Dear Dr. Tim Butler,

Many thanks for taking care of the review process of 'Comparison of tropospheric NO_2 columns from MAX-DOAS retrievals and regional air quality model simulations' (MS No.: acp-2016-1003). You can find detailed answers to the reviewer comments on the following pages together with a version of the manuscript in which changes are tracked in the text and references (former text in red, new text in blue font). Please note that changes applied to captions and Figures are not marked as these did not show up correctly in the changes tracked version.

The reviewer comments have been very helpful for further improving the scientific quality of the manuscript and many changes have been applied accordingly. These include:

- More weight is now put on individual model results in the main part of the manuscript (please see replies to referees #1 and #2 for details). For example, individual model results are shown in Figures 5, 8, 9, 10, 11 in the main part of the revised manuscript instead of standard deviations based on all model runs in the previous version. Moreover, statistical values of individual models are now listed in Tables 3 to 5 and results of individual models are described and discussed in several parts of the revised version.

- Subfigures showing means over different seasons of vertical profiles, seasonal cycles, diurnal cycles and weekly cycles were moved to the Appendix. Scatter density and wind directional distribution plots of surface partial columns have been removed as these were not substantial for the manuscript, statistical results on surface partial columns are now summarized in Table 4 instead. Figures showing non AVK-weighted tropospheric NO_2 VCDs were deleted as these do not differ substantially from AVK-weighted ones. This freed up space for Figures in the main part of the manuscript which are now larger in size and it should now be easier for the reader to concentrate on details and also on differences between individual model runs.

- The location of MAX-DOAS sites plotted on top of a mean map of tropospheric NO_2 columns from OMI as well as on top of TNO/MACC-II anthropogenic NOx emissions is now shown in Figure 1 of the revised version.

- Due to the large number of model evaluation points identified by the MAX-DOAS based comparisons, it is beyond the scope of this study to investigate all of these in depth. However, following suggestions of the referees, changes were applied to the manuscript in order to further investigate the differences found where possible and a final synthesis on how to track down reasons for disagreement has been added to the summary and conclusions section. These include investigation of contributions of seasonal, diurnal and weekly cycles to overall correlations, further investigations of the weekend effect, a discussion of model resolution and averaging volume of MAX-DOAS measurements, suggestions for sites to investigate in future studies and a paragraph on OMI satellite comparisons. We hope that the large number of model evaluation points identified by our MAX-DOAS based comparisons stimulate future dedicated studies for improving model performance.

Apart from these substantial changes, the wording has been improved, some typos were corrected, appearance of Tables improved and references have been updated. The terms 'validation' and 'intercomparison' were replaced as suggested by referee #3 and section 2.2 describing MAX-DOAS retrievals has been harmonised. Note that model results from method 1 and 2 are now termed non AVK-weighted and AVK-weighted ones, respectively.

Sincerely,

Anne-Marlene Blechschmidt

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.”

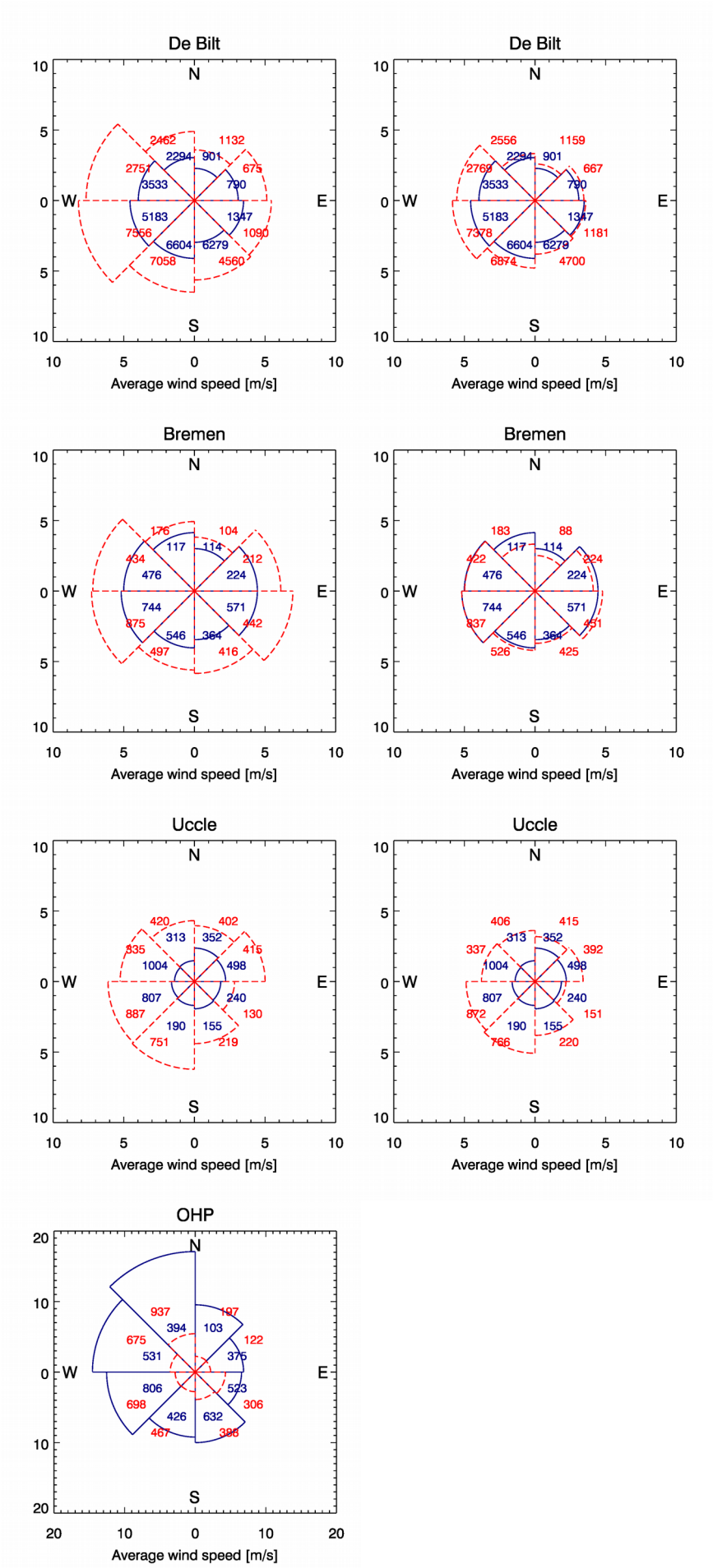

Figure R1: Average wind speed in 45° wide wind direction bins from (blue solid lines) weather station measurements and (red dashed lines) model output for (left) LOTOS-EUROS and (right) MOCAGE for (first row) De Bilt, (second row) Bremen, (third row) Uccle and (bottom row) OHP. Wind directions correspond 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.

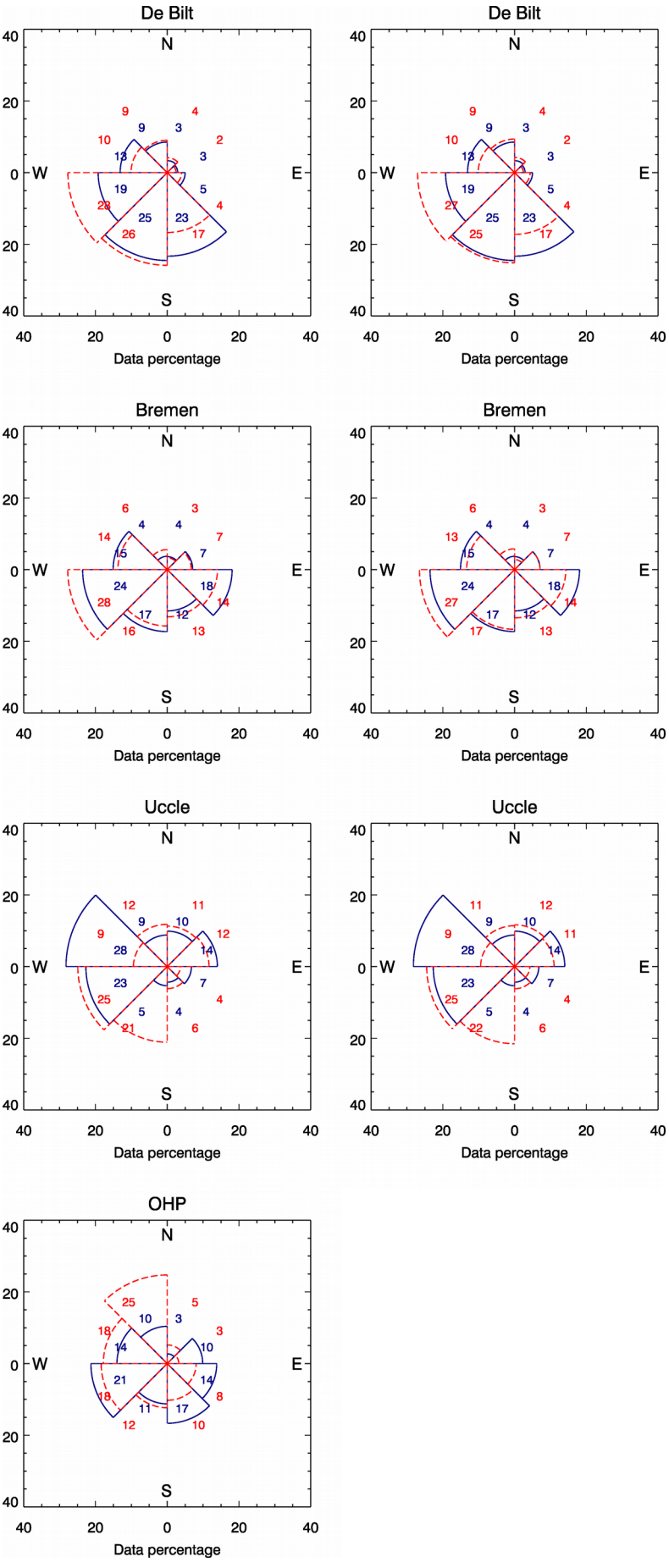

Figure R2: As in Figure R1 but for average percentage of data values. The printed numbers given for each bin were rounded to its closest integer value.

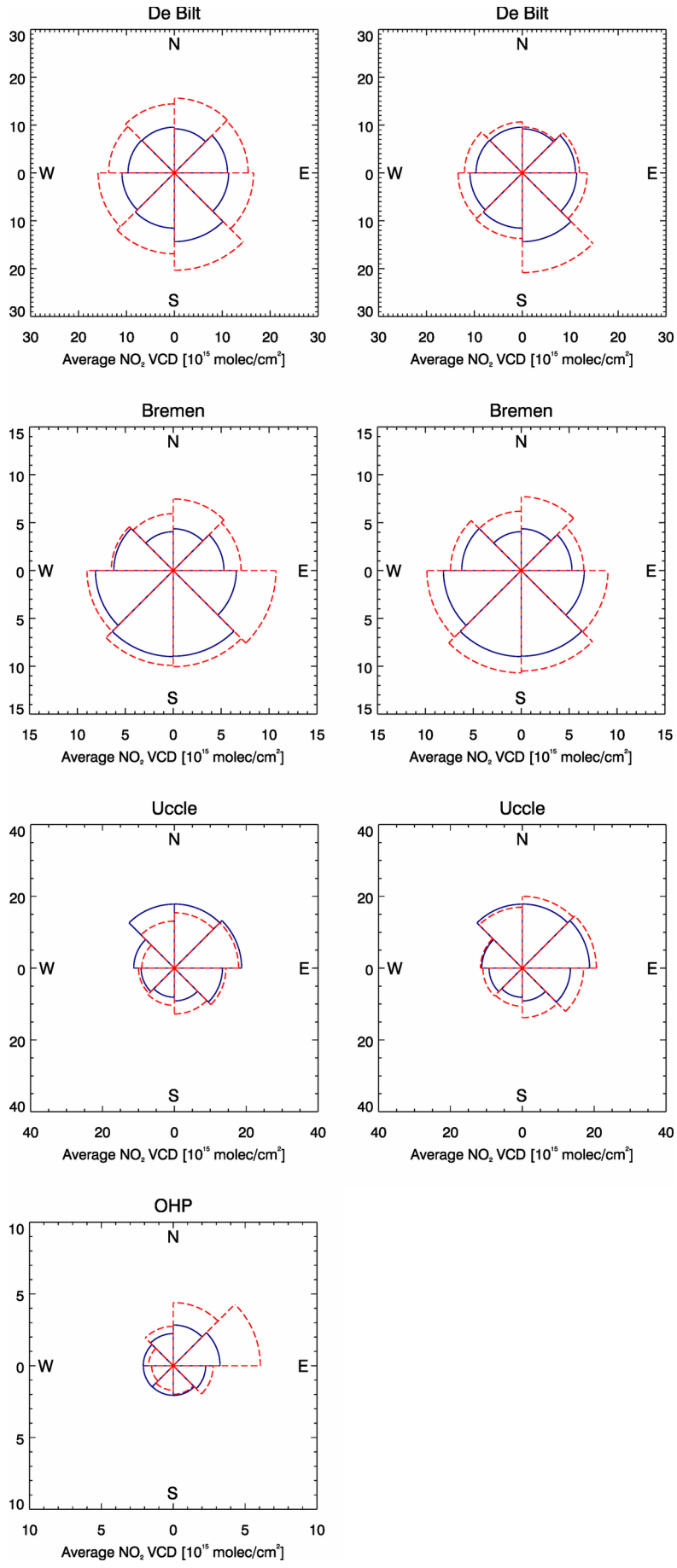

Figure R3: As in Figure R1 but for average AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}]. Wind directions correspond to the direction towards the station and are taken from weather station measurements for both MAX-DOAS retrieved and model simulated values.

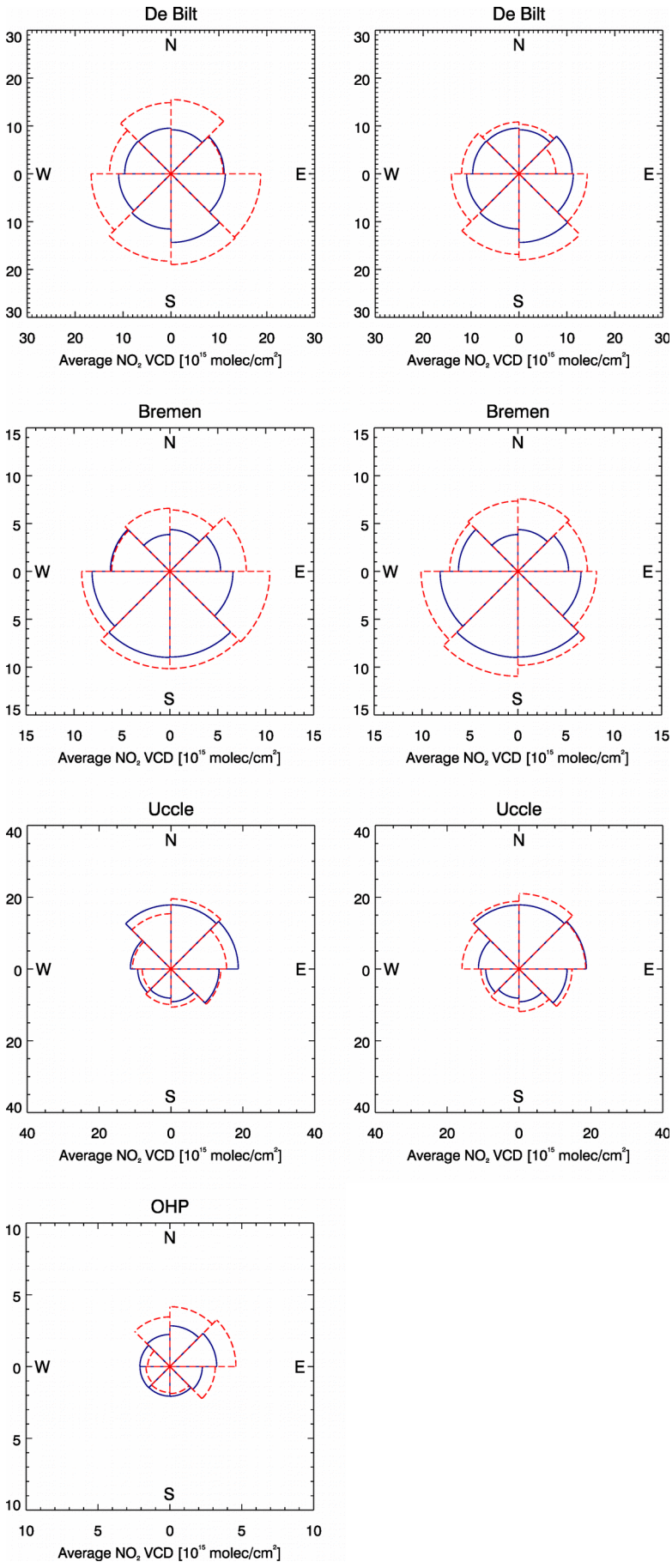

Figure R4: As in Figure R1 but for average AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}]. Wind directions correspond to the direction towards the station and are taken from weather station measurements for MAX-DOAS retrieved and from model output for simulated values.

Response to anonymous referee #2:

We thank referee #2 for constructive and helpful review comments, to which we hope to have responded appropriately. A list of comments including our response is given below.

In "Comparison of tropospheric NO2 columns from MAX-DOAS retrievals and regional air quality model simulations," the authors provide a nice overview of 1) long-term MAX-DOAS records of NO2 in northwest and southwest Europe, 2) a description of regional air quality models used in the CAMS ensemble, and 3) a description of past comparisons of regional CTMs and MAX-DOAS with in situ and satellite data. The comparison of the model ensemble and the four MAX-DOAS NO2 datasets showed general agreement in a broad sense. The authors highlight when and where there are discrepancies between ensemble median model results and MAX-DOAS observations (e.g., seasonal cycle, diurnal cycle), but do not offer ideas on potential approaches for disentangling the causes of these discrepancies.

I felt that the paper lacked a final synthesis, written in more general language, of how future simulations and MAX-DOAS deployments like these can isolate effects from individual processes. I hope that the authors consider adding a broader synthesis of their results to the end of section 4, offering possible paths forward for future analyses: what common and distinct attributes of these four sites share? How might these differences and similarities be exploited to investigate chemistry? Emissions? Meteorology? Where might the authors propose future MAX-DOAS instruments be located? Should one expect an ensemble median to capture hourly NO2 variations? Monthly averages? What is the native scale of NO2 spatial variations at the MAX-DOAS sites inferred from the time scale of NO2 variation and wind speed?

Many changes have been applied to the summary and conclusions section in the revised version including for example a discussion of model resolution and averaging volume of MAX-DOAS measurements, suggestions for sites to investigate in future studies (i.e. stations affected by different meteorological and pollution conditions for example at pollution hotspots in the Mediterranean with strong smog conditions especially during summer and clean mountain sites), a paragraph on OMI satellite comparisons with similar results as in Huijnen et al. (2010), as well as further suggestions on how to track down reasons for differences between model runs and MAX-DOAS retrievals (please see Section 5 of revised manuscript for further details).

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.

Comments: P2, L10-11: NO2 lifetime is much longer in the upper troposphere, primarily because its chemical family, NOx, is mostly present as NO at high altitudes, which has far fewer permanent sinks.

Changed to (p 2 l 14-16): "The lifetime of NOx is only a few hours in the boundary layer but a few days in the upper troposphere, where less OH radicals are present (Ehhalt et al., 1992) to react with NO_2 and more NOx is present as NO which has fewer permanent sinks than NO_2."

P3, L29: "focusses" – typo

This sentence has been deleted in response to a comment by referee #3.

P4: There is no discussion of model resolution. The NO2 lifetime is a function of model resolution. Also, median values may be biased towards coarser models as those with finer resolution may produce highs when a plume passes and lows when not.

In response to this question, the following text has been added to p 17 l 19 - p 18 l 9 of the revised manuscript (as response to a comment by referee #1, this is combined with a description on how the horizontal sensitivity range of MAX-DOAS compares to model resolution):

"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 as expected 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.

P5, L7: These sites, with exception of OHP, seem to be in very similar physical settings, with likely similar meteorology (e.g., vertical mixing characteristics). If so, this fact should be mentioned.

This is now mentioned in the summary and conclusions section together with suggestions for MAX-DOAS sites to be incorporated in future comparison studies (p 18 l 16-19):

"As the stations investigated in the present study have, apart from the rural background station OHP, rather similar meteorological and pollution conditions, investigation of stations over a broader range of different conditions would be desirable. Further comparison studies could for instance include stations at pollution hotspots in the Mediterranean such as Athens with strong smog conditions especially during summer and clean mountain sites."

Please also consider including a map of the region with sites indicated on a backdrop of satellite-based tropospheric NO2 column measurements.

The location of the MAX-DOAS stations is now shown in Figure 1 of the revised version, plotted on top of mean tropospheric columns of NO_2 from OMI for February 2011 as well as on top of TNO/MACC-II anthropogenic NOx emissions as an indicator of pollution levels in these and surrounding regions. The spatial distribution of NOx emissions agrees well with pollution hotpots and cleaner areas identified by OMI. Corresponding text has been added on p 4 l 1-4 of the revised version. 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.

Minor comment: I did not see Lat/Lon values reported for Uccle.

Added to revised version on p 6 l 33

Page 6, Line 29: Has there been any side-by-side operation and comparison of these two instruments? If so, please provide the reference.

The Uccle and OHP MAXDOAS instruments are a commercial mini-MAX-DOAS from Hoffmann Messtechnik GmbH and a BIRA research-grade spectrometer, respectively. Although there has not been formal side-by-side operation of both instruments for verification purpose, a good overall agreement has been obtained between the mini-DOAS and other BIRA research-grade spectrometers similar to the one operated at OHP, e.g. like during the CINDI campaign (see Roscoe et al., 2010). The last sentence has been added to p 7 l 14-17 of the revised manuscript.

Roscoe, H. K., Van Roozendael, M., Fayt, C., du Piesanie, A., Abuhassan, N., Adams, C., Akrami, M., Cede, A., Chong, J., Clémer, K., Frieß, U., Gil Ojeda, M., Goutail, F., Graves, R., Griesfeller, A., Grossmann, K., Hemerijckx, G., Hendrick, F., Herman, J., Hermans, C., Irie, H., Johnston, P. V., Kanaya, Y., Kreher, K., Leigh, R., Merlaud, A., Mount, G. H., Navarro, M., Oetjen, H., Pazmino, A., Perez-Camacho, M., Peters, E., Pinardi, G., Puentedura, O., Richter, A., Schönhardt, A., Shaiganfar, R., Spinei, E., Strong, K., Takashima, H., Vlemmix, T., Vrekoussis, M., Wagner, T., Wittrock, F., Yela, M., Yilmaz, S., Boersma, F., Hains, J., Kroon, M., Piters, A., and Kim, Y. J.: Intercomparison of slant column measurements of NO_2 and O_4 by MAX-DOAS and zenith-sky UV and visible spectrometers, Atmos. Meas. Tech., 3, 1629–1646, doi:10.5194/amt-3-1629-2010, 2010.

P9, L5: "As the typical error on MAX-DOAS retrieved VCDs is around 20%" – please describe this statement in more detail: at what time scale? Random or systematic uncertainty? Based on measurement intercomparisons or fitting statistics?

Uncertainty discussion of MAX-DOAS measurements is complex but has been done in previous studies (e.g. Hendrick et al., 2014; Wang et al., 2014; Franco et al., 2015). Briefly, uncertainties are a combination of small systematic errors (for example from the cross-sections used), random errors resulting from the DOAS retrieval, errors introduced by the profile retrieval and a priori assumptions made. In particular the latter contribution can vary depending on aerosol loading, vertical NO_2 profile and cloud contamination. In polluted conditions, uncertainties from profiling dominate. In clean situations, random errors from the fit can become significant. In general, uncertainties can be considered as random or pseudo-random, but systematic errors can result from, for example, the presence of elevated aerosol layers.

Quantification of uncertainties not only from error propagation but also from validation with independent measurements would be desirable, but very few suitable validation measurements are available, and differences are usually dominated by differences in measurement volume. Intercomparisons of different DOAS instruments show excellent (a few percent deviations) agreement on the level of slant columns (e.g. Roscoe et al., 2010) but substantial (20% - 50%) differences at the level of profiles.

Here, a simplified and conservative estimate of 30% uncertainty on all MAX-DOAS measurements has been assumed. Data products with more detailed uncertainty information are currently in development for example in the framework of the FRM4DOAS project (http://frm4doas.aeronomie.be/), and once available, this data and related uncertainty information should be used in future comparison studies.

The last sentence of the previous paragraph has been added on p 10 l 1-3 of the revised version.

