For each station, the most important retrieval and measurement site information are then given (such as instrument type, location and pointing direction of instrument, wavelength window of instrument and of the NO_2 DOAS fit, the radiative transfer model used, cross sections of gases included in the fit, how a-priori profiles were derived). Moreover, the retrieval procedure for De Bilt is now described in more detail.

Section 2.2: It is stated that for Uccle, cloud information was retrieved. Was this information also used for the selection of the measurements? What about the retrieval of cloud information for the other stations?

The following text is now given in the last paragraph of Section 2.2 (p 7 l 22-27, revised version):

"For Uccle, information on cloud conditions was retrieved according to the method by Gielen et al. (2014) which is based on analysis of the MAX-DOAS retrievals, but not applied for results shown in the present study. No cloud flags are available for Bremen, De Bilt and OHP. Larger uncertainties are associated with retrievals under cloudy conditions in particular as clouds are not included in the MAX-DOAS forward calculations. However, MAX-DOAS retrievals are usually filtered for patchy cloud situations by comparing radiative forward calculations of O_4 to retrieved O_4 columns and removing cases from the data with larger than expected differences."

Note that the discussion and analysis of the impact of clouds on comparison results has been removed from the results section (as suggested by anonymous referee #2) and regarded as a topic for future studies, which is now mentioned on p 7 l 34 and p 18 l 21 of the revised manuscript.

Gielen, C., Van Roozendael, M., Hendrick, F., Pinardi, G., Vlemmix, T., De Bock, V., De Backer, H., Fayt, C., Hermans, C., Gillotay, D., and Wang, P.: A simple and versatile cloud-screening method for MAX-DOAS retrievals, Atmos. Meas. Tech., 7, 3509-3527, doi:10.5194/amt-7-3509-2014, 2014.

Section 2.3: How does the wind data compare to the wind fields used in the models?

As described in section 2.1, all models use ECMWF-IFS as meteorological input and boundary conditions. As the models are run with differing horizontal and vertical resolution (see Table 1), wind data from the model output is expected to differ among the models. Wind speed and direction was provided as an output parameter for two of the model runs (LOTOS-EUROS and MOCAGE) of the present study. Figure R1 below shows wind directional distributions of wind speeds from the weather station data and the ones from the model output (near surface level) for the four MAX-DOAS stations (note that MOCAGE data is not available for OHP). Figure R2 shows corresponding wind directional distributions of the data percentage in each bin (e.g., a value of 10 for the 0 to 45° wind direction bin means that during 10% of the time period the wind was blowing from north to north-east). Statistical values of the wind speed comparisons were calculated along with the plots. Wind speed correlations are high for De Bilt and Bremen for both models (~0.8) and moderate for Uccle and OHP (~0.5-0.6). Wind speeds are positively biased for the three urban stations, with the largest biases for Uccle (on the order of 3 m/s), while there is a negative bias at OHP (~ -7 m/s). Note that the negative bias may result from the fact that wind speeds and directions from near sur-

face level were taken for the comparisons which should be comparable to measurements at meteorological sites. However, this is probably not representative of winds at the small hill where the OHP station is located (~650 m above mean sea level) since the orography of the IFS model is a smoothed version of the real orography. Thus, IFS simulates wind speeds for a more flat terrain, which are therefore lower than the measured ones.

Not considering the magnitude of values, wind directional distributions of wind speed from the models agree well with the ones from the weather station data for all stations apart from Uccle. For the latter, the model output shows the highest average wind speeds to the west/south-west of the station, while the measurements show the highest ones to the north-east. As for wind speeds, wind directional distributions also agree well in general for the data percentage. Larger differences occur for Uccle for south to south-westerly and west to north-westerly wind directions and for OHP for west to north-westerly winds.

Note that wind directional distributions shown in the manuscript (Figures 7 and A3 of revised version) are (as described in the corresponding Figure captions) based on wind directions from weather station measurements solely. However, due to the generally good agreement between measured and simulated wind speeds and directions described above, this is not expected to have a strong impact on the data analysis and conclusions given in the manuscript. This is demonstrated by Figures R3 and R4 below which show wind directional distributions of tropospheric NO_2 VCDs for (left) LOTOS-EUROS and (right) MOCAGE based on wind directions from measurements only (as in the manuscript) as well as based on measured wind directions for MAX-DOAS retrieved values of NO_2 and based on model output for simulated NO_2 values, respectively. Overall both Figures show a good agreement between measured and simulated wind directional distributions of NO_2.