Franco, B., Hendrick, F., Van Roozendael, M., Müller, J.-F., Stavrakou, T., Marais, E. A., Bovy, B., Bader, W., Fayt, C., Hermans, C., Lejeune, B., Pinardi, G., Servais, C., and Mahieu, E.: Retrievals of formaldehyde from ground-based FTIR and MAX-DOAS observations at the Jungfraujoch station and comparisons with GEOS-Chem and IMAGES model simulations, Atmos. Meas. Tech., 8, 1733-1756, doi:10.5194/amt-8-1733-2015, 2015.

Hendrick, F., Müller, J.-F., Clémer, K., Wang, P., De Mazière, M., Fayt, C., Gielen, C., Hermans, C., Ma, J. Z., Pinardi, G., Stavrakou, T., Vlemmix, T., and Van Roozendael, M.: Four years of ground-based MAX-DOAS observations of HONO and NO_2 in the Beijing area, Atmos. Chem. Phys., 14, 765–781, doi:10.5194/acp-14-765-2014, 2014.

Roscoe, H. K., Van Roozendael, M., Fayt, C., du Piesanie, A., Abuhassan, N., Adams, C., Akrami, M., Cede, A., Chong, J., Clémer, K., Frieß, U., Gil Ojeda, M., Goutail, F., Graves, R., Griesfeller, A., Grossmann, K., Hemerijckx, G., Hendrick, F., Herman, J., Hermans, C., Irie, H., Johnston, P. V., Kanaya, Y., Kreher, K., Leigh, R., Merlaud, A., Mount, G. H., Navarro, M., Oetjen, H., Pazmino, A., Perez-Camacho, M., Peters, E., Pinardi, G., Puentedura, O., Richter, A., Schönhardt, A., Shaiganfar, R., Spinei, E., Strong, K., Takashima, H., Vlemmix, T., Vrekoussis, M., Wagner, T., Wittrock, F., Yela, M., Yilmaz, S., Boersma, F., Hains, J., Kroon, M., Piters, A., and Kim, Y. J.: Intercomparison of slant column measurements of NO_2 and O_4 by MAX-DOAS and zenith-sky UV and visible spectrometers, Atmos. Meas. Tech., 3, 1629–1646, doi:10.5194/amt-3-1629-2010, 2010.

Wang, T., Hendrick, F., Wang, P., Tang, G., Clémer, K., Yu, H., Fayt, C., Hermans, C., Gielen, C., Pinardi, G., Theys, N., Brenot, H., and Van Roozendael, M.: Evaluation of tropospheric SO2 retrieved from MAX-DOAS measurements in Xianghe, China, Atmos. Chem. Phys. Discuss., 14, 6501-6536, doi:10.5194/acpd-14-6501-2014, 2014.

Page 9, L17-21: See comment on "page 4" above. NOx lifetime depends on model resolution, and NO2 maxima will be diluted in coarser models. Model resolution needs to be better reported.

See reply above.

P10, L5-17: I have a hard time following the language and reasoning behind this conclusion. Please consider clarifying. Is this because the a priori profiles are generated from similar models

as those included in the comparison? Are any systematic effects buried below random sources of uncertainty?

Multiplying simulated NO_2 partial columns by column AVKs of the retrievals prior to summing up partial columns in the vertical does not have a big impact on derived tropospheric NO_2 VCDs. One of the reasons for this is that (as shown by Figure 5 and A1, revised version), AVKs are close to 1 around the boundary layer where MAX-DOAS instruments have the highest sensitivity (generally a bit larger than one close to the surface and smaller than one higher up which has a balancing effect) and that the vertical shape of the column AVK curve is in principal agreement with the shape of simulated NO_2 partial columns. At altitudes above roughly 1 km, AVKs are on average for some stations significantly smaller than one, but simulated NO_2 partial columns are also significantly smaller at these altitudes compared to lower levels, so that the contribution to the tropospheric column is limited. At higher altitudes, MAX-DOAS retrievals tend to follow the a-priori, while retrievals in the boundary layer are not much influenced by the a-priori in general. This is in contrast to the situation for satellite observations of tropospheric NO_2, which usually have a minimum of the AVK in the boundary layer, i.e. where the largest fraction of NO_2 is usually located in polluted situations. A-priori profiles used within the MAX-DOAS retrievals (see Section 2.2) are in principal agreement with the ones simulated by the models. The vertical weighting caused by application of AVKs to partial columns does therefore not significantly impact on derived tropospheric NO_2 VCDs.

The information given in the paragraph above has been added to the results section and the corresponding text changed accordingly (see p 11 l 19 - p 12 l 2, revised version). Note that no profile retrievals are performed at De Bilt, which is therefore not shown in Figure 5.

Information on how a-priori profiles were derived for each station has been added to section 2.2. For Uccle and OHP, exponentially decreasing a-priori profiles were constructed based on an estimation of vertical column densities derived by so-called geometrical approximation (Hönninger et al., 2004; Brinksma et al., 2008) using scaling heights of 1 km and 0.5 km, respectively. For Bremen, an a-priori profile which is constant with height has been assumed in the retrieval. For De Bilt, a-priori profiles of NO_2 are based on a block-profile with NO_2 present the boundary layer, boundary layer heights were taken from a climatology based on ECMWF data.

Brinksma, E.J., Pinardi, G. J., Braak, R., Volten, H., Richter, A., Dirksen, R. J., Vlemmix, T., Swart, D. P. J., Knap, W. H., Veefkind, J. P., Eskes, H. J., Allaart, M., Rothe, R., Piters, A. J. M., and Levelt, P.F.: The 2005 and 2006 DANDELIONS NO_2 and Aerosol Intercomparison Campaigns. J. Geophys. Res.,113, D16S46, doi:10.1029/2007JD008808, 2008.

Hönninger, G., von Friedeburg, C., and Platt, U.: Multi axis differential optical absorption spectroscopy (MAX-DOAS), Atmos. Chem. Phys., 4, 231-254, doi:10.5194/acp-4-231-2004, 2004.

Page 10, L21-31: This analysis and discussion is tangential to the broader scope of the paper and should be removed, as earlier noted by the authors "The impact of clouds on MAX-DOAS retrievals is described in detail by Vlemmix et al. (2015)" I do consider the comparison of model and MAX-DOAS NO2 columns under different cloud conditions to be an interesting topic for its own manuscript.

The discussion and analysis of the impact of clouds on comparison results has been removed from the results section as suggested and is 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.

P10-11, L34-11: How much of the correlation is determined by seasonal and weekly cycle? Consider isolating correlation at one time of day, one season and one set of weekdays (e.g., M-F)

In response to this comment and comment b) by referee #1 three Tables have been added to the manuscript (note that in these Tables also results of individual model runs are summarized, in response to the requests by the other two referees to put more weight on 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

The following text has been added on p 15 l 22-24 of the revised version:

“Comparing Table 3 and 5 shows, that the overall correlations reached at all stations are mainly driven by seasonal and weekly cycles, while significantly lower and in many cases negative correlations are found for diurnal cycles which decreases overall correlations. An exception for the latter is Uccle, where good correlations are also found for diurnal cycles. ”

P12, L35: Consider a reference to Beirle et al. (2003). I think that this paragraph could be expanded. Day-of-week effects, over the long-term, are independent of meteorology and driven entirely by variations of emissions and chemistry. Future day-of-week comparisons would be one means of providing systematic approaches to quantify the many processes affecting NO2 (emissions, meteorology uncertainty, chemistry, observational uncertainty)

A reference to Beirle et al. (2003) has been added to p 15 l 19-21 of revised version:

“Beirle et al. (2003) investigated weekly cycles of tropospheric NO_2 based on GOME satellite observations and found a decrease in values of up to about 50 % towards Sundays over polluted regions and cities in Europe. This is in principal agreement with results of the present study, although the choice of the cities is different.”

Differences in diurnal cycles derived for weekdays and derived for weekends only 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. As expected, diurnal cycles retrieved from MAX-DOAS 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).

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.

Comparison of tropospheric NO_2 columns from MAX-DOAS retrievals and regional air quality model simulations

Anne-Marlene Blechschmidt[1], Joaquim Arteta[2], Adriana Coman[3], Lyana Curier[4,*], Henk Eskes[5], Gilles Foret[3], Clio Gielen[6], Francois Hendrick[6], Virginie Marécal[2], Frédérik Meleux[7], Jonathan Parmentier[2], Enno Peters[1], Gaia Pinardi[6], Ankie J. M. Piters[5], Matthieu Plu[2], Andreas Richter[1], Arjo Segers[4], Mikhail Sofiev[8], Álvaro M. Valdebenito[9], Michel Van Roozendael[6], Julius Vira[8], Tim Vlemmix[10], and John P. Burrows[1]

[1]Institute of Environmental Physics, University of Bremen, IUP-UB, Bremen, Germany
[2]Centre National de Recherches Météorologiques, Météo-France-CNRS, UMR 3589, Toulouse, France
[3]Laboratoire Interuniversitaire des Systèmes Atmosphériques, CNRS/INSU UMR7583, Université Paris-Est Créteil et Université Paris Diderot, Institut Pierre Simon Laplace, Créteil, France
[4]TNO, Climate Air and Sustainability Unit, Utrecht, the Netherlands
[5]Royal Netherlands Meteorological Institute, KNMI, De Bilt, the Netherlands
[6]Royal Belgian Institute for Space Aeronomy, BIRA-IASB, Brussels, Belgium
[7]Institut National de l'Environnement et des RISques industriels, INERIS, Verneuil en Halatte, France
[8]Finnish Meteorological Institute, FMI, Helsinki, Finland
[9]Norwegian Meteorological Institute, MetNo, Oslo, Norway
[10]TU-Delft, Delft, the Netherlands
[*]now at: Faculty of Humanities and Sciences, Department of Biobased Materials, Maastricht University, Geleen, the Netherlands

Correspondence to: A.-M. Blechschmidt (anne.blechschmidt@iup.physik.uni-bremen.de)

Abstract. Tropospheric NOx (NO+NO_2) is hazardous to human health and can lead to tropospheric ozone formation, eutrophication of ecosystems and acid rain production. It is therefore important to establish accurate data based on models and observations to understand and monitor tropospheric NO_2 concentrations on a regional and global scale.

In the present study, MAX-DOAS tropospheric NO_2 column retrievals from four European measurement stations are compared to ~~regional model ensemble simulations . The latter are based on~~ simulations from five regional air quality models which contribute to the European regional ensemble forecasts and reanalyses of the operational Copernicus Atmosphere Monitoring Service (CAMS). Compared to other observational data usually applied for regional model ~~validation~~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.

In general, there is ~~a~~ good agreement between simulated and retrieved NO_2 column values for individual MAX-DOAS measurements with correlations ~~between 45 and~~ roughly between 35 and 70 % for individual models and 45 to 75 % for the ensemble median for tropospheric NO_2 VCDs, indicating that ~~the model ensemble represents the emission~~ emissions, transport and tropospheric chemistry of NOx ~~(NO+NO_2) well. Pollution transport towards the stations is~~ are on average well ~~represented by the models~~simulated. However, large differences are found for individual pollution plumes ~~. Seasonal cycles are overestimated~~observed by MAX-DOAS. Most of the models overestimate seasonal cycles for the majority of MAX-DOAS sites

investigated. At the urban stations, weekly cycles are reproduced well but the decrease towards the weekend is underestimated and diurnal cycles ~~poorly representedby the model ensemble~~are overall not well represented. In particular, simulated morning rush hour peaks are not confirmed by MAX-DOAS retrievals ~~. Our results demonstrate that a~~ and models fail to reproduce observed changes in diurnal cycles for weekdays versus weekend. A large number of ~~validation points are available from~~ evaluation points arise from the comparison to MAX-DOAS measurements ~~, which should therefore be used more extensively~~ which should be used in future regional air quality modelling studies to track down reasons of disagreement.

1 Introduction

Nitrogen dioxide (NO_2) is a key species for atmospheric chemistry. Photolysis of NO_2 leads to formation of tropospheric ozone. The latter is a major greenhouse gas and the main precursor of OH, which itself determines the oxidising capacity of the atmosphere. Oxidation to HNO_3 via reaction with OH (daytime) or ozone (nighttime) is the major sink of NO_2 in the troposphere (Jacob, 1999) and results in acid rain and eutrophication of ecosystems, which are both harmful for the environment. Moreover, NO_2 can cause irritation of respiratory organs (http://www3.epa.gov/).

Within the troposphere, conversion of NO to NO_2 only takes about a minute during daytime. The sum of NO and NO_2 is called NOx, which is mainly emitted in the form of NO to the atmosphere. Main sources of NOx are fossil fuel combustion and biomass burning. Some NOx is also produced from lightning and microbial activity in soils.

The lifetime of ~~NO_2~~ NOx is only a few hours in the boundary layer but a few days in the upper troposphere, where less OH radicals are present (Ehhalt et al., 1992) to react with NO_2 and more NOx is present as NO which has fewer permanent sinks than NO_2. Several studies (e.g. Stohl et al., 2003; Zien et al., 2014) have shown that in the free troposphere, NO_2 can be transported over larger distances and is hence not only important for regional but also for global air quality. Peroxyacyl nitrate (PAN) produced by photochemical oxidation of carbonyl compounds is not much affected by wet scavenging and can act as a reservoir of NO_2, especially during long-range transport. If the air masses descend away from their source regions, PAN will decompose to NOx under the influence of, on average, higher temperatures at lower altitudes (Jacob, 1999).

Given the influence of NOx on air quality and climate through effects on radiation, it is of high environmental and scientific interest to accurately observe and simulate spatial distribution and time evolution of NO_2 concentrations in the troposphere. Simulating NO_2 is a challenge for numerical models as it is chemically very active and depends on many factors including for example cloud cover which affects photolysis of this trace gas. Moreover, ~~correct~~ representation of NOx emissions adds a large uncertainty to the model output.

MAX-DOAS (Multi Axis Differential Optical Absorption Spectroscopy; e.g. Hönninger et al., 2004; Wittrock et al., 2004) measurements have been used to investigate air pollution in many studies, including the FORMAT campaign in Northern Italy (Heckel et al., 2005; Wagner et al., 2011), the CINDI campaign in the Netherlands (Piters et al., 2012), campaigns in Canada (Halla et al., 2011; Mendolia et al., 2013), China (e.g. Irie et al., 2011; Hendrick et al., 2014; Ma et al., 2013; Wang et al., 2014), during ship-borne measurements (Leser et al., 2003; Takashima et al., 2012; Peters et al., 2012).

MAX-DOAS observations of atmospheric composition are performed by taking measurements of the scattered sunlight at different elevation and sometimes also azimuthal angles. Depending on the viewing angle and solar position, the light path through the atmosphere is different, with the observation in the zenith direction usually providing the shortest light path through the lower troposphere. Therefore, using ~~zenith measurements as intensity of incident radiation and observations in other angles as intensity of transmitted radiation~~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. These can be inverted to tropospheric columns and lower altitude tropospheric profiles by radiative transfer modelling and optimal estimation techniques.

A large number of studies applied MAX-DOAS data for satellite validation (e.g. Celarier et al., 2008; Valks et al., 2011; Irie et al., 2008; Irie et al., 2012; Ma et al., 2013; Lin et al., 2014; Kanaya et al., 2014; Pinardi et al., 2014) but up to now, comparisons to regional air quality model simulations of tropospheric NO_2 have, to our knowledge, only been carried out by Vlemmix et al. (2015) and Shaiganfar et al. (2015). Several studies compared regional air quality model simulations to satellite data (e.g. Huijnen et al., 2010), although satellite data are usually only available at much coarser time steps compared to regional model data. In this respect, the advantage of MAX-DOAS retrievals compared to satellite retrievals is the high resolution in time. Moreover, several studies compared in-situ NO_2 data to regional model results (e.g. Vautard et al., 2009; Colette et al., 2011; Mues et al., 2014)~~to regional model results~~, although in-situ data usually refer to a specific location (point measurements), whereas regional model results are available for a specific horizontal grid resolution and area depending on the model set up. As MAX-DOAS data represents a larger volume of air, it is much better suited for investigating performance of regional models than in-situ data. According to Richter et al. (2013) the horizontal averaging volume of MAX-DOAS data depends on aerosol loading, wavelength and viewing direction and ranges from a few kilometres in the polluted boundary layer up to 80 km from the top of a mountain under clean air conditions. ~~As MAX-DOAS data represents a larger volume of air, it is much better suited for regional model validation than in-situ data.~~ Another advantage of MAX-DOAS measurements is their ability to observe several pollution related species at the same time (e.g. NO_2, HCHO, CHOCHO, SO_2, aerosols, potentially also O_3) and to provide NO_2 data which is virtually free of interferences from other species or nitrogen compounds such as NOy (NOx and other oxidised nitrogen species). 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. Vertical profiles of trace gases can be retrieved from MAX-DOAS measurements, which is another advantage for model comparison studies.

~~The purpose of~~ In the present study~~is to investigate the usefulness of applying~~ , MAX-DOAS measurements ~~for validation of~~ are compared to regional air quality ~~models~~model simulations to investigate model performance. Parts of this approach are already applied within scientific reports of the operational Copernicus Atmosphere Monitoring Service (CAMS, http://atmosphere.copernicus.eu/), see e.g. Blechschmidt et al. (2015) and ~~, and in parts~~ Eskes et al. (2018), but mainly to model results provided on 8 output levels only, which introduces uncertainty to comparison results. CAMS is the operational follow-up of the former GEMS (Global and regional Earth-system Monitoring using Satellite and in-situ data) (Hollingsworth et al., 2008) and three succeeding MACC (Monitoring Atmospheric Composition and Climate, http://www.gmes-atmosphere.eu/)

projects. The global component of CAMS extends weather services of the ECMWF (European Centre for Medium-Range Weather Forecasts) with simulations of atmospheric trace gases and aerosols, while operational air quality forecasts and analyses for Europe are provided at much higher resolution through the regional component. Hourly NO_2 vertical column densities (VCDs) from 6 different regional model runs based on 5 models which are used within CAMS will be compared to MAX-DOAS measurements from three urban and one rural European station: Bremen (operated by IUP-Bremen), De Bilt (operated by KNMI), Uccle and OHP (Observatoire de Haute-Provence) (the latter two operated by BIRA-IASB). Location of the stations are plotted on top of mean NO_2 tropospheric columns from OMI (Levelt et al., 2006) satellite observations for February 2011 as well as on a map of anthropogenic NOx emissions used by the models in Figure 1 as an indicator of pollution levels in these and surrounding regions. The spatial distribution of NOx emissions agrees well with pollution hotpots and cleaner areas identified by OMI.

~~This study focusses on evaluating the usefulness of validating regional air quality models with MAX-DOAS observations in terms of validation~~ Due to the large number of model evaluation points arising from the ~~comparisons. The~~ MAX-DOAS based comparisons, the reasons for differences between model results and observations found by the comparisons are discussed here ~~only~~ in a general sense and need to be further investigated e.g. by carrying out additional dedicated model runs in future modelling studies.

The manuscript starts with an overview of regional model and MAX-DOAS data (Section 2) followed by a description of the ~~intercomparison~~ comparison method (Section 3). ~~Intercomparison results~~ Results are described and discussed in Section 4. Finally, a summary and conclusions are given in Section 5.

2 Data basis

2.1 Regional air quality model simulations

CHIMERE (Menut et al., 2013), LOTOS-EUROS (LOng Term Ozone Simulation - EURopean Operational Smog) (Schaap et al., 2008), EMEP MSC-W (European Monitoring and Evaluation Programme Meteorological Synthesizing Centre - West) (Simpson et al., 2012), MOCAGE (Model Of atmospheric Chemistry At larGE scale) (Josse et al., 2004; Guth et al., 2016) and SILAM (System for Integrated modeLling of Atmospheric coMposition) (Sofiev et al., 2006; Sofiev et al., 2015) contributed to the European regional ensemble forecasts (Marécal et al., 2015) and reanalyses of the former MACC projects and are currently used within CAMS. These models have been used in many studies for investigating atmospheric composition on a regional scale (e.g. Drobinski et al., 2007; Huijnen et al., 2010; Lacressonnière et al., 2014; Petetin et al., 2015; Solazzo et al., 2012; Watson et al., 2016; Zyryanov et al., 2012).

All of these models use ECMWF-IFS and MACC reanalysis (Innes et al., 2013) data as meteorological and chemical input data and boundary conditions, respectively. Anthropogenic emissions are taken from the MACC emissions database (Kuenen et al., 2011), GFAS (Kaiser et al., 2012) is used to account for fire emissions. The input to these models is thus consistent and hence, differences in model results are due to differences in the modelling code, model set up or due to different scalings of emissions e.g. to account for seasonal, diurnal and weekly cycles as well as emission heights. The model runs investigated

in the present study were performed by different European institutions and are based on different horizontal and vertical grid spacings and chemistry schemes (see Table 1 for further details). Apart from SILAM, the models were run without chemical data assimilation. The SILAM simulations included assimilation of surface observations of NO_2 as described in Vira and Sofiev (2015).

Two different sets of EMEP model runs are investigated in this study. The first one uses the same setup as the other regional models described above and is termed EMEP-MACCEVA in the following. EVA (validated assessments for air quality in Europe) was a subproject of MACC dedicated to the development and implementation of operational yearly production of European air quality assessment reports (https://www.gmes-atmosphere.eu). The second set of simulations (called EMEP in the following) uses the same set-up as in the EMEP status reports (see http://www.emep.int) for each year based on the EMEP subdomain, ECMWF-IFS as meteorological driver, EMEP emissions, Fire INventory from NCAR version 1.0 (FINNv1; Wiedinmyer et al., 2011), initial conditions described by Schulz et al. (2013) for the years 2010-2011 and Fagerli et al. (2014) for 2012 and climatological boundary conditions described by Simpson et al. (2012).

According to Mues et al. (2014), chemistry transport models in general account for seasonal, daily and diurnal emission changes by applying average time profiles given for different energy sectors and regions to totals of annual emissions across the model domain. Temporal emission patterns used by the regional air quality models listed above are country and SNAP (Selected Nomenclature for Sources of Air Pollution) sector dependent and are based on Denier van der Gon et al. (2011). A list of the SNAP sectors is given by Bieser et al. (2011). Moreover, different vertical emission profiles are applied for each regional model. These are described in more detail by Bieser et al. (2011) for EMEP and CHIMERE, Simpson et al. (2003) for SILAM and Thunis et al. (2010) for LOTOS-EUROS. For MOCAGE, emissions are injected into the five lowest model levels using a hyperbolic decay.

More details on specific model setups and scores with respect to surface observations, can be found in Marécal et al. (2015) and in the model specification/validation dossiers which are available online at:
http://www.gmes-atmosphere.eu/about/documentation/regional/.

2.2 MAX-DOAS retrievals

This study makes use of MAX-DOAS measurements from four ~~different~~ European stations: Bremen (Germany), De Bilt (the Netherlands), Uccle (Belgium), and OHP (France). Characteristics of the data available from the stations, such as exact location and time period of retrievals investigated here, are briefly summarized in Table 2 and will be described ~~in the following.~~ below.

For Bremen, Uccle and OHP, NO_2 slant column densities (SCDs) are obtained by a DOAS analysis for a specific wavelength window using a series of low elevation angles as measurement intensity and zenith measurements as reference intensity. Cross sections of different trace gases are accounted for in the retrieval. Resulting SCDs of NO_2 and O_4 are then used as input for a radiative transfer model which is a two-step approach. First, an aerosol extinction profile is retrieved by comparing the measured O_4 SCDs to O_4 SCDs simulated by the radiative transfer models SCIATRAN (Rozanov et al., 2005) for Bremen and bePRO (Clémer et al., 2010) for Uccle and OHP. In the second step, the derived aerosol extinction, measured NO_2 SCDs and an

a-priori NO_2 profile are used to retrieve the NO_2 profile of interest. This is an inverse problem solved by means of the optimal estimation method (Rodgers , 2000). The Maxdoas Retrieval algorithm of KNMI (MARK) uses a least squares minimization of the differences between measured and modeled differential slant column densities, by interpolation of look-up tables. The look-up tables are calculated with the radiative transfer model DAK (Doubling Adding KNMI; De Haan et al., 1987; Stammes , 2001). With this method, a maximum of four parameters are retrieved, which together determine the profile shape: tropospheric vertical column, boundary layer height, gradient in the boundary layer, fraction of NO_2 in the free troposphere.