What about wind data for KNMI?

The following sentence has been added to section 2.3 (p 8 l 10-11, revised version):

“For De Bilt, wind measurements (within 300 m from the MAX-DOAS instrument) carried out by KNMI were downloaded from https://www.knmi.nl/nederland-nu/klimatologie/uurgegevens.”

Page 8, line 22: ‘Only those model values closest to the measurement time are used’. Why is no interpolation in time of neighbouring model output values performed?

This was mainly done to save computation time. As the time difference between simulations and retrievals is shorter than half an hour, interpolation in time is not expected to have a major impact on conclusions of this study.

Page 9, line 10: What is the vertical extension of the lowest measurement layer?

Bremen 50 m, De Bilt 180 m, Uccle 180 m, OHP 150 m above ground. This has been added to p 10 l 29-30 of the revised version.

Page 9, line 12: ‘comparisons of profiles’? No comparison of profiles is shown in Figs. 1 and 2.

This was done in order to explain why surface partial columns are not shown in Figure 2 of previous version (Figure 3 of revised version) for De Bilt. Surface partial columns have been derived for

stations with vertical profile retrievals only. The sentence was however misleading and has been replaced by the following text in the revised version (p 10 l 28-31):

“In the present study, surface partial columns refer to the partial column of the lowest measurement layer (Bremen 50 m, De Bilt 180 m, Uccle 180 m, OHP 150 m above ground). As vertical profiles are not available from the MAX-DOAS output for De Bilt, comparisons of surface partial columns are not given for this station in the present manuscript.”

Page 10, line 5: ‘As the sensitivity of MAX-DOAS retrievals is largest in the boundary layer’ Is this also true for the ‘de Bilt measurements’?

Yes, the sensitivity to NO_2 in the boundary layer is intrinsic for the measurement method. Differences in retrieval methods will not change this. The corresponding sentence has been changed to (p 11 l 19-21, revised version) :

“As the sensitivity of MAX-DOAS retrievals is largest in the boundary layer, a feature which is independent of the retrieval method, we initially expected the application of column AVKs from the measurements to model simulations to be of crucial importance for evaluation results.“

Page 10, lines 23,24: ‘On average, observed NO2 partial columns are higher in the lowest observation layers during cloudy conditions compared to clear-sky conditions’ I guess that no clouds are considered in the MAX-DOAS forward model. How reliable are then the MAX-DOAS NO2 results under cloudy conditions?

As described above, the discussion and analysis of the impact of clouds on comparison results has been removed from the results section (as suggested by anonymous referee #2) and regarded as a topic for future studies (see p 7 l 34 and p 18 l 21of revised manuscript).

Larger uncertainties are associated with retrievals under cloudy conditions in particular as clouds are not included in the MAX-DOAS forward calculations. However, MAX-DOAS retrievals are usually filtered for patchy cloud situations by comparing radiative forward calculations of O_4 to retrieved O_4 columns and removing cases from the data with larger than expected differences. This is now mentioned on p 7 l 24-27 of the revised manuscript.

Page 11, line 3: What is exactly meant with ‘correlation’? r or r squared?

Correlations calculated in this study refer to the pearson correlation coefficient, i.e. r not squared. The latter was mentioned in the caption of Figure 5 only of the previous manuscript version, but is now mentioned in several parts of the revised manuscript (i.e. p 11 l 24, p 12 l 5, caption of Figure 6, caption of Figure A2, caption of Table 3).

Page 11, line 12: How consistent are the wind data from the weather stations with the wind fields used in the models? Can you show a similar plot as Fig. 6 based on the wind fields from the models?

See response to comment on section 2.3 above and corresponding Figures below. Note that this sentence has been changed to (p 12 l 26-28, revised version):

“Figure 7 shows comparisons between MAX-DOAS and the model ensemble of wind directional distributions of average tropospheric NO_2 VCDs based on wind measurements from station data

(note that further analysis has shown a good agreement between measured wind speeds and wind directions and those of the simulations). ”

Page 13, line 15: ‘However, many validation points arise from the MAX-DOAS based comparisons which could improve model performance substantially.’ This sentence is not clear to me. Please clarify.

Although there is good agreement between MAX-DOAS retrievals and model simulations of tropospheric NO_2 in a general sense, differences have been found for example for individual pollution plumes observed by MAX-DOAS, seasonal, weekly and diurnal cycles. The reasons for the differences should be identified in future studies and several aspects of simulations could be changed in order to achieve a better agreement to MAX-DOAS retrievals. The corresponding sentence has been changed, we hope it is now more clear (p 16 l 2-4, revised version):

“However, many points to evaluate arise from the MAX-DOAS based comparisons. Tracking down the reasons for differences between simulations and retrievals and adjusting model runs accordingly (in case of differences caused by errors in simulations rather than uncertainties of the retrievals) could improve model performance substantially.“

Text on how a better agreement to MAX-DOAS (where desirable) could be achieved has been added to section 5 (p 18 l 22-30, revised version):

“To track down reasons for the reported uncertainties of regional model simulations constitutes the main challenge for future studies. This could be achieved by running models with different chemistry schemes combined with different resolutions where possible (uncertainties in chemistry such as lifetime of NO_2), running models with and without scaling of emissions in time and for specific seasons or days only (uncertainties in seasonal, diurnal and weekly cycles related to emissions), performing runs with varying vertical scalings of emissions (uncertainties in injection heights) and carrying out runs with varying boundary layer physics (uncertainties of NO_2 profiles due to mixing of emissions in the boundary layer and transport therein). Especially LOTOS-EUROS and MOCAGE showed large differences to the MAX-DOAS retrieved seasonal and diurnal cycles for Bremen and De Bilt and also EMEP-MACCEVA for Bremen, so that the impact of different set-ups in emissions and chemistry is expected to be more pronounced compared to the other models at these stations.”