De Bilt (52.10° N, 5.18° E; see Figure 1) is the home town of KNMI, and located just outside the city of Utrecht. The De Bilt experimental research site is surrounded by local and regional roads, with a lot of traffic which can affect regional air quality significantly. According to Vlemmix et al. (2015), it can also be affected by pollution sources which are located more far away in the Rotterdam region to the south-west, Amsterdam to the north-west and the German Ruhr region to the ~~south-west~~ south-east of De Bilt. The MAX-DOAS instrument operated at De Bilt is a commercial system obtained from Hoffmann Messtechnik. It has an Ocean Optics spectrograph, diffraction grating and a CCD detector. It operates at a wavelength range of 400-600 nm. ~~Differential slant columns are retrieved by the DOAS method, wavelength~~ The pointing direction of the instrument is 80° (east to north-east), the wavelength window of the DOAS fit for NO_2 is 425-490nm. Wavelength calibration and slit-function width are determined using a high-resolution solar spectrum. Cross sections of O_3, NO_2, O_4, H_2O and a pseudo cross-section accounting for the Ring effect are applied. The choice of fitting parameters complies with the standards agreed by the MAX-DOAS community, following from homogenization efforts within e.g. CINDI, GEOMON, NORS and QA4ECV as much as possible. Air mass factor (AMF) calculations are performed ~~with the DAK (Doubling Adding KNMI; , ; ,) radiative transfer model.~~ by the DAK model. A-Priori profiles of NO_2 are based on a block-profile with NO_2 present in the boundary layer, boundary layer heights were taken from a climatology based on ECMWF data. For De Bilt, averaging kernels refer to the altitude-dependent (or box-)differential AMFs divided by the total differential AMF. The differential AMF is derived at a specific altitude by simulating the radiance with and without an added partial column of NO_2 at this altitude with the DAK ~~model.~~ radiative transfer model. NO_2 columns are retrieved from the measurements at De Bilt, NO_2 profiles are not available.

The IUP-Bremen MAX-DOAS instrument consists of an outdoor telescope unit collecting light in different directions, and an indoor grating spectrometer (Shamrock 163 equipped with an Andor LOT257U CCD with 2048x512 pixels) covering a wavelength interval from 430–516 nm at a resolution of approximately 0.7 nm. Both components are connected via an optical fiber bundle which simplifies handling and overcomes polarization effects. The telescope unit is installed at an altitude of approximately 20 m above ground level at the roof of the Institute of Environmental Physics building (53.11° N, 8.86° E) at the University of Bremen which is located to the north-east of the city centre. The azimuthal pointing direction is north-west, which means that some of the measured pollution peaks are due to the exhaust of an industrial area, predominantly a steel plant, as well as a near-by highway. However, ~~averages~~ over longer time periods the retrievals should be dominated by pollution from the city centre. NO_2 ~~slant column densities are obtained from a~~ SCDs are derived by DOAS analysis using a fitting window ~~from~~ of 450-497 nm~~and elevation angles ranging from 0° to 15° in 1° steps as well as at 30° elevation angle and zenith direction (used as a reference)~~. Cross sections of O_3, NO_2, O_4, H_2O and a pseudo cross-section accounting for the Ring effect are ~~applied. Resulting slant columns~~ accounted for in the fit. Profiles of NO_2 ~~and O_4 are then input to~~ are derived from SCDs applying the

BRemian Advanced MAX-DOAS retrieval algorithm (BREAM) ~~, which is a two-step approach. First, an aerosol extinction profile is retrieved by comparing the measured O_4 slant columns to O_4 slant columns simulated using the radiative transfer model SCIATRAN . In the second step, the derived aerosol extinction, measured~~ which incorporates SCIATRAN radiative transfer simulations. An NO_2 ~~slant columns and an apriori NO_2 profile are used to retrieve the NO_2 profile of interest. This is an inverse problem solved by means of the optimal estimation method .~~ a-priori which is constant with altitude is assumed and iterated in the retrieval. Detailed information about the profile retrieval ~~can be found in~~ is given by Wittrock et al. (2006) and Peters et al. (2012).

BIRA-IASB operates a MAX-DOAS instrument at OHP (Observatoire de Haute-Provence; 43.92° N, 5.7° E) since 2005. OHP is a background remote site in the south of France, temporarily affected by transport of pollution from regional sources (e.g. from the petrochemical plants of Etang de Berre close to Marseille in the south-west) and the Po valley (Italy) to the north-east of the station. The MAX-DOAS instrument, which points towards the SSW direction, consists of a grating spectrometer ~~covering a wavelength range of 330-390 nm and collecting photons at 410001000~~Jobin-Yvon Triax 180 (1800 grooves/mm) covering the 330-390nm wavelength range coupled to a thermo-electrically cooled (-40°, ~~5°, 6°, 8°, 10°, 15°, 30° and 90° (zenith)viewing elevation angles~~C) Hamamatsu CCD detector (1024 pixels). NO_2 ~~differential slant column densities (DSCDs)~~ SCDs are obtained by applying the DOAS technique to a 364-384 nm wavelength interval, taking into account spectral signatures of O_3, O_4, the Ring effect and NO_2~~at 298 K. For the retrieval of aerosol extinction profiles (see below), O_4 is fitted to a wavelength interval of 338-370 nm including O_3, HCHO, BrO, the Ring effect and O_4 absorption cross-sections~~.

~~At Uccle,~~ At Uccle (50.8° N, 4.32° E), which is located south-west of the Brussels city centre, a mini-MAX-DOAS from Hoffmann Messtechnik GmbH covering the 290-435 nm wavelength range is operated by BIRA-IASB since 2011. The instrument is pointing ~~northwards~~ north to north-east towards the city centre~~and scans the following elevation angles: 3°, 4°, 5°, 6°, 7°, 9°, 11°, 13°, 16°, 31° and zenith~~. NO_2 ~~and O_4 DSCDs~~ SCDs are retrieved in a 407-432 ~~and 350-384 nm wavelength intervals, respectively,~~ wavelength interval including the same spectral signatures as for OHP. It should be noted that a sequential zenith reference spectrum has been implemented in order to minimise the impact of changes in shift and resolution due to temperature instabilities. The DOAS fit for NO_2 has also been improved by introducing pseudo-absorber cross-sections derived from principal component analysis of residuals on days affected by large thermal instabilities. This approach allows for a better correction of fast-changing slit-function variations, resulting in more stable residuals and therefore more realistic random uncertainty estimates.

For NO_2 vertical profile retrievals at both stations, the bePRO radiative transfer code (Clémer et al., 2010) is used~~, which is an inversion algorithm based on the optimal estimation method and consists of a two-step approach. Firstly, the model uses observed MAX-DOAS O_4 DSCDs to derive vertical profiles of aerosol extinction at different wavelengths. In the second step, NO_2 vertical profiles are derived from NO_2 DSCDs and the previously retrieved information on aerosol profiles. A more detailed description of the model and trace gas profile retrievals can be found in~~ .

NO_2 ~~profiles are retrieved~~ profiles are retrieved at 420 nm for Uccle and 372 nm for OHP. For NO_2 vertical profile retrievals, exponentially decreasing a-priori profiles have been constructed, based on ~~a first~~ an estimation of NO_2 vertical column densities derived from the so-called geometrical approximation (Hönninger et al., 2004; Brinksma et al., 2008) and using scaling heights

of 0.5 and 1 km for OHP and Uccle, respectively. A-priori and measurement-uncertainty covariance matrices are constructed as by Clémer et al. (2010) with adopted correlation lengths of 0.05, and covariance scaling values of 0.5 and 0.35 for Uccle and OHP, respectively. ~~Pressure and temperature profiles were taken from the US Standard Atmosphere and the retrieval grid consists of ten layers of 200 m thickness between the station altitude and 2 km altitude, two layers of 500 m thickness between 2 and 3 km and 1 layer between 3 and 4 km altitude.~~ For this study, only retrievals with a residual of the optimal estimation method retrieval fit to the DSCDs smaller than 50 % and degrees of freedom for signal larger than 1 are used. A more detailed description of the model and trace gas profile retrievals can be found in Hendrick et al. (2014). Although there has not been formal side-by-side operation of both instruments for verification purpose, a good overall agreement has been obtained between the mini-DOAS and other BIRA research-grade spectrometers similar to the one operated at OHP, e.g. like during the CINDI campaign (Roscoe et al., 2010).

Previous studies (e.g. Hendrick et al., 2014; Wang et al., 2014; Franco et al., 2015) have shown that the typical error on MAX-DOAS retrieved VCDs is around 20 %, including uncertainties related to the optimal estimation method, trace gas cross sections and aerosol retrievals, and can be higher for sites with low trace gas concentrations like OHP or due to instrumental conditions. Moreover, the uncertainty of the retrieval is increased in cloudy conditions.

For Uccle, information on cloud conditions ~~(i.e. clear-sky, thin clouds, thick clouds, broken clouds)~~ 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. The presence of clouds may alter MAX-DOAS retrievals in several ways: (1) If clouds are present at both zenith and horizon viewing directions, NO_2 within and above the clouds is shielded from the MAX-DOAS view whereas the sensitivity is slightly increased below the cloud, (2) if a cloud is present at the zenith/non-zenith viewing direction only, the sensitivity is reduced/enhanced at the height of the cloud and slightly enhanced/reduced below the cloud compared to the cloud free case. The impact of clouds on MAX-DOAS retrievals is described in detail by Vlemmix et al. (2015). In addition to the direct effect of clouds on the measurements, clouds also affect photolysis rates and hence NOx chemistry and NO to NO_2 partitioning, which may have an impact on tropospheric NO_2 columns and profiles retrieved under cloudy weather conditions. The influence of clouds on comparison results is hence complex and regarded as a topic for future studies.

2.3 Wind measurements

In order to investigate the ability of the models to reproduce transport of NO_2 towards the stations, the MAX-DOAS data described above is complemented by meteorological in-situ station data of wind speed and wind direction. Wind data for Bremen was provided by the German Weather Service/ Deutscher Wetterdienst through their website at http://www.dwd.de. The weather station in Bremen is located at the main airport, approximately 9 km southwards of the MAX-DOAS station. This may result in some differences to the actual wind direction and wind speed at time and location of the MAX-DOAS

retrievals. Wind data for OHP was taken from the weather station at the observatory and downloaded from the corresponding website at http://pc-meteo.obs-hp.fr/intervalle.php. Wind speed and direction measurements at Uccle are performed using a commercial rugged wind sensor from Young (model 05103) and were provided by BIRA-IASB through their webpage at http://uvindex.aeronomie.be. 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.

3 ~~Intercomparison method~~Methodology for regional model evaluation

The sensitivity of MAX-DOAS retrievals is largest in the boundary layer, which needs to be taken into account when comparing MAX-DOAS retrievals to model simulated values. This is achieved here, by applying column averaging kernels (AVKs) to the model data prior to comparison. The AVKs are part of the MAX-DOAS profiling output and represent the sensitivity of the retrieved column to the amount of NO_2 at different altitudes. Note that no profile data is available for De Bilt and AVKs were derived based on (box-)differential AMFs at that station (see Section 2.2).

In this study, model VCDs are derived by two different methods in order to test the influence of AVKs on the data analysis. ~~Model~~ Non AVK-weighted model VCDs are calculated by simply summing up NO_2 partial columns (VCD_i) over all N model levels in the vertical~~(method 1)~~:

$$VCD^{model}_{method1 nonAVK-weighted} = \sum_{i=1}^{Nmodel} VCD^{model}_i \quad (1)$$

In addition, model VCDs are calculated by applying column AVKs of the retrievals to model NO_2 partial columns before summing up NO_2 partial columns in the vertical~~(method 2)~~. The following data processing steps were carried out prior to the application of column AVKs:

(1) Conversion of provided model NO_2 partial columns [molec cm^{-2}] to concentrations [molec cm^{-3}] using model layer thicknesses.

(2) Deriving model concentrations on measurement altitudes assuming that model concentrations are constant within a specific model layer. If a measurement layer overlaps with more than one model layer, the result is a weighted mean over the model layer concentrations. If the highest measurement altitude is above the model top, the concentration at the model top level is used. It is assumed here that the latter has no significant impact on the data analysis, as NO_2 concentrations are in general small towards higher elevation levels compared to lower levels.

(3) Conversion of derived NO_2 concentrations on measurement altitudes to partial columns [molec cm^{-2}] using observation layer thicknesses.

~~Model VCDs for method 2~~ AVK-weighted model VCDs were then calculated using the following equation:

$$VCD^{model}_{method2 AVK-weighted} = \sum_{i=1}^{Nobs} AVK_i * VCD^{model}_i \quad (2)$$

where Nobs is the number of measurement altitudes.

Note that ~~method 1 and method 2 use~~ non AVK-weighted and AVK-weighted model VCDs are based on the model output at original vertical resolution. VCDs are calculated separately for each model and constitute the basis for calculating ensemble mean values which are described at the end of this Section.

Only those model values closest to the measurement time are used below. As the model output is given in hourly time steps, the maximum possible time difference between measurements and simulations is 30 minutes.

Following studies by e.g. ~~,~~ Marécal et al. (2015), Langner et al. (2012), Solazzo et al. (2012), Vautard et al. (2009), the present manuscript focuses on results of the model ensemble, i.e. the median of individual model results of a given quantity~~, for the sake of simplicity and in order to reduce individual model outliers~~. As an even number of 6 different model runs (based on 5 different models) constitute the model ensemble in the present study, the median is calculated by ordering the 6 different model values (e.g. for seasonal cycles, these values refer to the average of individual model runs for each month) in terms of magnitude and taking the average of the two middle numbers. An exception is OHP as MOCAGE data is not available for this station so that the median refers to the middle number here. ~~Standard deviations are calculated based on results from individual ensemble members (i.e. results prior to calculation of model ensemble mean values) and are used as an indicator of how much individual ensemble members differ from each other.~~ In addition, results from separate models are briefly discussed ~~where needed~~ and shown in the main part of the manuscript to understand characteristics of the model ensemble output. However, it is beyond the scope of this study to describe the performance of each individual model in detail. The reader is referred to the Appendix for additional comparison Figures of individual model simulations and MAX-DOAS data.

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.

As the typical error on MAX-DOAS retrieved VCDs is around 20 %, but can be higher for sites with low trace gas concentrations like OHP or due to instrumental conditions (see Section 2.2), a conservative overall uncertainty of MAX-DOAS retrievals of 30 % is assumed for all stations within this manuscript and given along with the data plots, where appropriate. Data products with more detailed uncertainty information are currently in development for example in the framework of the FRM4DOAS project (http://frm4doas.aeronomie.be/), and once available, this data and related uncertainty information should be used in future comparison studies.

4 ~~Intercomparison results~~Results

~~Figures ?? and ?? show~~ Figure 2 shows time series of AVK-weighted tropospheric NO_2 VCDs ~~derived by method 1 and 2 as well as surface partial columns (i.e. the partial column of the lowest measurement layer)~~ from MAX-DOAS and model ensemble data. ~~As vertical profiles are not available from the MAX-DOAS output for De Bilt, comparisons of profiles and surface partial columns are not given for this station in the present manuscript.~~ The magnitude of VCDs from the measurements for Bremen and OHP is reproduced by the model ensemble~~(using either method 1 or method 2). Model ensemble values are generally lower than the observed ones for Uccle and especially for De Bilt. However, at~~ . For Uccle and De Bilt, retrieved values tend to be larger than simulated ones. Low retrieved values appear overestimated at De Bilt and Bremen. At all of the four stations, measurements and simulations show large deviations for some of the time steps investigated. ~~Some of the larger~~ Larger NO_2 values inside individual pollution plumes are generally underestimated by the model ensemble~~.~~ , especially at Uccle and De Bilt. This is in agreement with Shaiganfar et al. (2015) who compared car MAX-DOAS measurements and OMI retrievals with a regional model (CHIMERE) and found that values inside emission plumes are systematically underestimated. The model ensemble may fail to reproduce these peaks due to errors in transport of NO_2 towards the stations or incomplete representation of atmospheric chemistry. An example of the latter would be overestimation of conversion to HNO_3, which may result in lower tropospheric NO_2 VCDs compared to MAX-DOAS if the transport is not happening quickly enough. Moreover, differences between simulations and retrievals may also arise from uncertainties of anthropogenic NOx emissions and horizontal resolution of model results (e.g. pollution sources may not be sufficiently resolved by the model simulations). Colette et al. (2014) compared regional model simulations with differing horizontal resolution and found that an increase in resolution leads to a better agreement with NO_2 in-situ data. However, as described in Section 1, MAX-DOAS observations are closer to regional model output in horizontal resolution than in-situ data.

As expected, the magnitude of NO_2 VCDs is lowest at the rural station OHP, which is sometimes affected by ~~local~~ near by pollution plumes that show up in the time series. Further investigation shows, that most of these peaks are associated with north-easterly wind directions and hence pollution sources to the north-east of the station such as the Po valley (Italy). ~~Applying column AVKs to model data for calculating VCDs (method 2) compared to method 1 does not have a big impact on validation results. Statistical values (root mean squared error, bias, correlation) which will be described below are quite similar for AVK weighted model ensemble values and those from method 1.~~ At OHP, retrieved tropospheric NO_2 columns are generally a bit higher than simulated ones. At least for the summer period, this is in agreement with Huijnen et al. (2010) who showed that the GEMS regional model ensemble median underestimates background values of tropospheric NO_2 columns compared to OMI satellite retrievals. Note that we carried out a similar comparison to OMI for the model runs of the present study, which showed similar results as Huijnen et al. (2010) and is therefore not shown here (see Section 5).

The evolution of time series of tropospheric NO_2 VCDs is largely determined by the evolution of surface partial columns ~~. Looking at the time series, surface partial columns~~ (see Figure 3) which already account for about 25 % of the magnitude of tropospheric NO_2 VCDs. 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. The same conclusions as for tropospheric NO_2 VCDs described in ~~this~~ the previous paragraph arise for surface partial columns when comparing model ensemble to MAX-DOAS data. The negative bias found for OHP for troposperic NO_2 VCDs is not present when looking at the surface partial column time series for this station (see also Table 3 and Table 4 where most models are negatively biased at OHP for tropospheric columns but not for surface partial columns), indicating that NO_2 lifted above the ground level is underestimated compared to MAX-DOAS, pointing at uncertainties related to the transport of pollution and/or chemical conversion during transport.

Although there are ~~large differences for individual data points~~larger differences between simulations and retrievals especially for individual pollution plumes, Figure 4 shows that frequency distributions of tropospheric NO_2 VCDs are similar for ensemble simulations and observations. However, for OHP the number of data values with tropospheric NO_2 VCDs lower than 1 x 10^{15} molec cm^{-2} is significantly larger for model simulated values (about 1400 model values compared to about 200 observed data counts) in agreement with the negative bias in tropospheric columns described above.

~~As the sensitivity of~~ Figure 5 shows model simulated and MAX-DOAS ~~retrievals is largest in the boundary layer, we initially expected the application of column AVKs from the measurements to model simulations to be of crucial importance for validation results. Therefore, the results described aboveat first instance are surprising but can be explained when comparing model ensemble to MAX-DOAS~~ retrieved vertical profiles of NO_2 partial columns averaged over the whole time period of measurements ~~shown in Figure 5 (a)~~together with a-priori profiles and AVKs for completeness. Averages of vertical profiles over ~~three months~~ different seasons are given in ~~Figures 5 (b) to (e)~~Figure A1, in order to investigate consistency between profiles throughout different ~~seasons. A-priori profiles assumed within the DOAS retrievals and AVKs are included in the plots for completeness. Differences~~ times of the year. In general, differences between retrievals and simulations are largest for larger NO_2 partial columns, which means for the lower altitude layers and during the colder winter and autumn seasons. Many of the ~~model simulated values~~ values simulated by individual models do not fall into the uncertainty range of MAX-DOAS retrievals assumed here. 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. Although model ensemble profiles show some differences to the retrievals regarding the ~~exact~~ shape and magnitude of the profiles, they also show the largest partial columns close to the surface for all of the three stations investigated. This result also shows up throughout different seasons. ~~This means that the main source of the scatter between measurements and simulations is not due to the vertical representativeness of the observations and as such,~~

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. However, further analysis showed that applying column AVKs to model NO_2 partial columns before summing these up in the vertical does not have a big impact on derived tropospheric NO_2 VCDs and therefore has a minor effect on the data analysis presented in this manuscript.

~~Individual model runs consistently show low partial columns at higher altitudes and disagree much more for values close to the surface, i.e. closer to NOx emission sources, which is expressed by generally larger standard deviations at lower altitudes. The seasonal variation of vertical profiles is reproduced by the model ensemble.~~

Only AVK-weighted simulations of tropospheric NO_2 VCDs are therefore shown here. Statistical values (root mean squared error, bias, Pearson correlation coefficient) which will be described below are quite similar for AVK weighted model ensemble VCDs and those from non AVK-weighted ones. One of the reasons for this is that (as shown by Figure 5 ~~(f)shows comparisons for Uccle only, but for MAX-DOAS measurements carried out under different cloud conditions (i.e., from left to right: clear-sky, thin clouds, thick clouds) as derived from the MAX-DOAS observations (~~and Figure A1), ~~; see Section 2.2) . On average , observed~~ AVKs are close to 1 around the boundary layer where MAX-DOAS instruments have the highest sensitivity (generally a bit larger than one close to the surface and smaller than one higher up which has a balancing effect) and that the vertical shape of the column AVK curve is in principal agreement with the shape of simulated NO_2 partial columns. At altitudes above roughly 1 km, AVKs are on average for some stations significantly smaller than one, but simulated NO_2 partial columns are ~~higher in the lowest observation layers during cloudy conditions compared to clear-sky conditions. Further investigation shows, that this feature is consistent throughout different seasons, except for MAM months for which only 3 of the observations were made under clear-sky conditions, whereas during other seasons, about 10 to 15 of the observations were made under clear-sky conditions. In theory, below a cloud less light and hence less OH is present in the lowest observations layers, which acts as a sink for~~ also significantly smaller at these altitudes compared to lower levels, so that the contribution to the tropospheric column is limited. At higher altitudes, MAX-DOAS retrievals tend to follow the a-priori, while retrievals in the boundary layer are not much influenced by the a-priori in general. This is in contrast to the situation for satellite observations of tropospheric NO_2~~during daytime. Moreover, less NO_2 is photolysed below a cloud.The model ensemble reproduces the overall change~~ , which usually have a minimum of the AVK in the boundary layer, i.e. where the largest fraction of NO_2 ~~partial columns from clear-sky to thin cloud conditions for the lowest observation layer. The strong simulated decrease in values from thin cloud to thick cloud conditions is not confirmed by the retrievals. To find out the reason for this would require further investigation, but could point to errors in simulating photochemistry under cloudy conditions.~~

~~In the following, only results from method 2 will be discussed. As shown above, these do not differ substantially from method 1 comparisons, which is true for all results presented below~~is usually located in polluted situations. A-priori profiles used within the MAX-DOAS retrievals (see Section 2.2) are in principal agreement with the ones simulated by the models. The vertical weighting caused by application of AVKs to partial columns does therefore not significantly impact on derived tropospheric NO_2 VCDs.

Scatter density plots of tropospheric NO_2 VCDs from MAX-DOAS against model values corresponding to the time series displayed by ~~Figures ?? and ?? (b)~~ Figure 2 are shown in Figure 6 (~~a). Scatter density plots of surface partial columns corresponding to time series in Figures ?? and ?? (c) are shown in Figure 6 (b~~see Figure A2 for individual model results). Statistical values (root mean squared error, bias, Pearson correlation) and least squares regression lines are given along with the plots ~~in Figure 6~~ to draw further conclusions on the ability of the model ensemble to reproduce MAX-DOAS retrievals ~~.~~ ~~Moderate correlations of~~ and are listed in Table 3 (together with statistics on ensemble members). Statistical values for surface

partial columns are given in Table 4. The ensemble median performs best in terms of overall correlation with values between 45 to 75 % ~~are found for~~ (compared to 35 to70 % for individual models) for tropospheric NO_2 VCDs for all stations, the highest correlation is found for Uccle. ~~Correlations are lower~~ Note however, that for other statistical indicators, some of the individual models perform better. Correlations are generally lower than the ones based on tropospheric columns for surface partial columns which are on the order of 40 % for Bremen and OHP, but much higher again for Uccle (∼60 %) ~~. The~~ for the ensemble. As expected from the comparisons described above, the model ensemble has a negative bias of about ~~0.3 and 2~~ -0.3 and -2 x 10^{15} molec cm^{-2} for OHP and Uccle, respectively, and a positive bias of about 1 x 10^{15} molec cm^{-2} for De Bilt and Bremen for tropospheric columns. 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). The spread between models and observations is large for some individual data points. Regression lines show that the model ensemble tends to overestimate low and underestimate high tropospheric NO_2 ~~VCD and surface partial column retrievals~~VCDs. The underestimation of larger tropospheric NO_2 VCDs is most pronounced for De Bilt, followed by Uccle.

Figure 7 shows comparisons ~~for~~ between MAX-DOAS and the model ensemble of wind directional distributions of average tropospheric NO_2 VCDs ~~and surface partial columns for the different stations~~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). Changes of NO_2 mean values from one wind direction bin to another are reproduced well by the model ensemble ~~,~~ (and in general also by ensemble members, see Figure A3), with an overall slightly better agreement with retrievals for tropospheric NO_2 VCDs compared to surface partial columns (not shown). Both, MAX-DOAS and model ensemble show the highest NO_2 mean values for wind directions mainly where influence from pollution sources is expected (i.e. Ruhr area to the south-east of De Bilt, the Bremen city centre to the south-west of the Bremen MAX-DOAS, Brussels city centre to the north-east of Uccle, the Po valley to the north-east of OHP, see Section 2.2). As for the time series comparisons described above, differences between observations and model results could be related to model uncertainties in simulating transport of pollution towards the measurement stations and chemistry. Uncertainties in anthropogenic emissions and background NO_2 VCDs may add up to differences between models and MAX-DOAS for wind directional distributions.

Comparisons for seasonal cycles (i.e. monthly averages) of tropospheric NO_2 VCDs are given in Figure 8 together with corresponding statistical values in Table 5. The number of MAX-DOAS measurements available for each month is given at the top y-axis of each seasonal cycle plot as an indicator of statistical significance. The number of data values is also shown for diurnal and weekly cycle Figures which will be discussed below. There is a good agreement between MAX-DOAS and the

model ensemble for Uccle regarding the magnitude of NO_2 VCDs and seasonality, with simulated ensemble median values within the estimated uncertainty interval of the retrievals. The same is true for De Bilt, apart from the strong overestimation of MAX-DOAS retrieved values for January, March and April. The latter may be explained by the low number of observations available during these compared to other months. The model ensemble overestimates seasonal cycles for Bremen and OHP. More explicitly, there is an overestimation of wintertime values while summertime values are better reproduced by the model ensemble. This may indicate that the model ensemble overestimates production of OH via photolysis of O_3 when less light is available, as OH acts as a sink for NO_2. The latter may also result from errors in simulating clouds and related photochemistry during the colder season. It may also point to an overestimation of anthropogenic emissions or inappropriate scalings of these. The former would be in agreement with Petetin et al. (2015), who found that anthropogenic NOx emissions from the TNO emission inventory (on which MACC emissions are based on) are overestimated, but those results apply to the Paris region only. Huijnen et al. (2010) compared an ensemble of regional and global models to satellite data over Europe and found an overestimation of seasonal cycles by the simulations, which is in agreement with results for Bremen and OHP shown in the present manuscript. However, according to Huijnen et al. (2010) model values were closer to satellite retrievals during winter, whereas for summer a strong underestimation was found, while comparisons to Dutch surface observations showed that this could be partly attributed to a high bias of satellite retrievals in summer at least over the Netherlands. In the present study, ~~wintertime standard deviations are~~ the spread between individual models is quite large for OHP indicating that some of the models perform better ~~compared to others(see Figure ?? for corresponding individual model results). Results for SILAM agree with Vira and Sofiev (2015)who found that this model tends to overestimate NO_2 at rural sites based on in-situ data~~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).

~~Figures ?? and ?? (a) show~~ 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). Although most of the model values fall within the estimated uncertainty interval of MAX-DOAS retrievals, the shape of diurnal cycles differs ~~from each other. While the model ensemble tends to simulate a decrease in tropospheric NO_2 VCDs from the morning to the afternoon for all stations, MAX-DOAS retrieved values generally increase towards the afternoon. Moreover, the ensemble~~ between observations and simulations. The ensemble shows a strong peak during the morning rush hour around 8 am for Bremen, which is not confirmed by MAX-DOAS retrievals. In contrast to this, measurements show a maximum around 2 pm in the afternoon which coincides with a very weak local maximum

simulated by the model ensemble. Looking at diurnal cycles for different seasons shown in ~~Figures ?? and ?? (b) to (e)~~ Figure A4 and A5 reveals that these are in general much better reproduced for spring and summer compared to autumn and winter for all stations. This is in agreement with results for seasonal cycles described in the previous paragraph. Weak morning rush hour peaks are also simulated for the rural station OHP, which is not in agreement with the measurements. ~~Inspection of corresponding individual model results (see Figures A4 and A5) shows that the~~ The morning rush hour peaks for Bremen and OHP occur for all models with the exception of SILAM for OHP, which however strongly overestimates values (by a factor of 1.5 to 2 for diurnal cycle values averaged over the whole time series) for this station, resulting in a bias of 1.3 x 10^{15} molec cm^{-2} (see Table 5). 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 ~~scenario~~ inventory as a basis, except ~~EMEP.~~ 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. The differences in diurnal cycles between model simulations and retrievals as well as between individual model runs could mean that the different scalings of NOx emissions applied by each model to account for diurnal variations are not appropriate, maybe in combination with uncertainties in vertical scalings. This should be investigated in future modelling studies. For example, according to Mailler et al. (2013) improving efficient emission heights is a key factor for improving background atmospheric composition simulated by chemistry transport models. However, the disagreement between simulated and measured values as well as the disagreement between individual model runs may also point to problems regarding photochemistry and treatment of boundary layer mixing. Differences in transport of pollution towards the stations during the morning and evening may add up to model uncertainties, especially for the rural station OHP where different shapes of diurnal cycles for individual model runs may also result from pollution transported from urban surrounding areas towards the station. Comparing to surface station measurements of ozone, Marécal et al. (2015) found that statistical indicators of model performance for MACC-II regional models show a pronounced diurnal cycle (best performances at 15 UTC, worst ones at 18 UTC) and attributed this to uncertainties in the diurnal cycle of ozone precursor emissions.

Figure 10 shows comparisons of diurnal cycles for weekends (Saturdays and Sundays) only. A Figure of diurnal cycles for weekdays only shows very similar results as Figure 9 (and is therefore not shown here), meaning that overall diurnal cycles are mainly driven by weekday emissions. At the three urban stations, MAX-DOAS retrieved diurnal cycles show a different shape for weekends compared to diurnal cycles for the whole week (and hence weekdays only). This is in contrast to model simulated diurnal cycles, which do not change much going from cycles for the whole week to cycles for weekends only, apart from a general decrease in values towards weekends for both retrieved and simulated tropospheric NO_2 VCDs. As expected, MAX-DOAS retrieved diurnal cycles are rather flat for weekends only at the urban stations, as emissions from traffic and industry are reduced during weekends compared to weekdays (e.g. Elkus et al., 1977; Beirle et al., 2003; Ialongo et al., 2016). As the shape of simulated diurnal cycles is similar for weekdays versus weekend, the difference between retrieved and simulated trends in tropospheric columns from morning to afternoon hours is reduced for weekends only resulting in

significantly higher and positive correlations for diurnal cycles during weekends compared to weekdays for the ensemble at these stations (see Table 5). At Uccle, correlations are equally high (about 80 %) for weekdays and weekends, which is due to the fact that the shape of retrieved diurnal cycles is also similar. Correlations are also significantly higher for the background station OHP for weekends for the ensemble, mainly due to a better agreement in the development from the afternoon towards the evening during weekends. However, visually/by eye, the agreement between simulations and retrievals is similar for weekdays and weekends for this station. The results described above show that models fail to reproduce observed changes in diurnal cycles towards the weekend at urban stations, indicating that different diurnal scalings should be applied to emissions for weekdays and weekends. It should be tested in future simulations if switching off diurnal scalings during weekends leads to an improvement in model performance compared to MAX-DOAS.

Weekly cycle comparisons are presented in ~~Figures ?? and ?? for the whole time series and for different seasons~~Figure 11 (see Figures A6 and A7 for different seasons). In contrast to diurnal cycles, weekly cycles and their seasonal variation measured by MAX-DOAS are much better simulated, ~~with a small underestimation by the models compared to the retrievals~~reflected by high correlations (Table 5) for the ensemble at all stations. Both, MAX-DOAS and the model ensemble, show a decrease in tropospheric NO_2 VCDs towards the weekend when there is less traffic especially for the urban stations De Bilt, Bremen and Uccle. However, ~~the observed decrease~~ this observed weekly cycle is stronger than the simulated one, a feature which is most pronounced for Bremen. This is in agreement with Vlemmix et al. (2015) who also found an underestimation of the weekly cycle when comparing LOTOS-EUROS simulations to MAX-DOAS retrievals for De Bilt. As expected, only a very weak weekly cycle is observed by MAX-DOAS and simulated by the models for the rural station OHP. Note that maxima of weekly cycles for specific days may just be coincidence due to data sampling times. Beirle et al. (2003) investigated weekly cycles of tropospheric NO_2 based on GOME satellite observations and found a decrease in values of up to about 50 % towards Sundays over polluted regions and cities in Europe. This is in principal agreement with results of the present study, although the choice of the cities is different.

Comparing Table 3 and 5 shows, that the overall correlations reached at all stations are mainly driven by seasonal and weekly cycles, while significantly lower and in many cases negative correlations are found for diurnal cycles which decreases overall correlations. An exception for the latter is Uccle, where good correlations are also found for diurnal cycles.

5 Summary and conclusions

In this study, comparisons between NO_2 columns simulated by ~~a regional model ensemble~~ five regional models and retrieved from MAX-DOAS measurements for four European MAX-DOAS stations have been presented.

~~This study focusses on evaluating the usefulness of validating regional air quality models with MAX-DOAS observations in terms of validation points arising from the comparisons.~~ The reasons for differences between model results and observations found by the comparisons are discussed here ~~only~~ in a general sense and need to be further investigated by carrying out additional dedicated model runs in future modelling studies. In general, differences between simulated and retrieved tropospheric NO_2 VCDs as well as surface partial columns found in this study could result from model uncertainties in chemistry and mete-

orology or a combination of both. Moreover, errors ~~in the~~ related to NOx emission inventories or uncertainties in tropospheric MAX-DOAS retrievals may also contribute to differences between simulated and retrieved values found in this study.

Our analysis shows that in general and on average the model ensemble does well represent ~~emissions and tropospheric chemistry of NOx~~tropospheric NO_2 amounts observed by MAX-DOAS. However, many ~~validation points~~ points to evaluate arise from the MAX-DOAS based comparisons~~which~~. 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. Moderate correlations around 60 % are found for tropospheric NO_2 VCDs at each station for the ensemble. Time series comparisons and corresponding scatterplots show that uncertainties in simulating pollution transport towards the stations is a likely reason for the underestimation of MAX-DOAS retrieved pollution peaks by the model ensemble. This may also lead to the weak simulated morning rush hour peak for the rural station OHP, which is not confirmed by the retrievals. In fact, for OHP, a diurnal cycle representative of a remote background NO_2 station would be expected. However, comparisons of wind directional distributions of tropospheric NO_2 VCDs and surface partial columns show a good agreement between simulations and measurements. This indicates that transport of pollution towards the stations is, on average, well represented by the models.

Comparisons of vertical profiles show that the main source of the scatter between measurements and simulations is not ~~the correct~~ due to incorrect representation of the vertical NO_2 distribution. Hence, there are no large differences between comparisons which do not make use of column AVKs for calculating model VCDs and those based on more accurate column ~~AVK weighted~~ AVK-weighted values. The latter result was not expected as the sensitivity of the MAX-DOAS profile retrievals is much larger close to the surface than at altitudes larger than approximately 1 km.

Seasonal cycles are overestimated by the model ensemble. Simulation uncertainties in photochemistry ~~are a conceivable explanation~~ and/or in monthly scalings of emissions are conceivable explanations for this. As MAX-DOAS measurements are carried out throughout the whole course of a day during daylight and are hence available with a comparatively high resolution in time, it is (in contrast to many other approaches) possible to compare diurnal cycles derived from simulations and measurements. This reveals that models fail to reproduce the shape of diurnal cycles for all stations as well as the observed change in diurnal cycles from weekdays towards weekends at urban stations, which most likely points to uncertainties in diurnal ~~scaling~~ scalings of emissions. Improving model results for diurnal cycles could potentially have a strong impact on all other comparisons shown in this manuscript and hence may further improve model performance. This is in agreement with Mues et al. (2014) who found an improvement of correlations between LOTOS-EUROS and in-situ data when applying a time profile to emissions. It should be tested in future studies if switching off diurnal scalings during weekends leads to an improvement in model performance compared to MAX-DOAS. 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 ~~well~~ better represented by the model ensemble, ~~indicating that~~ which indicates that applied scalings of emissions on a daily basis are ~~appropriate~~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.

In addition to the MAX-DOAS comparisons shown in the present study, we also carried out a comparison between the regional models and OMI (Levelt et al., 2006) satellite retrievals looking at maps of monthly means for a winter and summer month (February and August 2011, respectively) falling into the time period investigated by the present study. We found similar results as Huijnen et al. (2010) which are therefore not shown here, i.e. an underestimation of tropospheric NO_2 columns over background regions during summer (in agreement with the general underestimation of means over summer months compared to MAX-DOAS shown by seasonal cycles for OHP for all models except SILAM) and a generally better agreement between satellite retrievals and models over pollution hotspots around Benelux countries, an underestimation however of values over large parts of Germany and over the Po valley in many of the model runs. Some of the models also overestimated values to the south and south-east of OHP (roughly between Marseille and Genua along the southern coast of France) compared to OMI. 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).

Our evaluation demonstrates that the large number of measurements available from the current MAX-DOAS network constitutes a useful data source for ~~validation~~ investigating the performance of regional models. In contrast to other measurements usually applied for ~~validation~~ evaluation of regional models, MAX-DOAS data are available with comparatively high resolution in time. Furthermore, MAX-DOAS retrievals are representative of a larger volume of air and are therefore much better suited for regional model ~~validation~~ evaluation than in-situ data. ~~Nevertheless, it would~~

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 ~~complement and compare resultsof this study by other~~ 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. 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.

Comparison results of this study could be compared and complemented by further data sources where possible. Future investigations of regional model performance may also include application of stricter quality filters on the MAX-DOAS data to reduce the impact of retrieval uncertainty. As the discussion here is based on results of five regional models used within CAMS for four European stations, similar comparisons to other regional models or other model set-ups as well as for more MAX-DOAS sites should follow. As the stations investigated in the present study have, apart from the rural background station OHP, rather similar meteorological and pollution conditions, investigation of stations over a broader range of different conditions would be desirable. Further comparison studies could for instance include stations at pollution hotspots in the Mediterranean such as Athens with strong smog conditions especially during summer and clean mountain sites. The impact of different model set-ups and different anthropogenic emission inventories ~~as well as horizontal and vertical scalings of the latter on validation~~ on comparison results should be tested in order to improve model performance. ~~As the discussion here is based on results from five regional models used within CAMS for four European stations, similar comparisons to other regional models as well as for more MAX-DOAS stationsshould follow~~Moreover, the complex influence of clouds on comparison results could be investigated.

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.

6 Code availability

Source code and test data sets for the Open Source EMEP/MSC-W model are available at https://github.com/metno/emep-ctm or by contacting EMEP/MSC-W (emep.mscw@met.no). The SILAM code is available on request from the authors (mikhail.sofiev@fmi.fi, julius.vira@fmi.fi). The MOCAGE results in the present paper are based on source code which is presently incorporated in the MOCAGE model. The MOCAGE source code is the property of Météo-France and CERFACS, and it is based on libraries that belong to some other holders. The MOCAGE model is not open source and routines from MOCAGE cannot be freely distributed. CHIMERE is an open source code protected under the GNU General Public license. It can be found at http://www.lmd.polytechnique.fr/chimere/. LOTOS-EUROS is downloadable free of charge after signing a license agreement. All information concerning the LOTOS-EUROS code is available on the website (http://lotos-euros.nl), for further information the reader can contact Dr. A. Manders (astrid.manders@tno.nl).

Acknowledgements. This study was funded by the European Commission under the EU Seventh Research Framework Programme (grant agreement no. 283576, MACC II), the EU Horizon 2020 Research and Innovation programme (grant agreement no. 633080, MACC-III) and the Copernicus Atmosphere Monitoring Service (CAMS), implemented by the European Centre for Medium-Range Weather Forecasts (ECMWF) on behalf of the European Commission. It was also funded in part by the University of Bremen. The LOTOS-EUROS work was carried out within the ESA project, GLOB-EMISSION (grant number AO/1-6721/11/I-NB). BIRA-IASB MAX-DOAS observations at Uccle and OHP were financially supported by the projects AGACC-II (BELSPO, Brussels) and NORS (EU FP7; contract 284421). We thank the German Weather Service/ Deutscher Wetterdienst for providing wind in-situ data for Bremen through their website at http://www.dwd.de. We are also gratefull to people behind the wind in-situ data at Uccle and OHP for providing these measurements through the webpages http://uvindex.aeronomie.be and http://pc-meteo.obs-hp.fr/intervalle.php, respectively.

References

~~Antonakaki, T., Blechschmidt, A.-M., Clark, H., Gielen, C., Hendrick, F., Kapsomenakis, J., Mortier, A., Peters, E., Piters, A., Richter, A., van Roozendael, M., Schulz, M., Wagner, A., Zerefos, C., and Eskes, H. J.: Validation of CAMS regional services: concentrations above the surface, CAMS report, 30 June 2016, 2016.~~

Beirle, S., Platt, U., Wenig, M., and Wagner, T.: Weekly cycle of NO_2 by GOME measurements: a signature of anthropogenic sources, Atmos. Chem. Phys., 3, 2225-2232, doi:10.5194/acp-3-2225-2003, 2003.

Bergström, R., Denier van der Gon, H.A.C., Prevot, A.S.H., Yttri, K.E. and Simpson, D.: Modelling of organic aerosols over Europe (2002-2007) using a volatility basis set (VBS) framework with application of different assumptions regarding the formation of secondary organic aerosol, Atmos. Chem. Physics, 12, 5425–5485, 2012.

Bieser, J., Aulinger, A., Matthias, V., Quante, M., and Denier van Der Gon, H. A. C.: Vertical emission profiles for Europe based on plume rise calculations, Environmental Pollution 159 (10), 2935-2946, 2011.

Blechschmidt, A.-M., Coman, A., Curier, L., Eskes, H. J., Foret, G., Gielen, C., Hendrick, F., Marecal, V., Meleux, F., Parmentier, J., Peters, E., Pinardi, G., Piters, A., Plu, M., Richter, A., Sofiev, M., Valdebenito, Á. M., Van Roozendael, M., Vira, J., and Vlemmix, T.: MAX-DOAS tropospheric NO_2 column retrievals as a validation tool for regional air quality models of the upcoming Copernicus Atmosphere Monitoring Service, CAMS report, 30 September 2015, 2015.

Brinksma, E.J., Pinardi, G. J., Braak, R., Volten, H., Richter, A., Dirksen, R. J., Vlemmix, T., Swart, D. P. J., Knap, W. H., Veefkind, J. P., Eskes, H. J., Allaart, M., Rothe, R., Piters, A. J. M., and Levelt, P.F.: The 2005 and 2006 DANDELIONS NO_2 and Aerosol Intercomparison Campaigns. J. Geophys. Res.,113, D16S46, doi:10.1029/2007JD008808, 2008.

Celarier, E. A., Brinksma, E. J., Gleason, J. F., Veefkind, J. P., Cede, A, Herman, J. R., Ionov, D., Goutail, F., Pommereau, J.-P., Lambert, J.-C., van Roozendael, M., Pinardi, G., Wittrock, F., Schönhardt, A., Richter, A., Ibrahim, O. W. , Wagner, T., Bojkov, B., Mount, G., Spinei, E., Chen, C. M. Pongetti, T. J. , Sander, S. P., Bucsela, E. J., Wenig, M. O. , Swart, D. P. J., Volten, H., Kroon, M., and Levelt, P. F.: Validation of Ozone Monitoring Instrument nitrogen dioxide columns, J. Geophys. Res., 113, D15S15, doi:10.1029/2007JD008908, 2008.

Clémer, K., Van Roozendael, M., Fayt, C., Hendrick, F., Hermans, C., Pinardi, G., Spurr, R., Wang, P., and De Mazière, M.: Multiple wavelength retrieval of tropospheric aerosol optical properties from MAXDOAS measurements in Beijing, Atmos. Meas. Tech., 3, 863-878, doi:10.5194/amt-3-863-2010, 2010.

Colette, A., Granier, C., Hodnebrog, Ø., Jakobs, H., Maurizi, A., Nyiri, A., Bessagnet, B., D'Angiola, A., D'Isidoro, M., Gauss, M., Meleux, F., Memmesheimer, M., Mieville, A., Rouïl, L., Russo, F., Solberg, S., Stordal, F., and Tampieri, F.: Air quality trends in Europe over the past decade: a first multimodel assessment, Atmos. Chem. Phys., 11, 11657–11678, doi:10.5194/acp-11-11657-2011, 2011.

Colette, A., Bessagnet, B., Meleux, F., Terrenoire, E., and Rouïl, L.: Frontiers in air quality modelling, Geosci. Model Dev., 7, 203-210, doi:10.5194/gmd-7-203-2014, 2014.

Curier, R. L., Timmermans, R., Calabretta-Jongen, S., Eskes, H., Segers, A., Swart, D., and Schaap, M: Improving ozone forecasts over Europe by synergistic use of the LOTOS-EUROS chemical transport model and in situ measurements. Atmos. Environ, 60, 217–226, 2012.

Curier, R. L., Kranenburg, R., Segers, A., Timmermans, R., and Schaap, M.: Synergistic use of OMI-NO_2 tropospheric columns and LOTOS-EUROS to evaluate the NOx emission trends across Europe, Remote Sens. Environ., 149, 58 69,doi:10.1016/j.rse.2014.03.032, 2014.

De Haan, J.F., Bosma, P.B., and Hovenier, J.W.: The adding method for multiple scattering calculations of polarized light, Astron. Astrophys., 183, 371-393, 1987.

Denier van der Gon, H. A. C., Hendriks, C., Kuenen, J., Segers, A. and Visschedijk, A.: Description of current temporal emission patterns and sensitivity of predicted AQ for temporal emission patterns, TNO report, EU FP7 MACC deliverable report D_D-EMIS_1.3, Utrecht, the Netherlands, 2011.

Drobinski, P., Saïd, F., Ancellet, G., Arteta, J., Augustin, P., Bastin, S., Brut, A., Caccia, J., Campistron, B., Cautenet, S., Colette, A., Coll, I., Corsmeier, U., Cros, B., Dabas, A., Delbarre, H., Dufour, A., Durand, P., Gu enard, V., Hasel, M., Kalthoff, N., Kottmeier, C., Lasry, F., Lemonsu, A., Lohou, F., Masson, V., Menut, L., Moppert, C., Peuch, V., Puygrenier, V., Reitebuch, O., and Vautard, R.: Regional transport and dilution during high-pollution episodes in southern France: Summary of findings from the Field Experiment to Constraint Models of Atmospheric Pollution and Emissions Transport (ESCOMPTE), J. Geophys. Res., 112, D13105, doi:10.1029/2006JD007494, 2007.

Ehhalt, D. H., Rohrer, F., and Wahner, A.: Sources and Distribution of NOx in the Upper Troposphere at Northern Mid-Latitudes, J. Geophys. Res., 97, 3725–3738, http://www.agu.org/journals/jd/v097/iD04/91JD03081/, 1992.

Elkus, B. and Wilson, K. R.: Photochemical air pollution: weekend-weekday differences, Atmos. Environ., 11, 509–515, 1977.

Eskes, H. J.,Douros, J., Akriditis, D., Antonakaki, T., Bennouna, Y., Blechschmidt, A.-M., Bösch, T., Clark, H., Gielen, C., Hendrick, F., Kapsomenakis, J., Kartsios, K., Katragkou, E., Melas, D., Mortier, A., Peters, E., Petersen, K., Piters, A., Richter, A., van Roozendael, M., Schulz, M., Sudarchikova, N., Wagner, A., Zanis, P., Zerefos, C., and Eskes, H. J.: Validation of CAMS regional services: concentrations above the surface, Status update for September-November 2017, CAMS report, 27 February 2018, 2018.

Fagerli, H., Solberg, S., Tsyro, S., Benedictow, A., Aas, W., Hjellbrekke, A.-G. and Posch, M.: Status of transboundary pollution in 2012, Transboundary particulate matter, photo-oxidants, acidifying and eutrophying components, EMEP Status Report 1/2014, The Norwegian Meteorological Institute, Oslo, Norway, 2014.

Franco, B., Hendrick, F., Van Roozendael, M., Müller, J.-F., Stavrakou, T., Marais, E. A., Bovy, B., Bader, W., Fayt, C., Hermans, C., Lejeune, B., Pinardi, G., Servais, C., and Mahieu, E.: Retrievals of formaldehyde from ground-based FTIR and MAX-DOAS observations at the Jungfraujoch station and comparisons with GEOS-Chem and IMAGES model simulations, Atmos. Meas. Tech., 8, 1733-1756, doi:10.5194/amt-8-1733-2015, 2015.

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.

Guth, J., Josse, B., Marécal, V., Joly, M., and Hamer, P.: First implementation of secondary inorganic aerosols in the MOCAGE version R2.15.0 chemistry transport model, Geosci. Model Dev., 9, 137-160, doi:10.5194/gmd-9-137-2016, 2016.