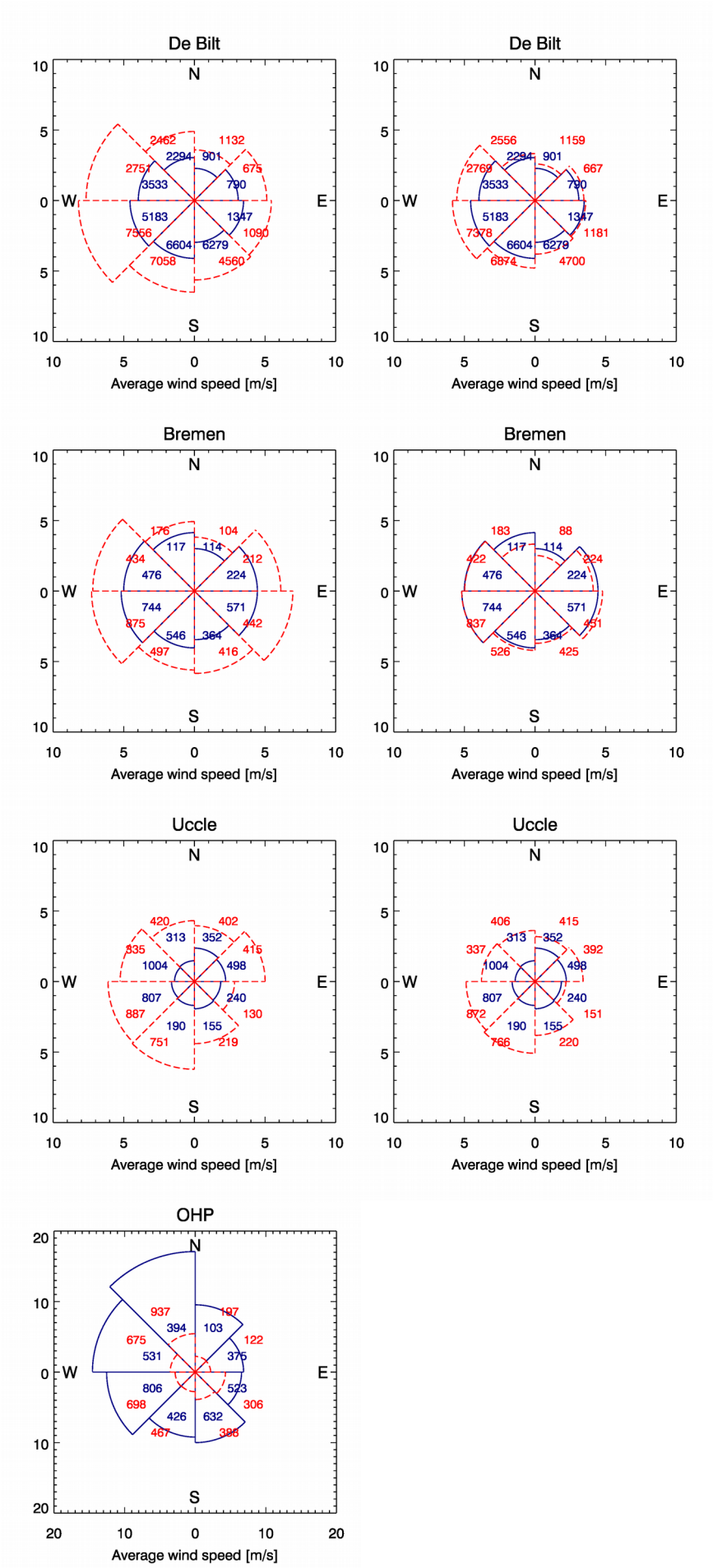

Figure R1: Average wind speed in 45° wide wind direction bins from (blue solid lines) weather station measurements and (red dashed lines) model output for (left) LOTOS-EUROS and (right) MOCAGE for (first row) De Bilt, (second row) Bremen, (third row) Uccle and (bottom row) OHP. Wind directions correspond to the direction towards the station and are taken from weather station measurements itself for measured and from model output for simulated wind speeds. The printed numbers in each bin refer to the number of data values used for calculating average values for each bin.