Halla, J. D., Wagner, T., Beirle, S., Brook, J. R., Hayden, K. L., O'Brien, J. M., Ng, A., Majonis, D., Wenig, M. O., and McLaren, R.: Determination of tropospheric vertical columns of NO_2 and aerosol optical properties in a rural setting using MAX-DOAS, Atmos. Chem. Phys., 11, 12475-12498, doi:10.5194/acp-11-12475-2011, 2011.

Heckel, A., Richter, A., Tarsu, T., Wittrock, F., Hak, C., Pundt, I., Junkermann, W., and Burrows, J. P.: MAX-DOAS measurements of formaldehyde in the Po-Valley, Atmos. Chem. Phys., 5, 909–918, 2005.

Hendrick, F., Müller, J.-F., Clémer, K., Wang, P., De Mazière, M., Fayt, C., Gielen, C., Hermans, C., Ma, J. Z., Pinardi, G., Stavrakou, T., Vlemmix, T., and Van Roozendael, M.: Four years of ground-based MAX-DOAS observations of HONO and NO_2 in the Beijing area, Atmos. Chem. Phys., 14, 765–781, doi:10.5194/acp-14-765-2014, 2014.

Hönninger, G., von Friedeburg, C., and Platt, U.: Multi axis differential optical absorption spectroscopy (MAX-DOAS), Atmos. Chem.

Phys., 4, 231-254, doi:10.5194/acp-4-231-2004, 2004.

Hollingsworth, A. R., Engelen, R. J., Textor, C., Benedetti, A., Boucher, O., Chevallier, F., Dethof, A., Elbern, H., Eskes, H., Flemming, J., Granier, C., Kaiser, J. W., Morcrette, J.-J., Rayner, P., Peuch, V.-H., Rouil, L., Schultz, M. G., Simmons, A. J., and Consortium, T. G.: Toward a monitoring and forecasting system for atmospheric composition: The GEMS project, B. Am. Meteorol. Soc., 89, 1147–1164, 2008.

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.

Ialongo, I., Herman, J., Krotkov, N., Lamsal, L., Boersma, K. F., Hovila, J., and Tamminen, J.: Comparison of OMI NO_2 observations and their seasonal and weekly cycles with ground-based measurements in Helsinki, Atmos. Meas. Tech., 9, 5203-5212, https://doi.org/10.5194/amt-9-5203-2016, 2016.

Inness, A., Baier, F., Benedetti, A., Bouarar, I., Chabrillat, S., Clark, H., Clerbaux, C., Coheur, P., Engelen, R. J., Errera, Q., Flemming, J., George, M., Granier, C., Hadji-Lazaro, J., Huijnen, V., Hurtmans, D., Jones, L., Kaiser, J. W., Kapsomenakis, J., Lefever, K., Leitão, J., Razinger, M., Richter, A., Schultz, M. G., Simmons, A. J., Suttie, M., Stein, O., Thépaut, J.-N., Thouret, V., Vrekoussis, M., Zerefos, C., and the MACC team: The MACC reanalysis: an 8 yr data set of atmospheric composition, Atmos. Chem. Phys., 13, 4073-4109, doi:10.5194/acp-13-4073-2013, 2013.

Irie, H., Kanaya, Y., Akimoto, H., Tanimoto, H., Wang, Z., Gleason, J. F., and Bucsela, E. J.: Validation of OMI tropospheric NO_2 column data using MAX-DOAS measurements deep inside the North China Plain in June 2006: Mount Tai Experiment 2006, Atmos. Chem. Phys., 8, 6577-6586, doi:10.5194/acp-8-6577-2008, 2008.

Irie, H., Takashima, H., Kanaya, Y., Boersma, K. F., Gast, L., Wittrock, F., Brunner, D., Zhou, Y., and Van Roozendael, M.: Eight-component retrievals from ground-based MAX-DOAS observations, Atmos. Meas. Tech., 4, 1027–1044, doi:10.5194/amt-4-1027-2011, 2011.

Irie, H., Boersma, K. F., Kanaya, Y., Takashima, H., Pan, X., and Wang, Z. F.: Quantitative bias estimates for tropospheric NO_2 columns retrieved from SCIAMACHY, OMI, and GOME-2 using a common standard for East Asia, Atmos. Meas. Tech., 5, 2403-2411, doi:10.5194/amt-5-2403-2012, 2012.

Jacob, D. J.: Introduction of Atmospheric Chemistry, Princeton Univ. Press, Princeton, NJ, 234-243, 1999.

Josse, B., Simon, P., and Peuch, V.-H.: Radon global simulations with the multiscale chemistry and transport model MOCAGE, Tellus B, 56, 339–356. doi: 10.1111/j.1600-0889.2004.00112.x, 2004.

Kaiser, J. W., Heil, A., Andreae, M. O., Benedetti, A., Chubarova, N., Jones, L., Morcrette, J.-J., Razinger, M., Schultz, M. G., Suttie, M., and van der Werf, G. R.: Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power, Biogeosciences, 9, 527–554, doi:10.5194/bg-9-527-2012, 2012.

Kanaya, Y., Irie, H., Takashima, H., Iwabuchi, H., Akimoto, H., Sudo, K., Gu, M., Chong, J., Kim, Y. J., Lee, H., Li, A., Si, F., Xu, J., Xie, P.-H., Liu, W.-Q., Dzhola, A., Postylyakov, O., Ivanov, V., Grechko, E., Terpugova, S., and Panchenko, M.: Long-term MAX-DOAS network observations of NO_2 in Russia and Asia (MADRAS) during the period 2007–2012: instrumentation, elucidation of climatology, and comparisons with OMI satellite observations and global model simulations, Atmos. Chem. Phys., 14, 7909-7927, doi:https://doi.org/10.5194/acp-14-7909-2014, 2014.

Kramer, L. J., Leigh, R. J., Remedios, J. J. and Monks, P. S.: Comparison of OMI and ground-based in situ and MAX-DOAS measure-

ments of tropospheric nitrogen dioxide in an urban area, J. Geophys. Res., 113, D16S39, doi:10.1029/2007JD009168, 2008.

Kuenen, J. J. P., Denier van der Gon, H. A. C., Visschedijk, A., Van der Brugh, H., and Van Gijlswijk, R.: MACC European emission inventory for the years 2003–2007, TNO report TNO-060-UT-2011-00588, Utrecht, 2011.

Lacressonnière, G., Peuch, V.-H., Vautard, R., Arteta, J., Déqué, M., Josse, B., Marécal, V., and Saint-Martin, D.: European air quality in the 2030s and 2050s: Impacts of global and regional emission trends and of climate change, Atmos. Environ., 92, 348–358, 2014.

Langner, J., Engardt, M., Baklanov, A., Christensen, J. H., Gauss, M., Geels, C., Hedegaard, G. B., Nuterman, R., Simpson, D., Soares, J., Sofiev, M., Wind, P., and Zakey, A.: A multi-model study of impacts of climate change on surface ozone in Europe, Atmos. Chem. Phys., 12, 10423-10440, doi:10.5194/acp-12-10423-2012, 2012.

Lefèvre, F., Brasseur, G., Folkins, I., Smith, A., and Simon, P.: Chemistry of the 1991–1992 stratospheric winter: three-dimensional model simulations, J. Geophys. Res.-Atmos., 99, 8183–8195, 1994.

Leser, H., Hönninger, G. and Platt, U.: MAX-DOAS measurements of BrO and NO_2 in the marine boundary layer, Geophys. Res. Lett., 30(10), 1537, doi:10.1029/2002GL015811, 2003.

Levelt, P. F., Van den Oord, G. H. J., Dobber, M. R., Malkki, A., Visser, H., de Vries, J., Stammes, P., Lundell, J. O. V and Saari, H.: The Ozone Monitoring Instrument, Ieee Trans. Geosci. Remote Sens., 44(5), 1093–1101, doi:Urn:nbn:nl:ui:25-648485, 2006.

Lin, J.-T., Martin, R. V., Boersma, K. F., Sneep, M., Stammes, P., Spurr, R., Wang, P., Van Roozendael, M., Clémer, K., and Irie, H.: Retrieving tropospheric nitrogen dioxide from the Ozone Monitoring Instrument: effects of aerosols, surface reflectance anisotropy, and vertical profile of nitrogen dioxide, Atmos. Chem. Phys., 14, 1441-1461, doi:10.5194/acp-14-1441-2014, 2014.

Ma, J. Z., Beirle, S., Jin, J. L., Shaiganfar, R., Yan, P., and Wagner, T.: Tropospheric NO_2 vertical column densities over Beijing: results of the first three years of ground-based MAX-DOAS measurements (2008–2011) and satellite validation, Atmos. Chem. Phys., 13, 1547-1567, doi:10.5194/acp-13-1547-2013, 2013.

Mailler, S., Khvorostyanov, D., and Menut, L.: Impact of the vertical emission profiles on background gas-phase pollution simulated from the EMEP emissions over Europe, Atmos. Chem. Phys., 13, 5987-5998, doi:10.5194/acp-13-5987-2013, 2013.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier van der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadygrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouïl, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, Á., van Velthoven, P., van Versendaal, R., Vira, J., and Ung, A.: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production, Geosci. Model Dev., 8, 2777-2813, doi:10.5194/gmd-8-2777-2015, 2015.

Mendolia, D., D'Souza, R. J. C., Evans, G. J., and Brook, J.: Comparison of tropospheric NO_2 vertical columns in an urban environment using satellite, multi-axis differential optical absorption spectroscopy, and in situ measurements, Atmos. Meas. Tech., 6, 2907-2924, doi:10.5194/amt-6-2907-2013, 2013.

Menut, L., Bessagnet, B., Khvorostyanov, D., Beekmann, M., Blond, N., Colette, A., Coll, I., Curci, G., Foret, G., Hodzic, A., Mailler, S., Meleux, F., Monge, J. L., Pison, I., Siour, G., Turquety, S., Valari, M., Vautard, R., and Vivanco, M. G.: CHIMERE 2013: a model for regional atmospheric composition modelling, Geoscientific Model Development, 6, 981-1028, doi:10.5194/gmd-6-981-2013, 2013.

Mues, A., Kuenen, J., Hendriks, C., Manders, A., Segers, A., Scholz, Y., Hueglin, C., Builtjes, P., and Schaap, M.: Sensitivity of air pollution simulations with LOTOS-EUROS to the temporal distribution of anthropogenic emissions, Atmos. Chem. Phys., 14, 939-955,

doi:10.5194/acp-14-939-2014, 2014.

Peters, E., Wittrock, F., Großmann, K., Frieß, U., Richter, A., and Burrows, J. P.: Formaldehyde and nitrogen dioxide over the remote western Pacific Ocean: SCIAMACHY and GOME-2 validation using ship-based MAX-DOAS observations, Atmos. Chem. Phys., 12, 11179-11197, doi:10.5194/acp-12-11179-2012, 2012.

Petetin, H., Beekmann, M., Colomb, A., Denier van der Gon, H. A. C., Dupont, J.-C., Honoré, C., Michoud, V., Morille, Y., Perrussel, O., Schwarzenboeck, A., Sciare, J., Wiedensohler, A., and Zhang, Q. J.: Evaluating BC and NOx emission inventories for the Paris region from MEGAPOLI aircraft measurements, Atmos. Chem. Phys., 15, 9799-9818, doi:10.5194/acp-15-9799-2015, 2015.

Pinardi, G., Van Roozendael, M., Lambert, J.-C., Granville, J., Hendrick, F., Tack, F., Yu, H., Cede, A., Kanaya, Y., Irie, H., Goutail, F., Pommereau, J.-P., Pazmino, A., Wittrock, F., Richter, A., Wagner, T., Gu, M., Remmers, J., Friess, U., Vlemmix, T., Piters, A., Hao, N., Tiefengraber, M., Herman, J., Abuhassan, N., Bais, A., Kouremeti, N., Hovila, J., Holla, R., Chong, J., Postylyakov, O. and Ma, J.: GOME-2 total and tropospheric NO_2 validation based on zenith-sky, direct-sun and Multi-Axis DOAS network observations, EUMETSAT conference, Geneva, Switzerland, 22-26 September 2014.

Piters, A. J. M., Boersma, K. F., Kroon, M., Hains, J. C., Van Roozendael, M., Wittrock, F., Abuhassan, N., Adams, C., Akrami, M., Allaart, M. A. F., Apituley, A., Beirle, S., Bergwerff, J. B., Berkhout, A. J. C., Brunner, D., Cede, A., Chong, J., Clémer, K., Fayt, C., Frieß, U., Gast, L. F. L., Gil-Ojeda, M., Goutail, F., Graves, R., Griesfeller, A., Großmann, K., Hemerijckx, G., Hendrick, F., Henzing, B., Herman, J., Hermans, C., Hoexum, M., van der Hoff, G. R., Irie, H., Johnston, P. V., Kanaya, Y., Kim, Y. J., Klein Baltink, H., Kreher, K., de Leeuw, G., Leigh, R., Merlaud, A., Moerman, M. M., Monks, P. S., Mount, G. H., Navarro-Comas, M., Oetjen, H., Pazmino, A., Perez-Camacho, M., Peters, E., du Piesanie, A., Pinardi, G., Puentedura, O., Richter, A., Roscoe, H. K., Schönhardt, A., Schwarzenbach, B., Shaiganfar, R., Sluis, W., Spinei, E., Stolk, A. P., Strong, K., Swart, D. P. J., Takashima, H., Vlemmix, T., Vrekoussis, M., Wagner, T., Whyte, C., Wilson, K. M., Yela, M., Yilmaz, S., Zieger, P., and Zhou, Y.: The Cabauw Intercomparison campaign for Nitrogen Dioxide measuring Instruments (CINDI): design, execution, and early results, Atmos. Meas. Tech., 5, 457-485, doi:10.5194/amt-5-457-2012, 2012.

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.

Rodgers, C. D.: Inverse Methods for Atmospheric Sounding – Theory and Practice, Series on Atmospheric, Oceanic and Planetary Physics, World Scientific, Singapore, 2000.

Rozanov, A., Rozanov, V., Buchwitz, M., Kokhanovsky, A., and Burrows, J. P.: SCIATRAN 2.0 – a new radiative transfer model for geophysical applications in the 175–2400 nm spectral region, Adv. Space Res., 36, 1015–1019, 2005.

Roscoe, H. K., Van Roozendael, M., Fayt, C., du Piesanie, A., Abuhassan, N., Adams, C., Akrami, M., Cede, A., Chong, J., Clémer, K., Frieß, U., Gil Ojeda, M., Goutail, F., Graves, R., Griesfeller, A., Grossmann, K., Hemerijckx, G., Hendrick, F., Herman, J., Hermans, C., Irie, H., Johnston, P. V., Kanaya, Y., Kreher, K., Leigh, R., Merlaud, A., Mount, G. H., Navarro, M., Oetjen, H., Pazmino, A., Perez-Camacho, M., Peters, E., Pinardi, G., Puentedura, O., Richter, A., Schönhardt, A., Shaiganfar, R., Spinei, E., Strong, K., Takashima, H., Vlemmix, T., Vrekoussis, M., Wagner, T., Wittrock, F., Yela, M., Yilmaz, S., Boersma, F., Hains, J., Kroon, M., Piters, A., and Kim, Y. J.: Intercomparison of slant column measurements of NO_2 and O_4 by MAX-DOAS and zenith-sky UV and visible spectrometers, Atmos. Meas. Tech., 3, 1629–1646, doi:10.5194/amt-3-1629-2010, 2010.

Schaap, M., Timmermans, R. M. A., Roemer, M., Boersen, G. A. C., Builtjes, P. J. H., Sauter, F. J., Velders, G. J. M., Beck, J .P: The LOTOS-EUROS Model: Description, validation and latest developments. Int. J. Environ. Pollut., 32, 270–290, 2008.

Schaap, M., Kranenburg, R., Curier, L., Jozwicka, M., Dammers, E., and Timmermans, R.: Assessing the Sensitivity of the OMI-NO_2

Product to Emission Changes across Europe, Remote Sensing, 5(9), 4187-4208, doi:10.3390/rs5094187, 2013.

Schmidt, H., Derognat, C., Vautard, R., and Beekmann, M.: A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in western Europe, Atmos. Environ., 35, 6277–6297, 2001.

Schulz, M., Benedictow, A., Schneider, P., Bartnicki, J., Valdebenito, Á., Gauss, M. and Griesfeller, J.: Modelling and evaluation of trends in the EMEP framework, Transboundary acidification, eutrophication and ground level ozone in Europe in 2011, EMEP Status Report 1/2013, The Norwegian Meteorological Institute, Oslo, Norway, 2013.

Shaiganfar, R., Beirle, S., Petetin, H., Zhang, Q., Beekmann, M., and Wagner, T.: New concepts for the comparison of tropospheric NO_2 column densities derived from car-MAX-DOAS observations, OMI satellite observations and the regional model CHIMERE during two MEGAPOLI campaigns in Paris 2009/10, Atmos. Meas. Tech., 8, 2827-2852, doi:10.5194/amt-8-2827-2015, 2015.

Simpson, D., Benedictow, A., Berge, H., Bergström, R., Emberson, L. D., Fagerli, H., Flechard, C. R., Hayman, G. D., Gauss, M., Jonson, J. E., Jenkin, M. E., Nyíri, A., Richter, C., Semeena, V. S., Tsyro, S., Tuovinen, J.-P., Valdebenito, Á., and Wind, P.: The EMEP MSC-W chemical transport model – technical description, Atmos. Chem. Phys., 12, 7825-7865, doi:10.5194/acp-12-7825-2012, 2012.

Simpson, D., Fagerli, H., Jonson, J., Tsyro, S., Wind, P., and Tuovinen, J.-P.: The EMEP Unified Eulerian Model. Model Description, EMEP MSC-W Report 1/2003, The Norwegian Meteorological Institute, Oslo, Norway, 2003.

Sofiev, M.: A model for the evaluation of long-term airborne pollution transport at regional and continental scales, Atmos. Env., 34(15), 24812493, doi:10.1016/S1352-2310(99)00415-X, 2000.

Sofiev, M., Siljamo, P., Valkama, I., Ilvonen, M., and Kukkonen, J.:. A dispersion modelling system SILAM and its evaluation against ETEX data, Atmos. Environ. 40, 674–685, doi:10.1016/j.atmosenv.2005.09.069, 2006.

Sofiev, M., Vira, J., Kouznetsov, R., Prank, M., Soares, J., and Genikhovich, E.: Construction of an Eulerian atmospheric dispersion model based on the advection algorithm of M . Galperin: dynamic cores, Geosci. Model Dev. Discuss., 8, 2905–2947, doi:10.5194/gmdd-8-2905-2015, 2015.

Solazzo, E., Bianconi, R., Vautard, R., Appel, K. W., Moran, M. D., Hogrefe, C., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Denier van der Gon, H., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Jeričević, A., Kraljević, L., Miranda, A. I., Nopmongcol, U., Pirovano, G., Prank, :, Riccio, A., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke , R., Yarwood, G., Zhang, J., Rao, S. T., and Galmarini, S.: Model Evaluation and Ensemble Modeling of Surface-Level Ozone in Europe and North America in the Context of the AQMEII, Atmos. Environ., 53, 60-74, 2012.

Solomon, S.: Stratospheric ozone depletion: a review of concepts and history, Rev. Geophys., 37, 275–316, 1999.

Stammes, P.: Spectral radiance modeling in the UV-visible range, in: IRS 2000: Current Problems in Atmospheric Radiation, edited by: Smith, W. and Timofeyev, Y. A. Deepak, Hampton, Va, 385–388, 2001.

Stockwell, W., Kirchner, F., Kuhn, M., and Seefeld, S.: A new mechanism for regional atmospheric chemistry modeling, J. Geophys. Res., 102, 25847–25879, 1997.

Stohl, A., Huntrieser, H., Richter, A., Beirle, S., Cooper, O. R., Eckhardt, S., Forster, C., James, P., Spichtinger, N., and Wenig, M.: Rapid intercontinental air pollution transport associated with a meteorological bomb, Atmos. Chem. Phys., 3, 969–985, http://www.atmos-chem-phys.net/3/969/2003/, 2003.

Takashima, H., Irie, H., Kanaya, Y., and Syamsudin, F.: NO_2 observations over the western Pacific and Indian Ocean by MAX-DOAS on Kaiyo, a Japanese research vessel, Atmos. Meas. Tech., 5, 2351-2360, doi:10.5194/amt-5-2351-2012, 2012.

Thunis, P., Cuvelier, C., Roberts, P., White, L., Stern, R., Kerschbaumer, A.,Bessagnet, B., Bergström, R., and Schaap, M.: EURODELTA: Evaluation of a Sectoral Approach to Integrated Assessment Modelling- Second Report. In: EUR -Scientific and Technical Research Se-

ries -24474 EN-2010, Publications Office of the European Union, Luxembourg, ISSN 1018-5593, 2010.

Valks, P., Pinardi, G., Richter, A., Lambert, J.-C., Hao, N., Loyola, D., Van Roozendael, M., and Emmadi, S.: Operational total and tropospheric NO_2 column retrieval for GOME-2, Atmos. Meas. Tech., 4, 1491-1514, doi:10.5194/amt-4-1491-2011, 2011.

Vautard, R., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., Builtjes, P. J. H., Christensen, J. H., Cuvelier, K., Foltescu, V., Graff, A., Kerschbaumer, A., Krol, M., Roberts, P., Rouil, L., Stern, R., Tarrasón, L., Thunis, P., Vignati, E., Wind, P.: Skill and uncertainty of a regional air quality model ensemble, Atmos. Env. 43, 4822-4832, doi:10.1016/j.atmosenv.2008.09.083, 2009.

Vira, J. and Sofiev, M.: Assimilation of surface NO_2 and O_3 observations into the SILAM chemistry transport model. Geosci. Model Dev. 8, 191–203, doi:10.5194/gmd-8-191-2015, 2015.

Vlemmix, T., Eskes, H. J., Piters, A. J. M., Schaap, M., Sauter, F. J., Kelder, H., and Levelt, P. F.: MAX-DOAS tropospheric nitrogen dioxide column measurements compared with the Lotos-Euros air quality model, Atmos. Chem. Phys., 15, 1313-1330, doi:10.5194/acp-15-1313-2015, 2015.

Wagner, T., Beirle, S., Brauers, T., Deutschmann, T., Frieß, U., Hak, C., Halla, J. D., Heue, K. P., Junkermann, W., Li, X., Platt, U., and Pundt-Gruber, I.: Inversion of tropospheric profiles of aerosol extinction and HCHO and NO_2 mixing ratios from MAX-DOAS observations in Milano during the summer of 2003 and comparison with independent data sets, Atmos. Meas. Tech., 4, 2685–2715, doi:10.5194/amt-4-2685-2011, 2011.

Wang, T., Hendrick, F., Wang, P., Tang, G., Clémer, K., Yu, H., Fayt, C., Hermans, C., Gielen, C., Pinardi, G., Theys, N., Brenot, H., and Van Roozendael, M.: Evaluation of tropospheric SO2 retrieved from MAX-DOAS measurements in Xianghe, China, Atmos. Chem. Phys. Discuss., 14, 6501-6536, doi:10.5194/acpd-14-6501-2014, 2014.

Watson, L. , Lacressonnière, G., Gauss, M., Engardt, M., Andersson, C., Josse, B., Marécal, V., Nyiri, A., Sobolowski, S., Siour, G., Szopa, S., Vautard, R.: The impact of emissions and +2◦ C climate change upon future ozone and nitrogen dioxide over Europe, Atmos. Env., 142, 271-285, 2016.

Wennberg, P. O., Cohen, R. C., Stimpfle, R. M., Koplow, J. P., Anderson, J. G., Salawitch, R. J., Fahey, D. W., Woodbridge, E. L., Keim, E. R., Gao, R. S., Webster, C. R., May, R. D., Toohey, D., Avallone, L., Proffitt, M. H., Loewenstein, M., Podolske, J. R., Chan, K. R., and Wofsy, S. C.: Removal of stratospheric O_3 by radicals: in situ measurements of OH, HO2 , NO, NO_2, ClO and BrO, Science, 266, 398–404, 1994.

Whitten, G. Z., Hogo, H., and Killus, J. P.: The Carbon-Bond Mechanism : A Condensed Kinetic Mechanism for Photochemical Smog, Environ. Sci., 14 (6), 690–700, doi: 10.1021/es60166a008, 1980.

Wiedinmyer, C., Akagi, S. K., Yokelson, R. J., Emmons, L. K., Al-Saadi, J. A., Orlando, J. J., and Soja, A. J.: The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning, Geosci. Model Dev., 4, 625-641, doi:10.5194/gmd-4-625-2011, 2011.

Wittrock, F., Oetjen, H., Richter, A., Fietkau, S., Medeke, T., Rozanov, A., and Burrows, J. P.: MAX-DOAS measurements of atmospheric trace gases in Ny-Ålesund - Radiative transfer studies and their application, Atmos. Chem. Phys., 4, 955-966, 2004.

Wittrock, F.: The retrieval of oxygenated volatile organic compounds by remote sensing techniques, Ph.D., University of Bremen, Bremen, Germany, available at: http://www.doas-bremen.de/paper/diss wittrock 06.pdf, 2006.

Zien, A. W., Richter, A., Hilboll, A., Blechschmidt, A.-M., and Burrows, J. P.: Systematic analysis of tropospheric NO_2 long-range transport events detected in GOME-2 satellite data, Atmos. Chem. Phys., 14, 7367-7396, doi:10.5194/acp-14-7367-2014, 2014.

Zyryanov, D., Foret, G., Eremenko, M., Beekmann, M., Cammas, J.-P., D'Isidoro, M., Elbern, H., Flemming, J., Friese, E., Kioutsioutkis, I., Maurizi, A., Melas, D., Meleux, F., Menut, L., Moinat, P., Peuch, V.-H., Poupkou, A., Razinger, M., Schultz, M., Stein, O., Suttie, A.

M., Valdebenito, A., Zerefos, C., Dufour, G., Bergametti, G., and Flaud, J.-M.: 3-D evaluation of tropospheric ozone simulations by an ensemble of regional Chemistry Transport Model, Atmos. Chem. Phys., 12, 3219–3240, doi:10.5194/acp-12-3219-2012, 2012.

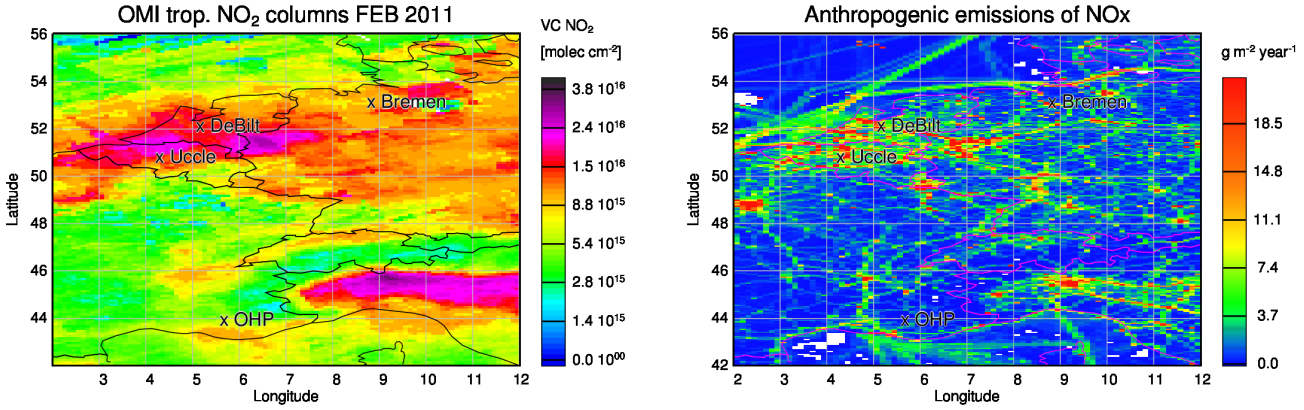

Figure 1. Maps of (left) average tropospheric NO_2 VCDs [molec cm^{-2}] observed by OMI for February 2011 and (right) TNO/MACC-II anthropogenic NOx emissions [g m^{-2} $year^{-1}$] over Europe. Location of MAX-DOAS measurement sites investigated in this study are marked by black crosses on the maps. The satellite data has been gridded to 0.1° lat x 0.1° lon, the resolution of the emission database is 0.125° lat x 0.0625° lon.

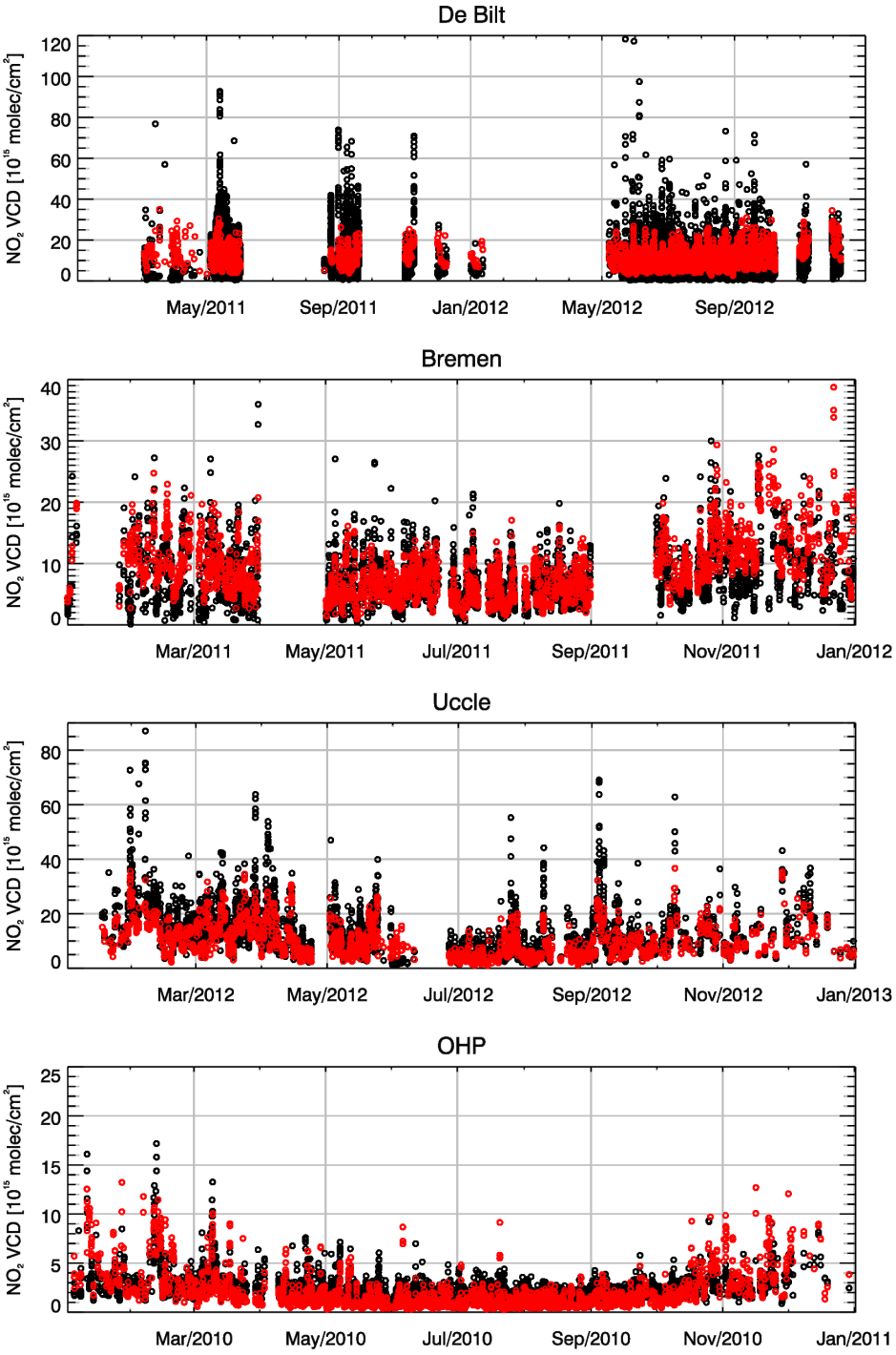

Figure 2. Time series of AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] from (black circles) MAX-DOAS and (colored circles) model ensemble hourly data for (from top to bottom) De Bilt, Bremen, Uccle and OHP.

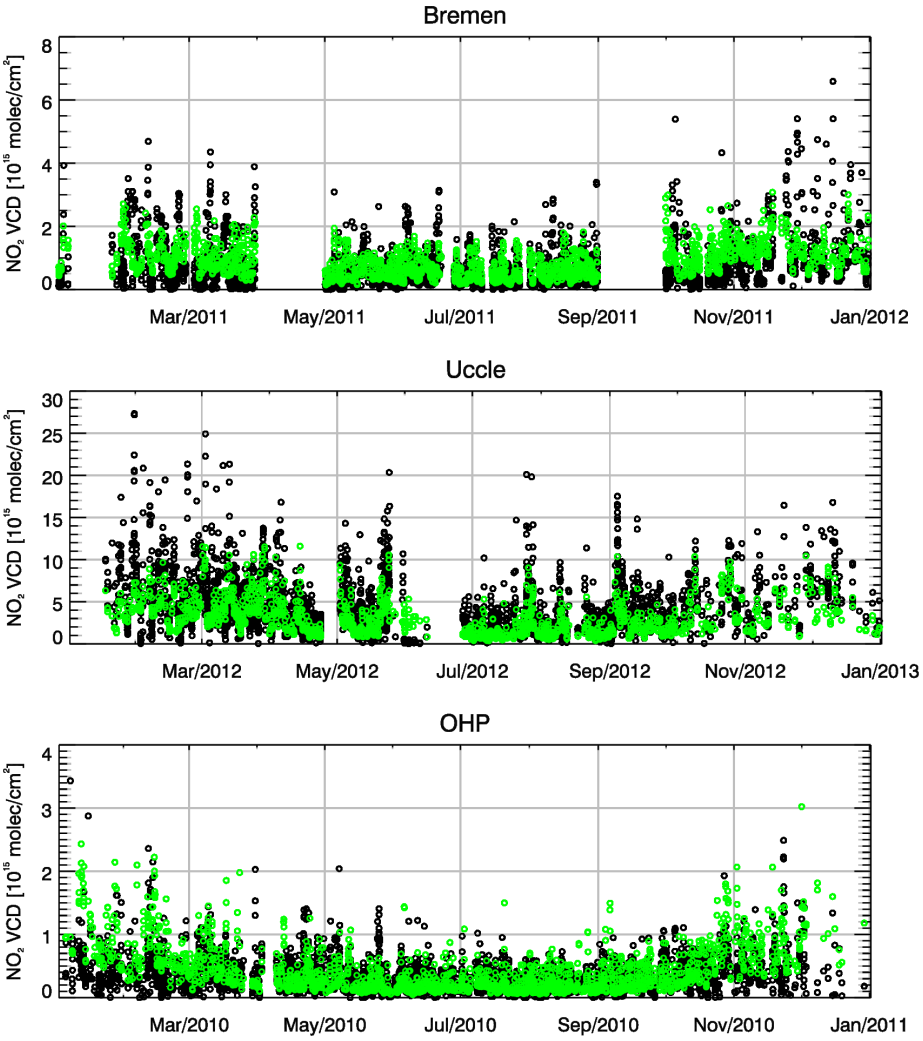

Figure 3. As in Figure 2 but for NO_2 surface partial columns [10^{15} molec cm^{-2}]. Surface partial columns from MAX-DOAS are not available for De Bilt for the investigated time period.

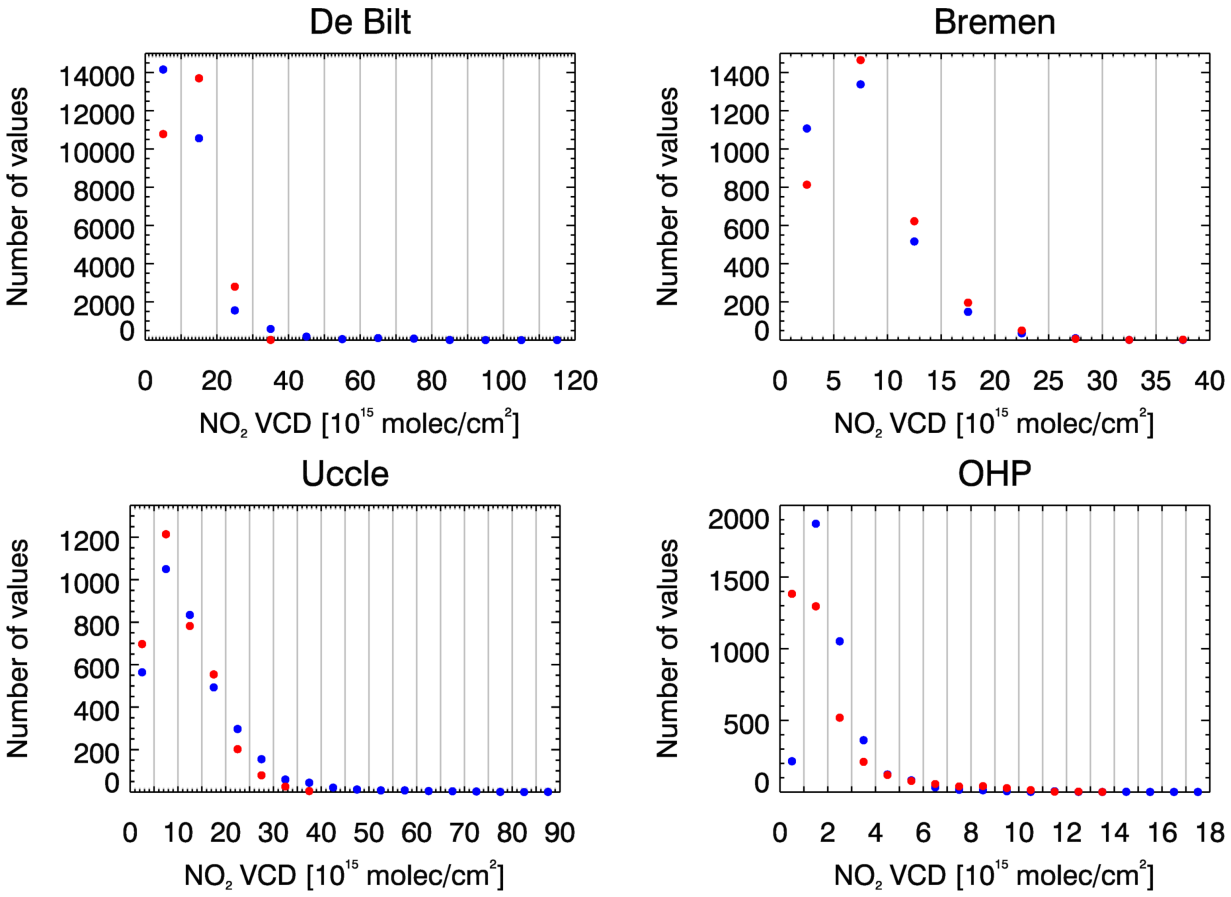

Figure 4. Frequency distributions of AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] from (blue) MAX-DOAS and (red) model ensemble data for (top left) De Bilt, (top right) Bremen, (lower left) Uccle and (lower right) OHP. The distance between vertical grey lines on the x-axis corresponds to the size of the bins used to calculate the number of values given on the y-axis.

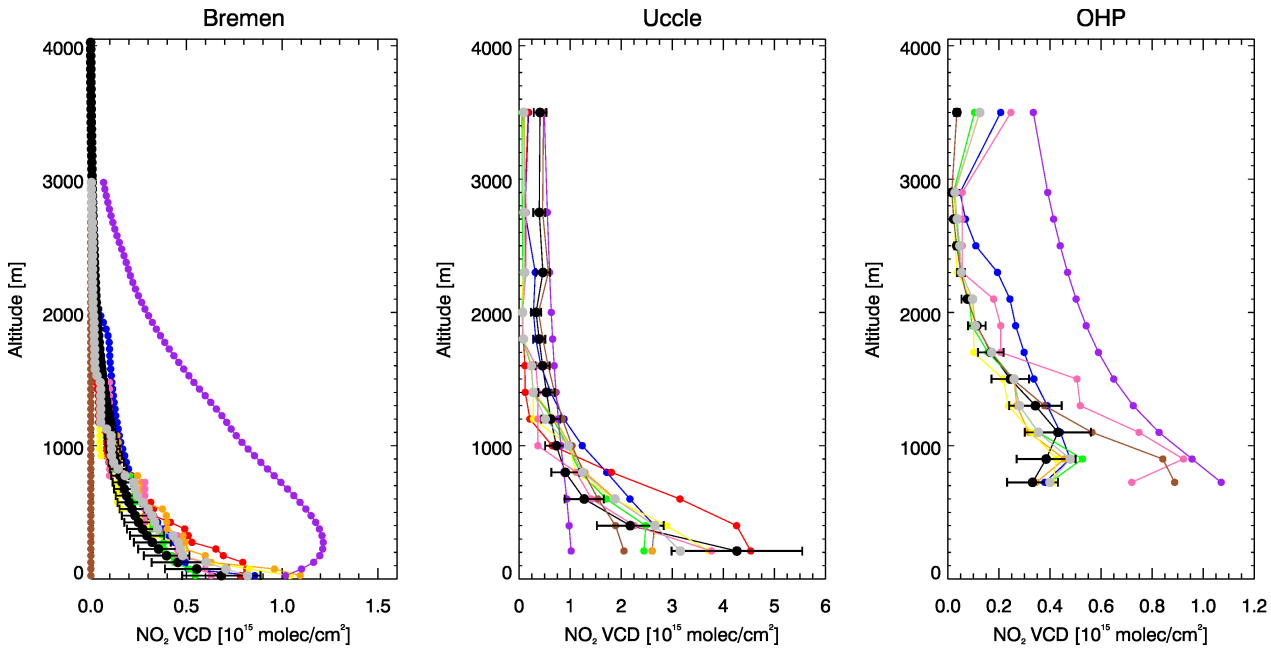

Figure 5. Average vertical profiles of NO_2 partial columns [10^{15} molec cm^{-2}] from (black) MAX-DOAS, (brown) a priori used for MAX-DOAS retrievals, (gray) model ensemble median, (blue) LOTOS-EUROS, (yellow) CHIMERE, (green) EMEP, (orange) EMEP-MACCEVA, (pink) SILAM and (red) MOCAGE as well as (purple) column averaging kernels [unitless] for (left) Bremen, (middle) Uccle and (right) OHP. Black error bars refer to the uncertainty associated with the MAX-DOAS retrievals (assumed to be 30 % for all stations). MAX-DOAS vertical profiles are not available for De Bilt for the investigated time period.

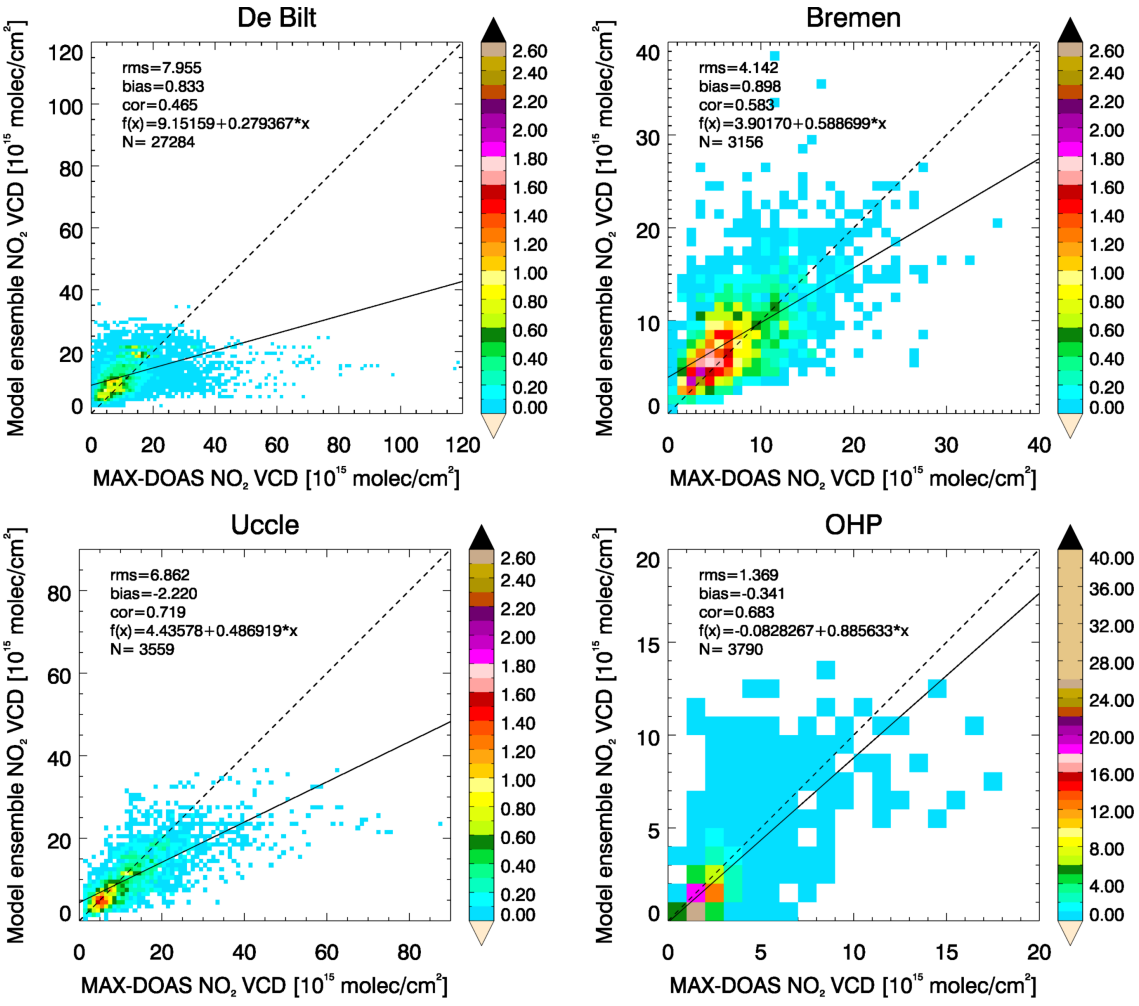

Figure 6. Scatter density plots of AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] from MAX-DOAS against model ensemble data for (top left) De Bilt, (top right) Bremen, (lower left) Uccle and (lower right) OHP. The data is shown for different bins with a size of 10^{15} molec cm^{-2} and is colored according to the number of data points per bin [%]. The dashed line is the reference line (f(x)=x). The solid line is the regression line (see top left of each plot for f(x) of this line). The root mean squared error (rms) [10^{15} molec cm^{-2}], bias [10^{15} molec cm^{-2}], Pearson correlation coefficient (cor, not squared) as well as the number of data points N are given at the top left of each plot.

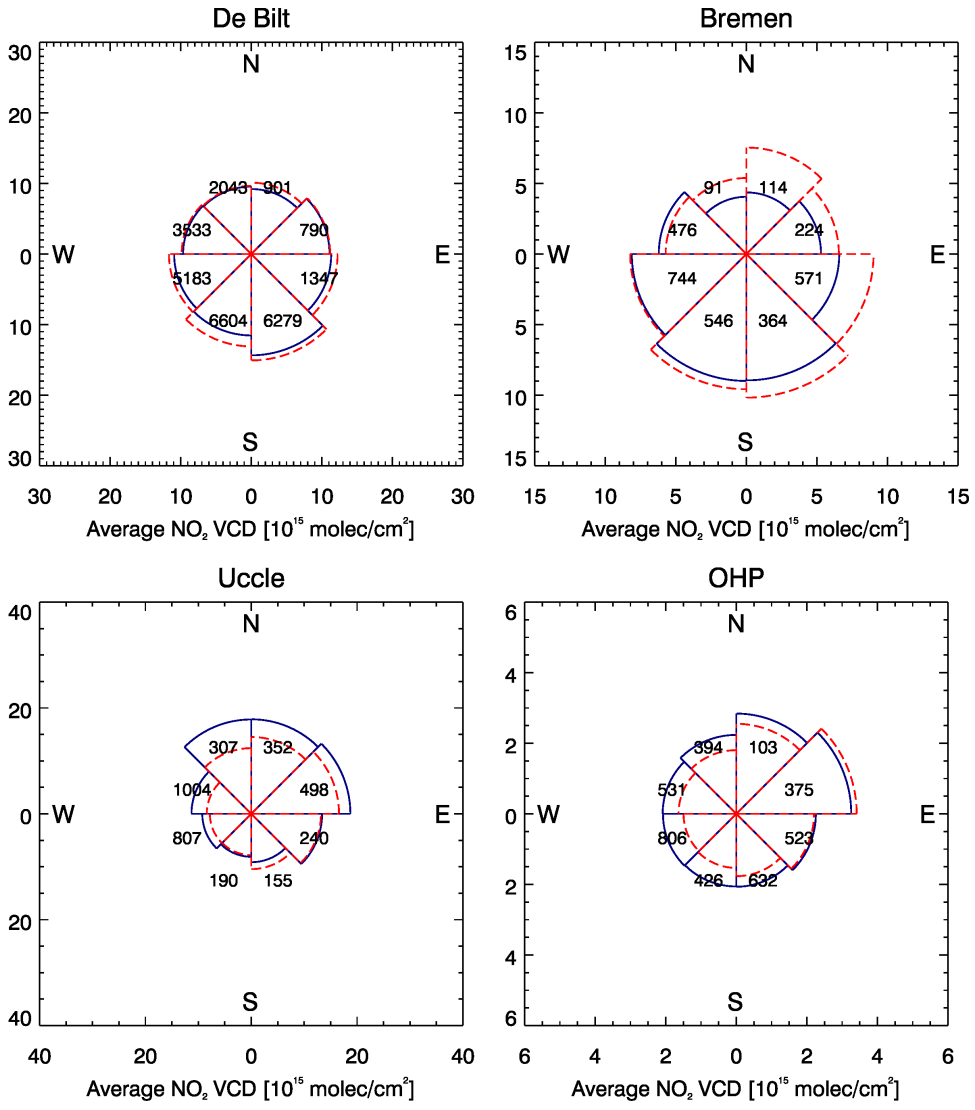

Figure 7. (a) Average AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] in 45° wide wind direction bins from (blue solid lines) MAX-DOAS and (red dashed lines) model ensemble data for (top left) De Bilt, (top right) Bremen, (lower left) Uccle and (lower right) OHP. Wind directions correspond to the direction towards the station and are taken from weather station measurements. The numbers close to the centre of each plot refer to the number of data values used for calculating average values for each bin.