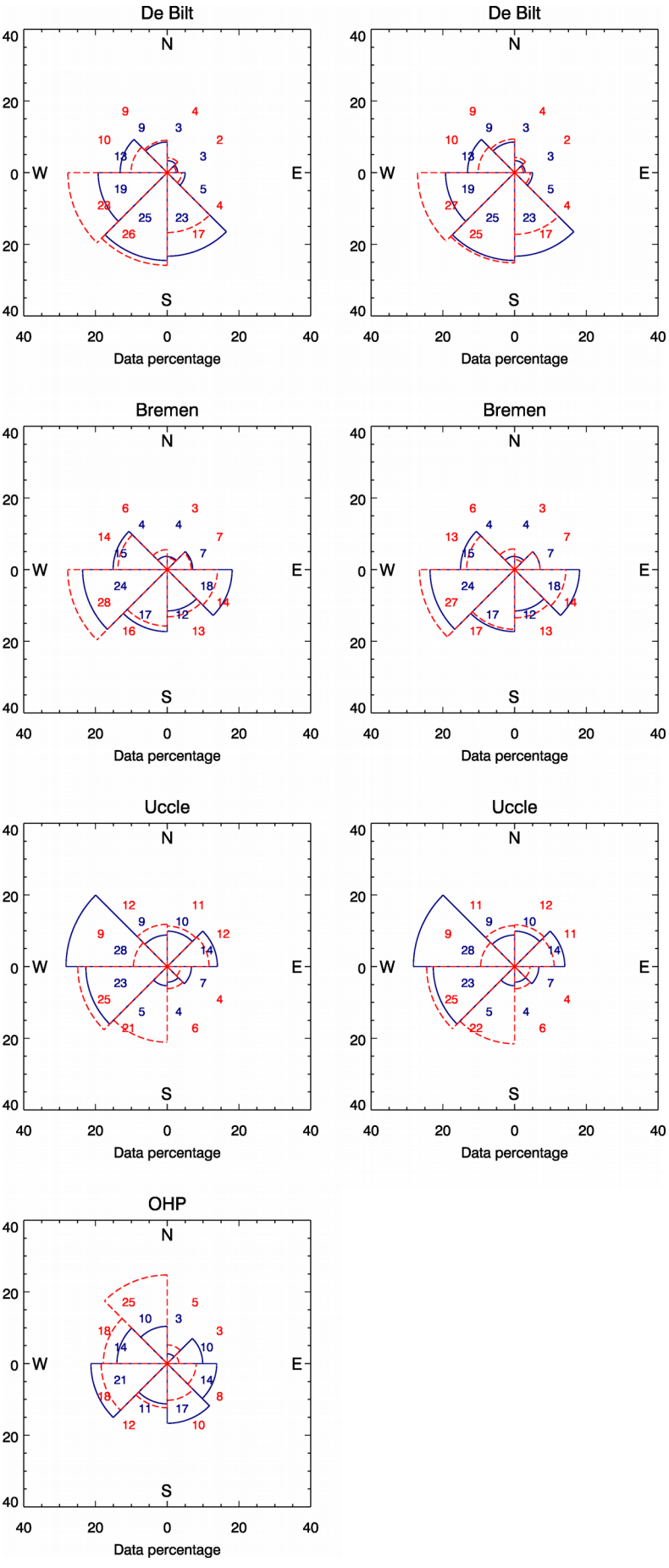

Figure R2: As in Figure R1 but for average percentage of data values. The printed numbers given for each bin were rounded to its closest integer value.