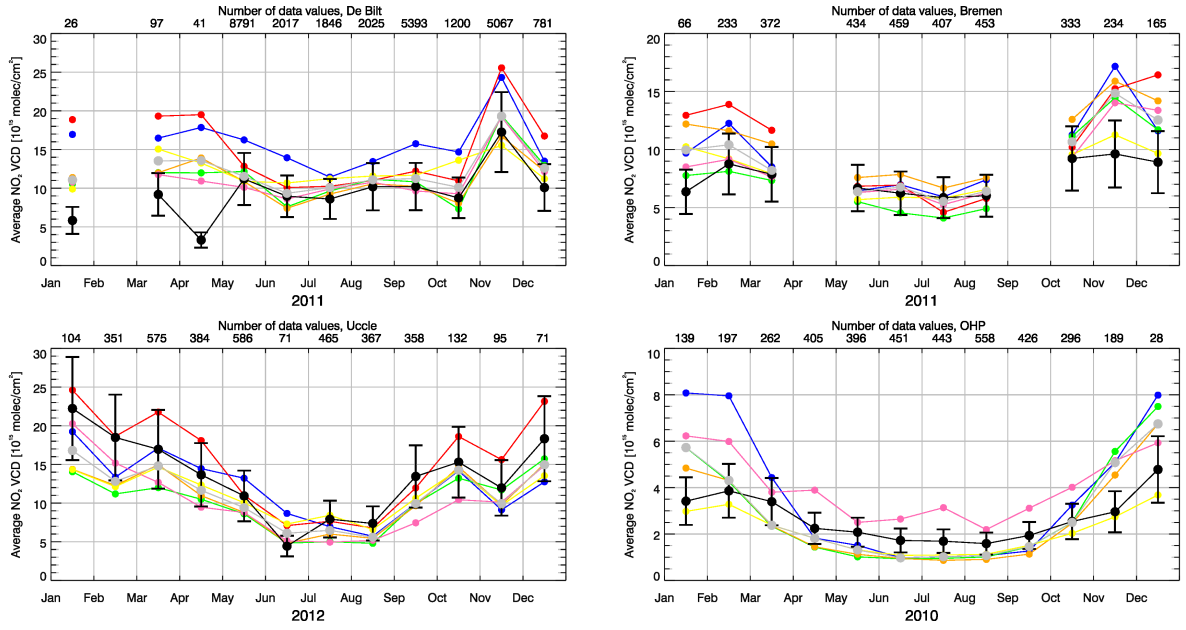

Figure 8. Seasonal cycles (monthly averages) of AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] from (black) MAX-DOAS, (gray) model ensemble median, (blue) LOTOS-EUROS, (yellow) CHIMERE, (green) EMEP, (orange) EMEP-MACCEVA, (pink) SILAM and (red) MOCAGE for (top left) De Bilt, (top right) Bremen, (lower left) Uccle and (lower right) OHP. Black error bars refer to the uncertainty associated with the MAX-DOAS retrievals (assumed to be 30 % for all stations). The number of data values used for calculating average values is shown at the upper x-axis of each plot.

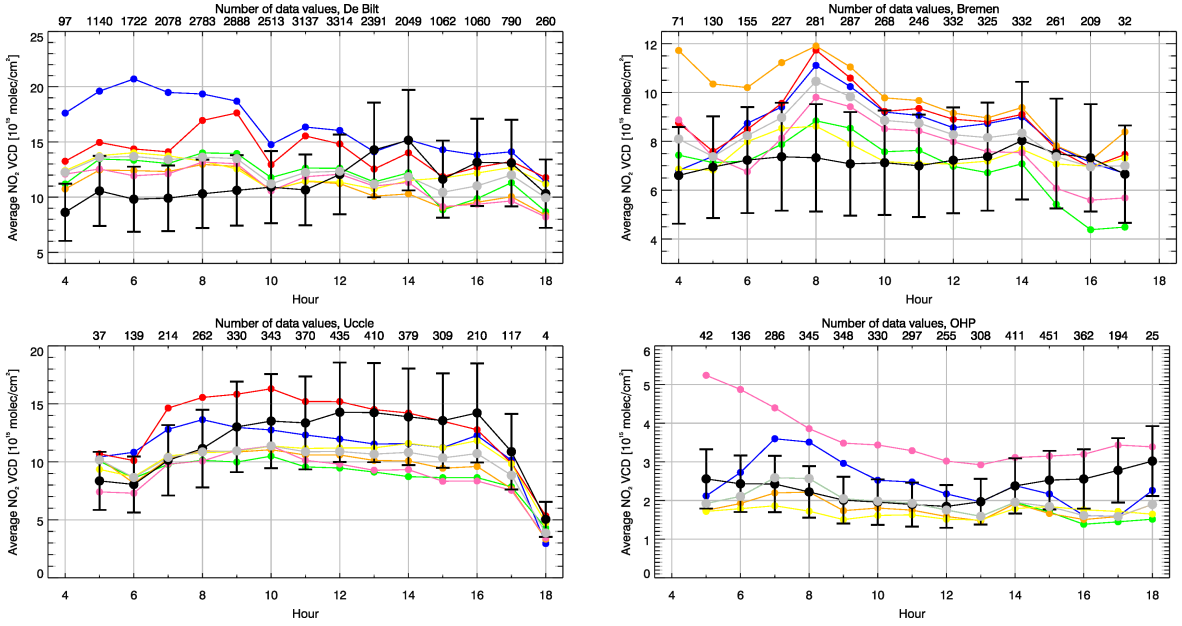

Figure 9. As in Figure 8 but for diurnal cycles (averages over hourly bins) of tropospheric NO_2 VCDs [10^{15} molec cm^{-2}].

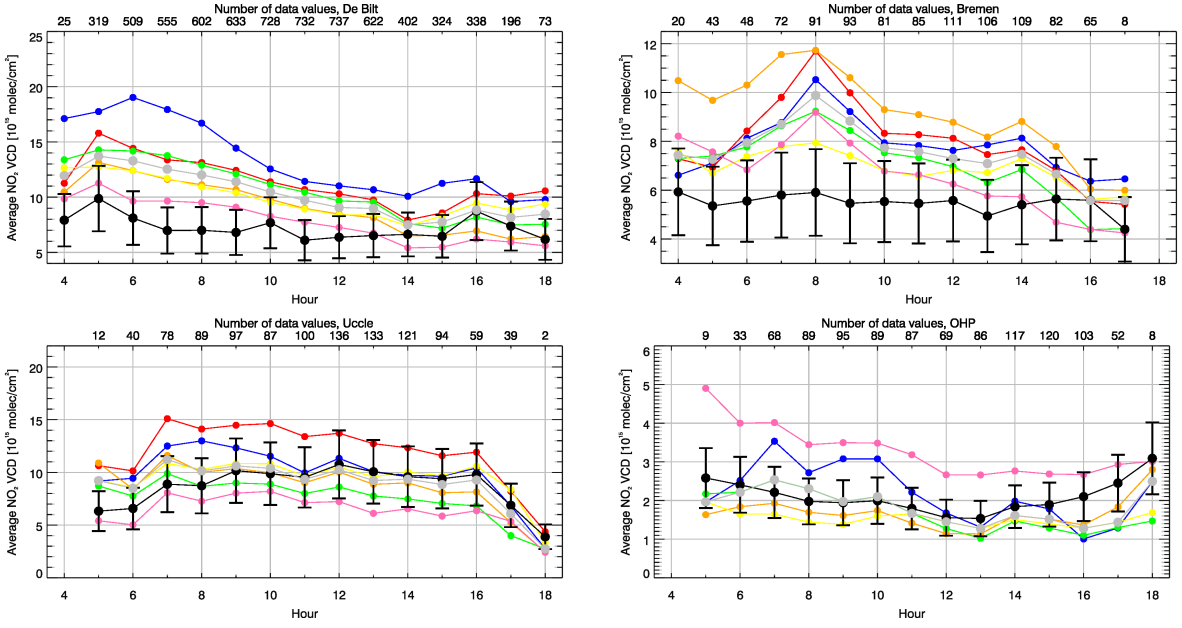

Figure 10. As in Figure 8 but for diurnal cycles (averages over hourly bins) of tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] during weekends only.

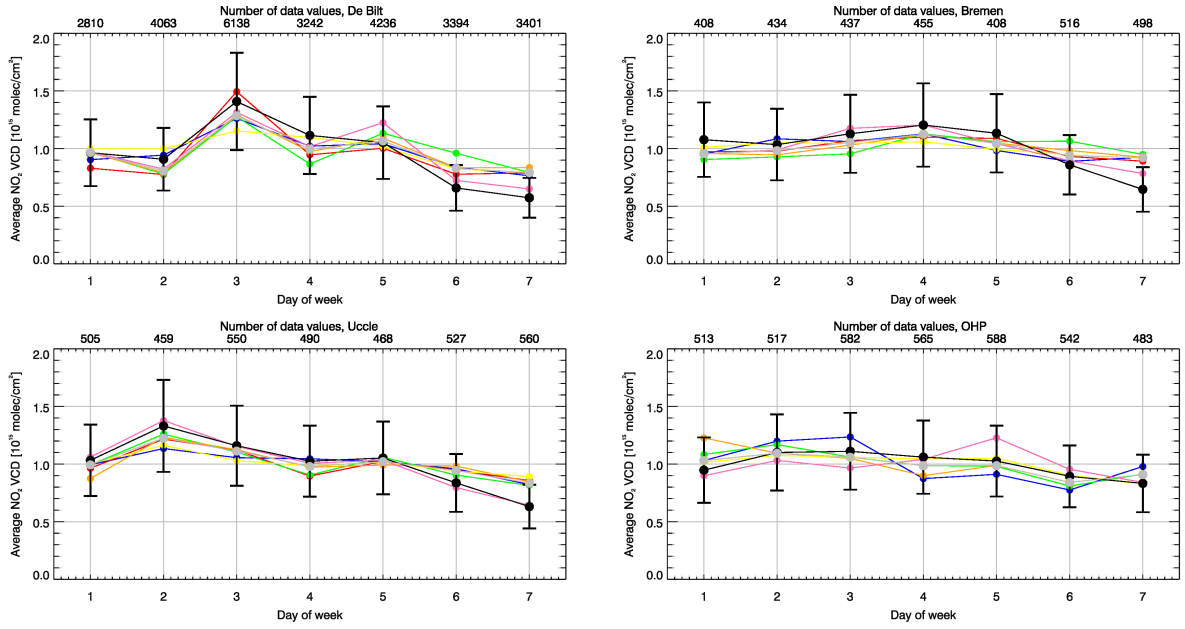

Figure 11. As in Figure 8 but for weekly cycles (averages over daily bins devided by mean over whole week, unitless values) of tropospheric NO_2 VCDs.

Table 1. Overview of regional air quality model simulations.

Model	**Institution**	**Grid spacing (zonal x meridional)**	**Number of vertical levels model top**	**Chemistry scheme**
CHIMERE	LISA-CNRS/UPEC/UPD INERIS	0.25° x 0.25° ($\sim$ 18 x 28 km^2)	8 500 hPa	MELCHIOR II (Schmidt et al., 2001)
EMEP-MACCEVA	MetNo	0.25° x 0.125° ($\sim$ 18 x 14 km^2)	20 100 hPa	EMEP-EmChem09soa (Simpson et al., 2012; Bergström et al., 2012)
EMEP	MetNo	50 x 50 km^2	20 100 hPa	EMEP-EmChem09soa (Simpson et al., 2012 Bergström et al., 2012)
LOTOS-EUROS	TNO	0.125° x 0.0625° ($\sim$ 9 x 7 km^2)	3 $\sim$ 3.5 km	TNO CBM-IV (Schaap et al., 2008; Whitten et al., 1980)
MOCAGE	CNRS-Météo-France	0.2° x 0.2° ($\sim$ 15 x 22 km^2)	47 5 hPa	troposphere: RACM (Stockwell et al., 1997) stratosphere REPROBUS (Lefèvre et al., 1994)
SILAM	FMI	years 2010/2011: 0.2° x 0.2° ($\sim$ 15 x 22 km^2) year 2012: 0.15° x 0.15° ($\sim$ 11 x 17 km^2)	9 (2010), 8 (2011-2012) 6.725 km (2010), 6.7 km (2011-2012)	DMAT (Sofiev, 2000)

Table 2. Overview of MAX-DOAS station data.

Station location	lat, lon height [masl]	Institution	Time period	Type	Retrieved quantitiy	number of layers layer top [km]	additional data
De Bilt Netherlands	52.1° N, 5.18° E $\sim$ 23 m	KNMI	03/2011–12/2012	urban	column	12 4.0 km	wind data (in-situ)
Bremen Germany	53.11° N, 8.86° E 21 m	IUP-UB	01/2011–12/2011	urban	column profile	81 4.025 km	wind data (in-situ data from airport weather station $\sim$ 9 km southwards at 53.05° N, 8.79° E)
Uccle Belgium	50.8° N, 4.32° E 120 m	BIRA-IASB	01/2012–12/2012	urban	column profile	13 3.5 km	wind data (in-situ) clouds from MAX-DOAS
OHP France	43.92° N, 5.7° E 650 m	BIRA-IASB	01/2010-12/2010	rural	column profile	13 3.5 km	wind data (in-situ)

Table 3. Statistics on how AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] from regional models compare to MAX-DOAS retrievals at the four MAX-DOAS stations. Each column entry shows from left to right: root mean squared error [10^{15} molec cm^{-2}], bias [10^{15} molec cm^{-2}] and Pearson correlation coefficient. MOCAGE data is not available for the measurement time period at OHP.

	De Bilt			**Bremen**			**Uccle**			**OHP**		
	rms	bias	r	rms	bias	r	rms	bias	r	rms	bias	r
ENSEMBLE	7.955	0.833	0.465	4.142	0.898	0.583	6.862	-2.220	0.719	1.369	-0.341	0.683
LOTOS-EUROS	10.516	5.254	0.352	5.180	1.598	0.461	7.815	-0.897	0.598	3.004	0.217	0.666
CHIMERE	8.221	0.573	0.402	3.960	0.198	0.533	7.059	-1.950	0.686	1.269	-0.563	0.627
EMEP	8.680	0.919	0.393	4.558	-0.167	0.521	8.134	-3.609	0.624	1.824	-0.338	0.600
EMEP-MACCEVA	8.340	-0.182	0.397	5.308	2.427	0.554	7.591	-2.777	0.659	2.012	-0.473	0.532
SILAM	8.516	0.115	0.408	4.506	0.570	0.550	7.985	-3.385	0.633	2.577	1.195	0.482
MOCAGE	9.731	3.082	0.427	5.651	1.801	0.520	7.413	1.476	0.692			

Table 4. As in Table 3 but for NO_2 surface partial columns [10^{15} molec cm^{-2}]. Surface partial columns from MAX-DOAS are not available for De Bilt for the investigated time period.

	Bremen			**Uccle**			**OHP**		
	rms	bias	r	rms	bias	r	rms	bias	r
ENSEMBLE	0.715	0.123	0.374	2.905	-1.124	0.586	0.351	0.058	0.439
LOTOS-EUROS	0.783	0.181	0.336	3.309	-1.659	0.509	0.517	0.048	0.393
CHIMERE	0.927	0.364	0.252	2.902	-0.554	0.531	0.337	0.081	0.400
EMEP	0.723	-0.133	0.318	3.330	-1.819	0.533	0.443	0.068	0.417
EMEP-MACCEVA	0.869	0.414	0.320	3.229	-1.658	0.548	0.428	0.014	0.344
SILAM	0.681	-0.054	0.397	2.910	-0.498	0.572	0.659	0.388	0.318
MOCAGE	0.750	0.101	0.372	2.886	0.272	0.596			

Table 5. As in Table 3 but for (upper rows) seasonal cycles, (middle rows) diurnal cycles and (lower rows) weekly cycles of AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}]. In addition, values for diurnal cycles are given based on data during weekdays only and during weekends only for the ensemble.

		De Bilt			**Bremen**			**Uccle**			**OHP**		
		rms	bias	r	rms	bias	r	rms	bias	r	rms	bias	r
seasonal	ENSEMBLE	3.750	2.494	0.511	2.218	1.347	0.710	2.857	-2.074	0.863	1.055	0.018	0.755
	LOTOS-EUROS	6.982	5.920	0.413	2.808	1.807	0.708	2.648	-0.892	0.777	1.975	0.752	0.859
	CHIMERE	4.001	2.592	0.343	1.291	0.519	0.627	3.337	-1.853	0.810	0.563	-0.507	0.878
	EMEP	3.237	1.837	0.575	1.960	0.337	0.743	4.000	-3.071	0.820	1.187	-0.020	0.733
	EMEP-MACCEVA	3.647	1.587	0.304	3.344	2.595	0.702	3.455	-2.581	0.832	0.903	-0.208	0.795
	SILAM	2.904	1.631	0.617	1.933	0.905	0.688	3.323	-2.759	0.853	1.506	1.270	0.798
	MOCAGE	7.443	5.315	0.214	3.823	2.414	0.643	2.621	1.594	0.853			
diurnal	ENSEMBLE	2.571	0.749	-0.389	1.499	1.134	0.148	2.401	-1.735	0.815	0.599	-0.371	-0.069
	ENSEMBLE (weekdays)	2.762	0.103	-0.237	1.343	0.752	0.125	3.181	-2.527	0.827	0.639	-0.390	-0.251
	ENSEMBLE (weekends)	3.438	3.006	0.559	2.664	2.463	0.405	1.294	0.341	0.806	0.461	-0.253	0.567
	LOTOS-EUROS	6.283	4.959	-0.525	1.777	1.362	0.413	2.055	-0.448	0.715	0.756	0.106	-0.255
	CHIMERE	2.560	0.813	-0.456	0.614	0.250	0.335	1.927	-1.379	0.919	0.703	-0.641	0.585
	EMEP	2.690	0.482	-0.219	1.339	-0.251	0.052	3.568	-2.739	0.567	0.711	-0.435	-0.298
	EMEP-MACCEVA	2.681	-0.450	-0.424	2.971	2.574	-0.235	2.911	-2.194	0.724	0.662	-0.540	0.128
	SILAM	2.620	-0.199	-0.327	1.429	0.497	-0.085	3.410	-2.895	0.763	1.468	1.301	0.270
	MOCAGE	3.782	2.639	-0.262	2.070	1.670	0.149	2.277	1.433	0.812			
weekly	ENSEMBLE	0.129	0.009	0.917	0.131	-0.009	0.820	0.101	0.006	0.967	0.060	-0.009	0.802
	LOTOS-EUROS	0.122	0.014	0.970	0.136	-0.008	0.699	0.122	-0.005	0.973	0.129	0.005	0.572
	CHIMERE	0.147	0.029	0.973	0.138	-0.009	0.878	0.134	-0.006	0.954	0.036	0.001	0.935
	EMEP	0.186	0.014	0.708	0.176	-0.013	0.278	0.095	-0.004	0.928	0.082	0.004	0.696
	EMEP-MACCEVA	0.146	0.014	0.865	0.141	-0.010	0.727	0.126	-0.006	0.828	0.130	0.005	0.335
	SILAM	0.097	0.007	0.930	0.083	-0.006	0.898	0.028	0.000	0.993	0.102	-0.001	0.557
	MOCAGE	0.139	-0.009	0.850	0.117	-0.008	0.899	0.120	-0.006	0.872			

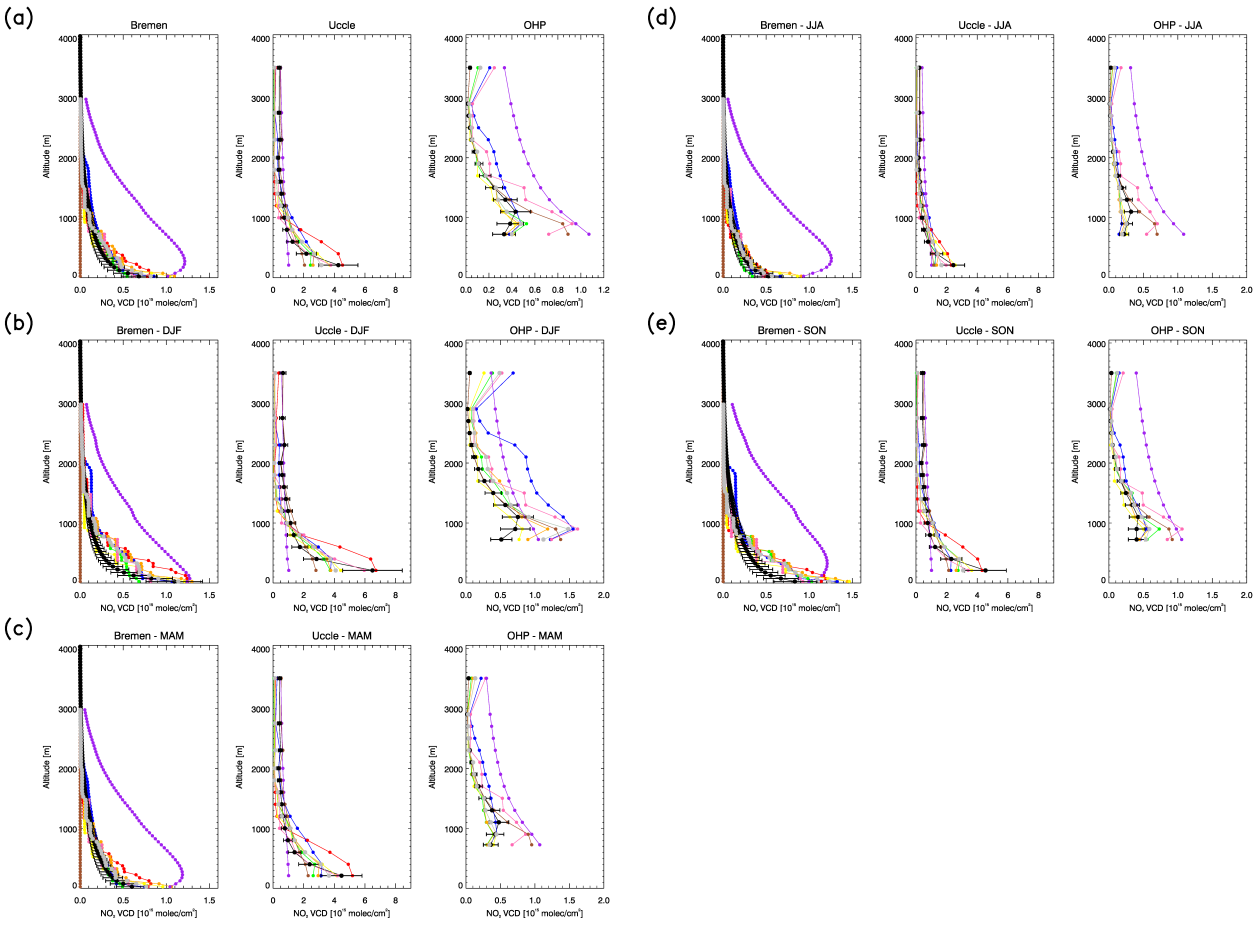

Figure A1. Average vertical profiles of NO_2 partial columns [10^{15} molec cm^{-2}] from (black) MAX-DOAS, (brown) a priori used for MAX-DOAS retrievals, (gray) model ensemble median, (blue) LOTOS-EUROS, (yellow) CHIMERE, (green) EMEP, (orange) EMEP-MACCEVA, (pink) SILAM and (red) MOCAGE. Black error bars refer to the uncertainty associated with the MAX-DOAS retrievals (assumed to be 30 % for all stations). Panel (a) shows profiles for data averaged over whole time series, panel (b) shows profiles for DJF (December, January February), (c) MAM (March, April, May), (d) JJA (June, July, August) and (e) SON (September, October, November) months only. Figures in panels (a) to (e) refer to (left) Bremen, (middle) Uccle and (right) OHP. MAX-DOAS vertical profiles are not available for De Bilt for the investigated time period.