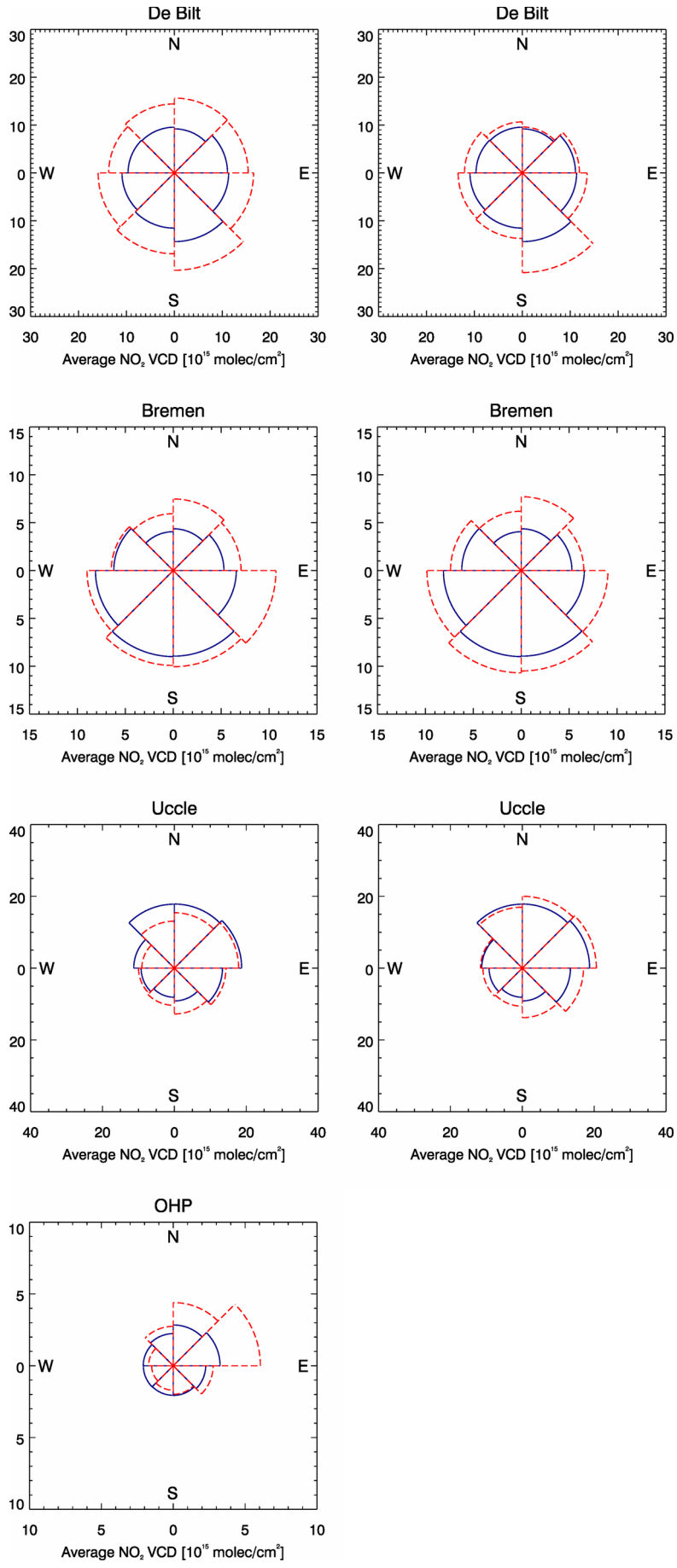

Figure R3: As in Figure R1 but for average AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}]. Wind directions correspond to the direction towards the station and are taken from weather station measurements for both MAX-DOAS retrieved and model simulated values.

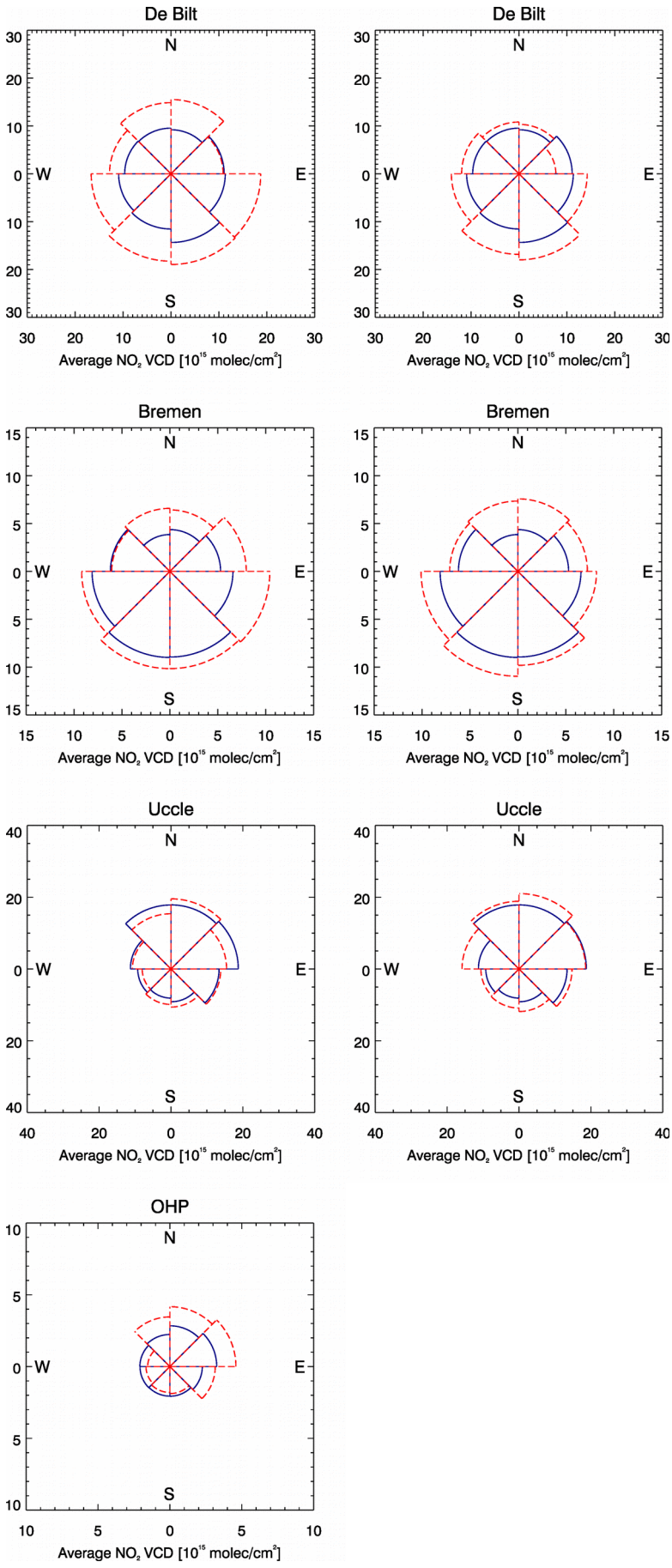

Figure R4: As in Figure R1 but for average AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}]. Wind directions correspond to the direction towards the station and are taken from weather station measurements for MAX-DOAS retrieved and from model output for simulated values.