Appendix A: Appendix

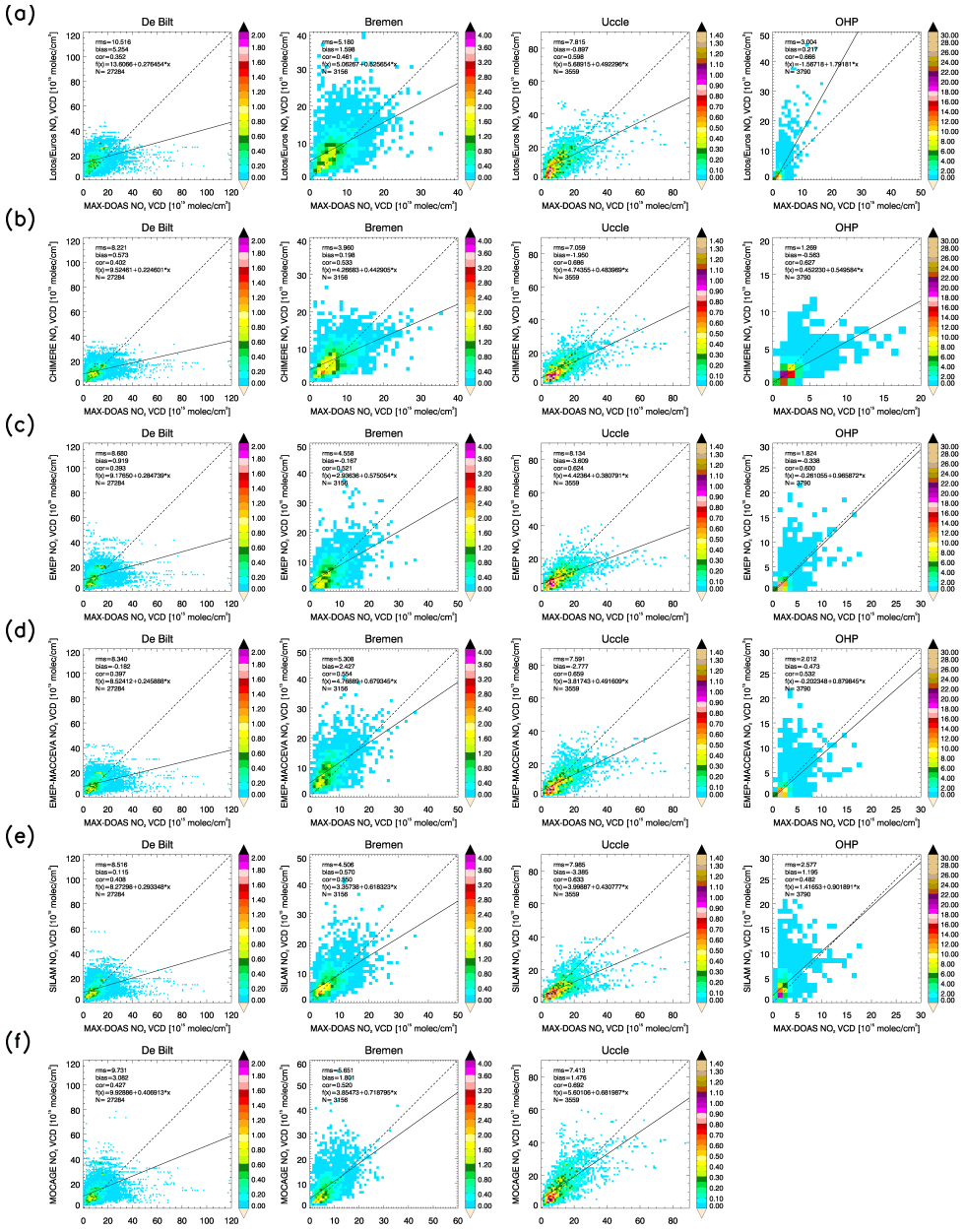

Figure A2. Scatter density plots of AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] from MAX-DOAS against model data for (from left to right) De Bilt, Bremen, Uccle and OHP. The data is shown for different bins with a size of 10^{15} molec cm^{-2} and is colored according to the number of data points per bin [%].The different panels show different model runs: (a) LOTOS-EUROS, (b) CHIMERE, (c) EMEP, (d) EMEP-MACCEVA, (e) SILAM and (f) MOCAGE. MOCAGE data is not available for the measurement time period at OHP. The dashed line is the reference line (f(x)=x). The solid line is the regression line (see top left of each plot for f(x) of this line). The root mean squared error (rms) [10^{15} molec cm^{-2}], bias [10^{15} molec cm^{-2}], Pearson correlation coefficient (cor, not squared) as well as the number of data points N are given at the top left of each plot.

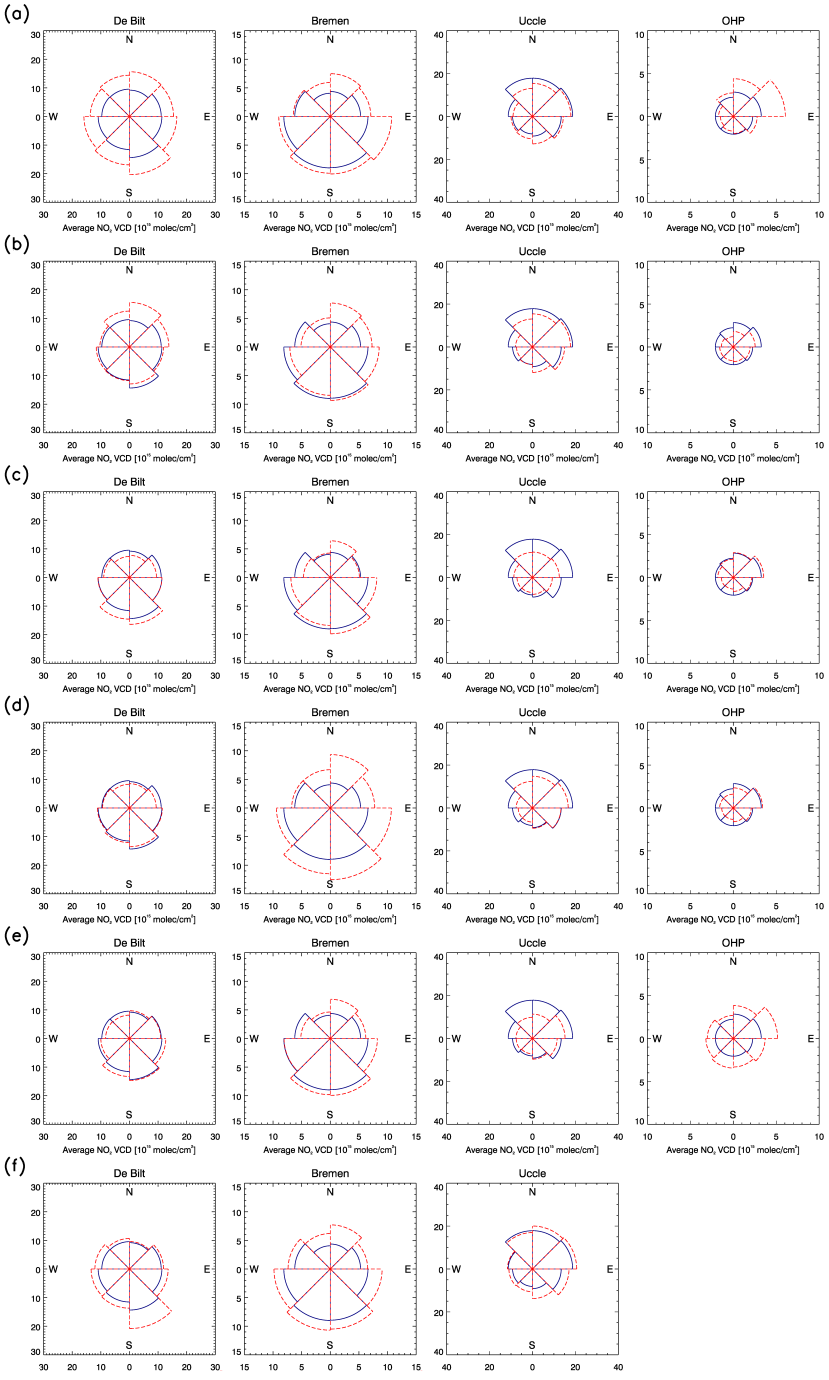

Figure A3. As in Figure A2 but for average AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] in 45° wide wind direction bins from (blue solid lines) MAX-DOAS and (red dashed lines) model data calculated. MOCAGE data is not available for the measurement time period at OHP.

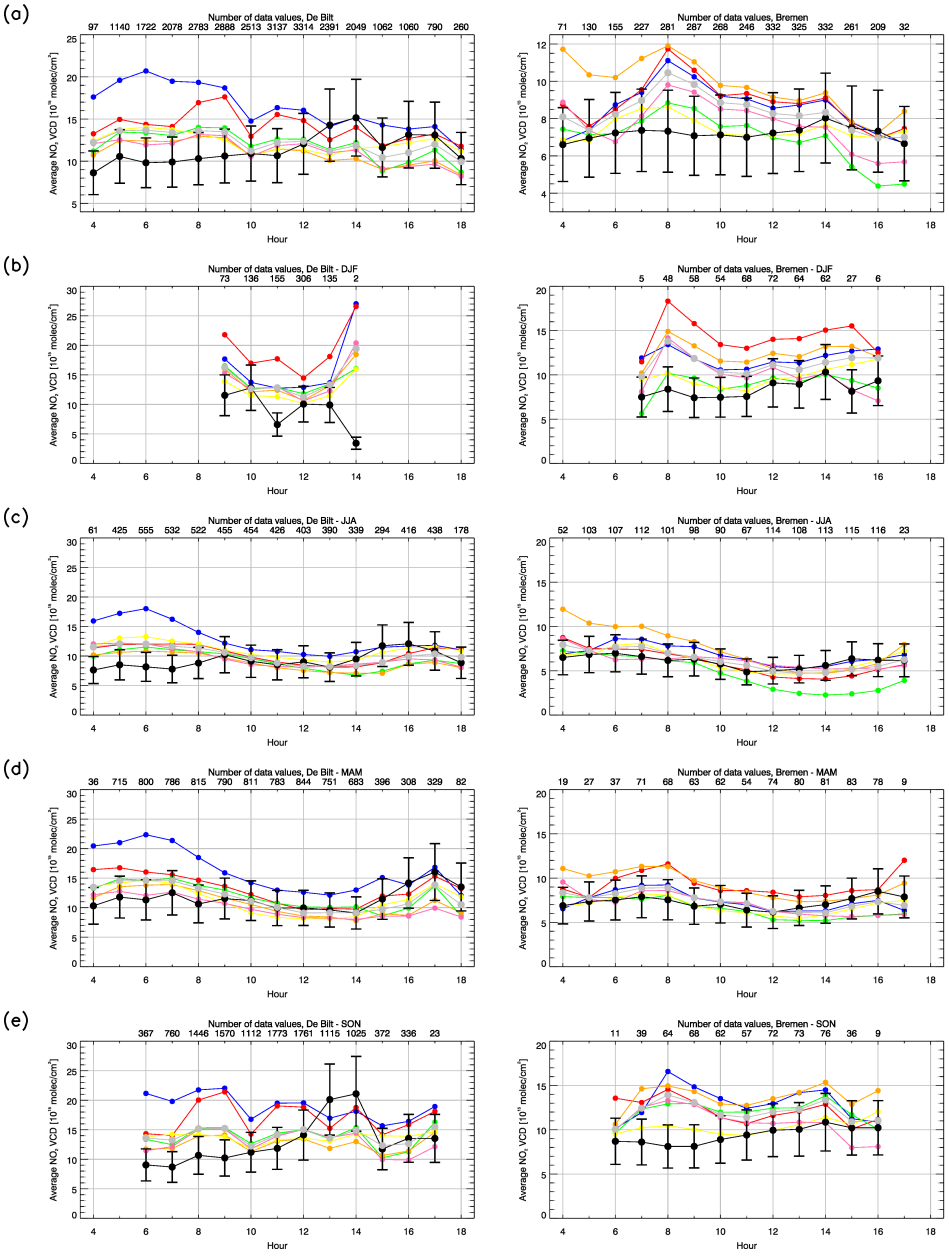

Figure A4. Diurnal cycles (averages over hourly bins) of AVK-weighted tropospheric NO_2 VCDs [10^{15} molec cm^{-2}] from (black) MAX-DOAS, (gray) model ensemble median, (blue) LOTOS-EUROS, (yellow) CHIMERE, (green) EMEP, (orange) EMEP-MACCEVA, (pink) SILAM and (red) MOCAGE for (left) De Bilt and (right) Bremen. Model based diurnal cycles were calculated from tropospheric NO_2 VCDs. Black error bars refer to the uncertainty associated with the MAX-DOAS retrievals (assumed to be 30 % for all stations). Panel (a) shows cycles for the whole time series, panel (b) shows cycles for DJF, (c) MAM, (d) JJA and (e) SON months only. The number of data values used for calculating average values is shown at the top x-axis of each plot.

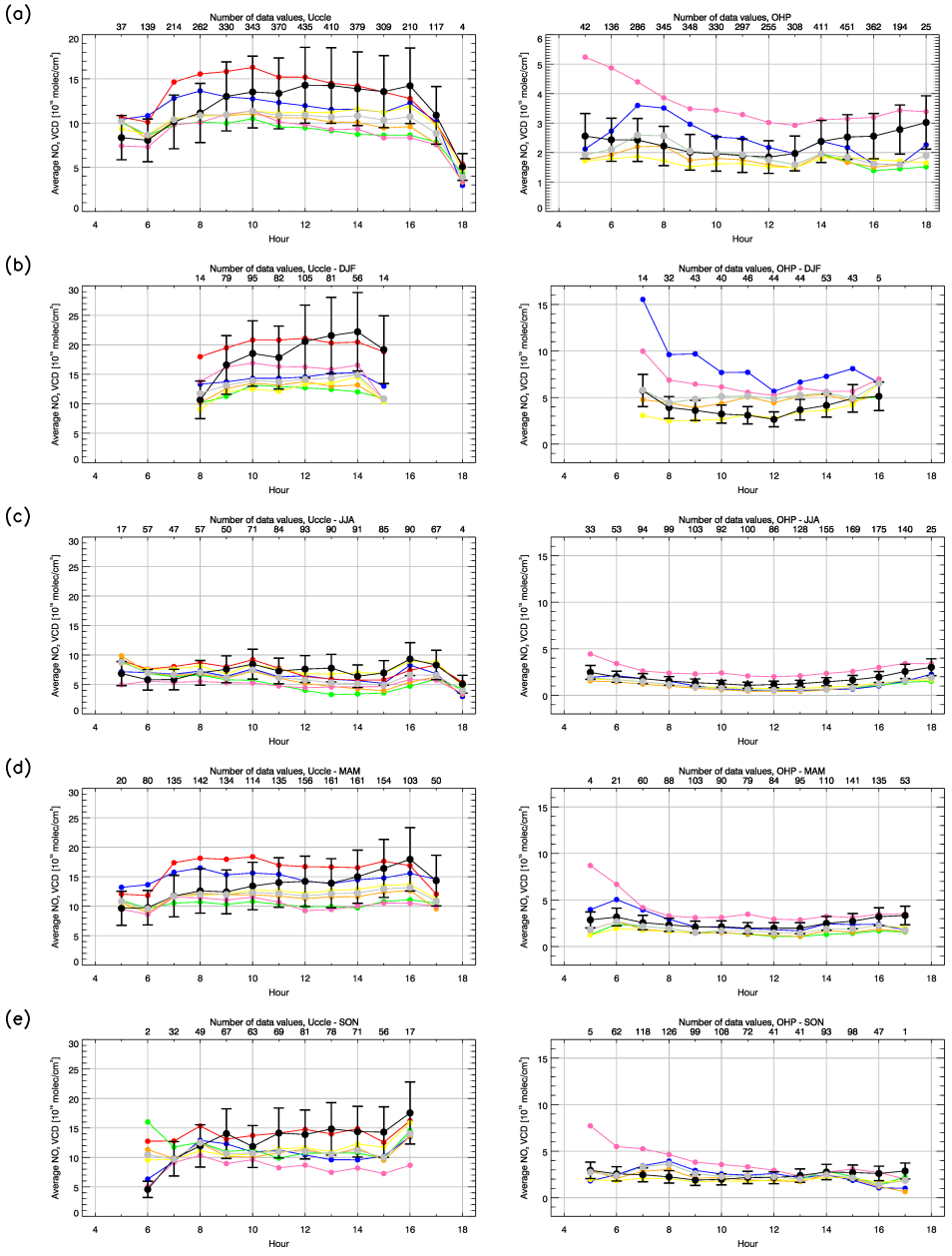

Figure A5. As in Figure A4 but for (left) Uccle and (right) OHP.

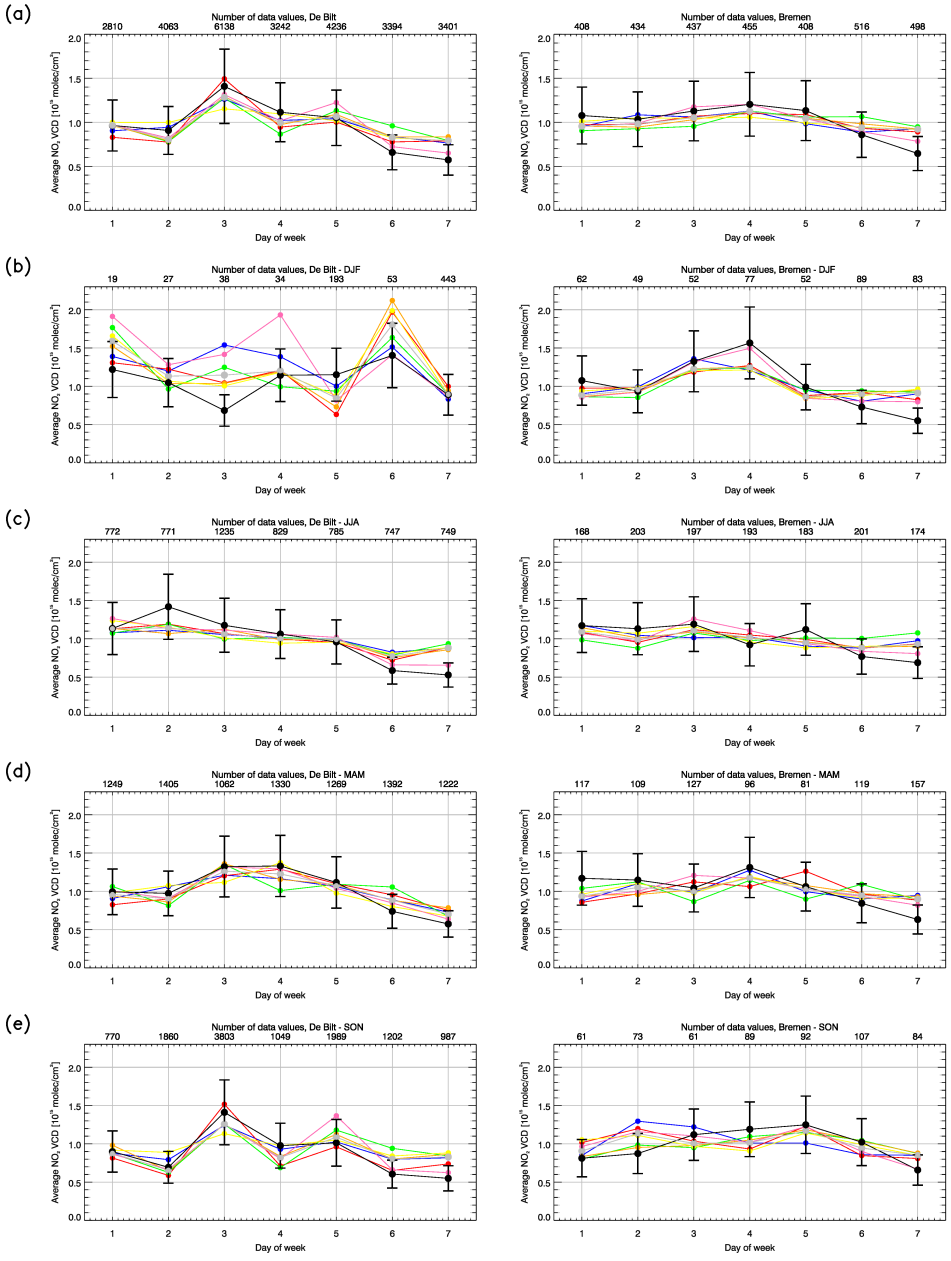

Figure A6. As in Figure A4 but for weekly cycles (averages over daily bins devided by mean over whole week, unitless values) of tropospheric NO_2 VCDs.

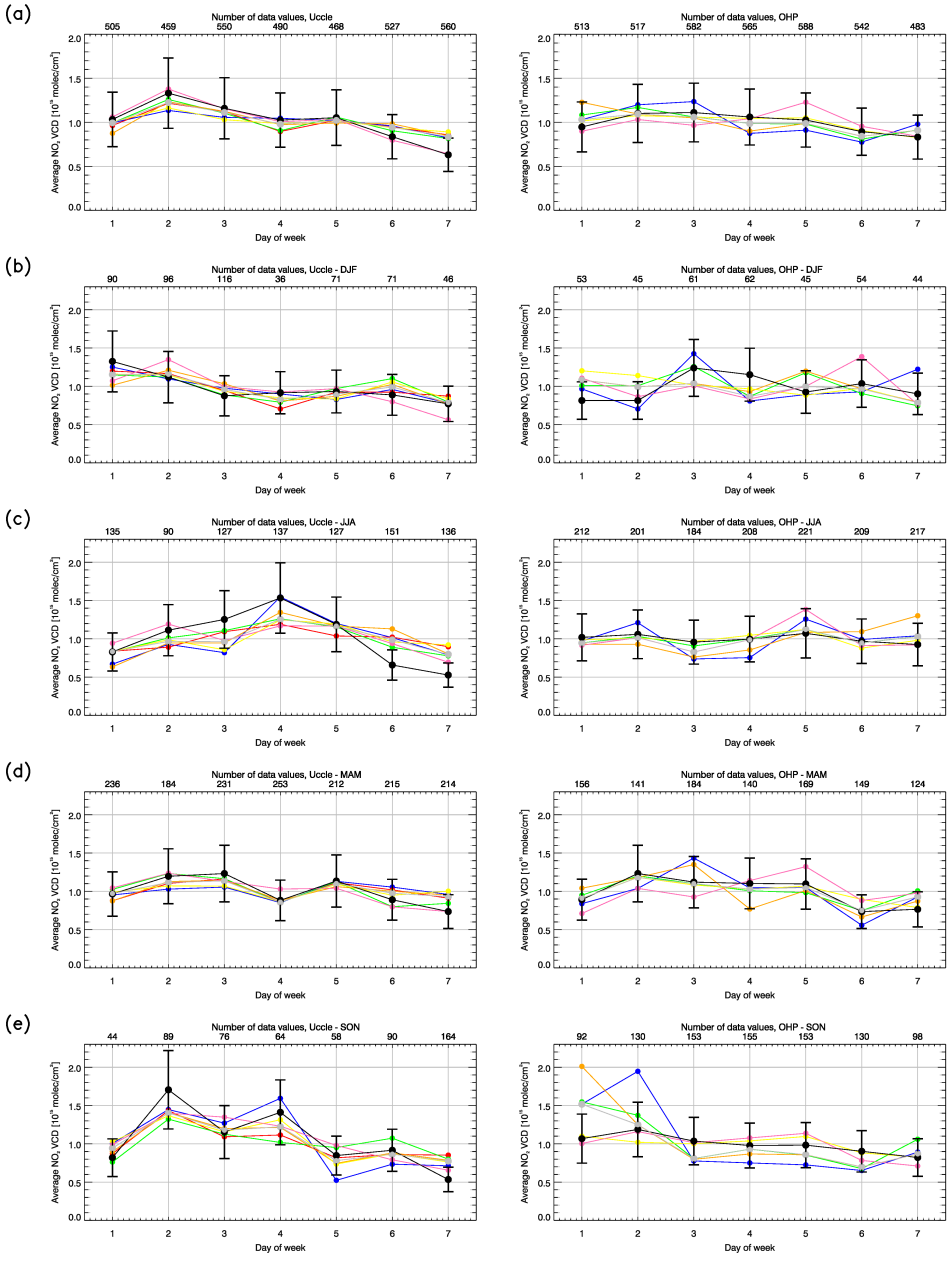

Figure A7. As in Figure A5 but for weekly cycles (averages over daily bins devided by mean over whole week, unitless values) of tropospheric NO_2 VCDs.