## General response

We thank the two reviewers for their comments and suggestions, which help improve the manuscript substantially. We first discuss our revisions in a general sense, and then respond to specific comments.

One purpose of this study is to retrieve NO2 over China from OMI using an improved set of ancillary assumptions, including the explicit treatment of aerosol optical effects, the consideration of surface reflectance anisotropy, the use of relatively-high-resolution a priori vertical profiles of NO2, and the use of cloud properties retrieved with consistent ancillary parameters. Combination of all these improvements is lacking in current literature concerning China (and probably other world regions as well, according to our best knowledge). Another purpose is to evaluate the sensitivity of NO2 retrieval to these ancillary assumptions. For these purposes we conduct radiative transfer modeling pixel by pixel and do not rely on a look-up table and interpolation (as in currently widely used products). Our pixel-specific radiative transfer calculation takes more time than look-up-table-based interpolation. However, we are not concerned with the computational cost, as we are not targeting on a near-real-time (NRT) retrieval. Rather, we think the more sophisticated retrieval here can be used to compare against the NRT retrieval that requires simplification of ancillary assumptions.

The MAX-DOAS NO2 dataset used here is indeed limited by its spatial representativeness, as stated in the original manuscript and now further clarified. The limitation reflects the fact that there is lack of such measurements (or other types of independent measurement that can be used to evaluate satellite NO2) over China. The spatial representativeness particularly affects the evaluation of the magnitude of OMI NO2. However, we find the day-to-day variability to be less subject to the spatial representativeness of MAX-DOAS data. And, our results clearly show that our reference retrieval captures the variability much better than DOMINO-2 ($R^2$ = 0.96 versus 0.72).

The current paper is our first step to improve the NO2 retrieval over China and better quantify the retrieval uncertainties. For this first step, we limit our analysis to locations near the MAX-DOAS sites. We will extend the retrieval and analysis to the entire country in later studies. We feel although current MAX-DOAS data have limitations, they are still useful for understanding NO2 in China and its retrieval from satellites, particularly on the day-to-day variability. In the future, we hope to use MAX-DOAS data in places with a more homogeneous distribution of NO2.

We have revised the descriptions of our analyses throughout the text. We have added more conceptual discussions and removed some of the statistical results and analyses that do not contribute to the overall analysis significantly. We have eliminated Figs. 6,7,12, revised the associated statistics, and moved relevant statistics to Table 4. We have also moved some ancillary information (including Figs. 5 and

10) to Appendix B-D to make the main text more focused.

**Reviewer #1**

The paper "Retrieving tropospheric nitrogen dioxide over China from the ozone monitoring instrument: effects of aerosols, surface reflectance anisotropy and vertical profile of nitrogen dioxide" describes strategies for an improved retrieval of NO2 from space based UV/Vis spectrometers and an evaluation of the improved retrieval. The paper contains interesting elements, however, I recommend major revisions prior to publication.

We thank the reviewer for the comments and suggestions, based on which we have improved the manuscript substantially.

1) The paper is not really about China per se, but rather about aspects of the retrieval. One does not expect any errors in retrieval that are country specific, but rather errors in the incorporation of aerosol, surface reflectance etc. I recommend removing the word China from the title.

We agree and have removed the word. However, the effects of aerosols are more important for China than most of other countries, concerning its high aerosol loadings.

2) There are two major themes in this paper a) improving a retrieval and evaluating the improvements by discussion of what is physically reasonable and by comparison to alternate assumptions and b) evaluating that improvement against MAX-DOAS observations. The two separate issues are muddled throughout the paper making it difficult to understand what has been learned. One solution would be to separate this into two papers. Another would be to more clearly identify a single best retrieval strategy in the first part of the paper and then to evaluate that against observations in the second. In addition, the paper makes weak statements about the utility of the MAX-DOAS evaluation, reporting 20-30% differences from spatial sampling and other sources of error. Given the size of this uncertainty, are any of the differences/similarities reported in this paper significant?

As now further clarified in the revised manuscript, our manuscript first presents a reference retrieval (i.e., our best estimate), compares it with MAX-DOAS, and then tests the ancillary assumptions that may affect the performance of the retrieval. These parts are all important to our overall analysis, and we feel they should be put together in a single paper. As stated in the general response, we have revised the writing significantly to increase the readability of the paper.

Please see our general response regarding the MAX-DOAS data and spatial representativeness.

3) There are too many figures, they present obscure and difficult to parse information

and it is not obvious what the specific point being made with each figure is. At a minimum, the figures should have words explaining the parameters instead of only symbols (e.g. Pressure instead of Ps). In Figure 2, it is not clear to me, that one expects the differences between the two retrievals to be described by a linear model and the low correlation in Panel a) confirms that notion. A different form of presentation and analysis would seem appropriate. Perhaps, % difference as a function of the parameter change? It's not clear all of the panels are useful. The purpose should be to give the reader some intuition about the role of the parameters being tested. Figure 4 could be a table or omitted. Figure 5 should be omitted, it is not informative. Fig 6 and 7 and 12 look like noise. Case number is not a useful coordinate for thinking about these issues. If there is no coherent statement to be made, no figure is needed. Figs 9 and 10 don't appear to be teaching the reader much.

As stated in the general response, we have revised the writing significantly to increase the readability of the paper.

We have removed Figs. 6,7,12 as suggested, and have moved some relevant statistics to Table 4. We have replaced the symbols with plain words in the figure legends.

Fig. 2 – we have replaced the symbols with plain words, and removed the RMA regression in Fig. 2a. We prefer scatter plots that present original information for individual pixels. We think all panels are important for understanding the differences in ancillary parameters between our retrieval and DOMINO-2.

Fig. 4 is a key figure showing the extent of consistency between OMI data (our retrieval and DOMINO-2) and MAX-DOAS for individual pixels/days. We think it should be kept. We have improved the analysis of the figure.

We have moved Fig. 5 and related analysis to Appendix B. We think the figure is still useful, because it shows the dependence of OMI versus MAX-DOAS differences on the spatial constraint.

Fig. 9 (now Fig. 6) is important for understanding the effect of aerosols on the cloud retrieval for individual pixels. We have improved the analysis of the figure.

The results in Fig. 10 are useful for understanding the cloud retrieval. Nonetheless, we have moved the figure and related analysis to Appendix D to increase the readability of the main text.

4) It is logically awkward to have a reference that is not the current standard for retrievals. It would make the paper much easier to read by starting with something standard that has been well documented in the literature and then beginning with a discussion of differences that emerge as a result of the improvements in the design of the retrieval presented herein.

The reference retrieval represents our best estimate. We would like to evaluate the

reference retrieval against the widely used DOMINO-2 (using MAX-DOAS data as reference), and use sensitivity tests to quantify the effects of ancillary assumptions affecting the NO2 retrieval. Therefore we would like to retain the approach to presenting the analyses. We have improved the overall writing to enhance the readability of the paper.

5) The introduction could do a better job of describing the current state of our understanding of how to build an optimal retrieval. For example, the introduction implies that it is an open question whether higher spatial resolution a priori information gives a better retrieval. I think the literature is clear that higher spatial resolution inputs than in the current global retrievals are essential. There are at least 3 papers that make that point.

We have improved the introduction. We agree that higher spatial resolution generally gives a better vertical profile, other factors unchanged. This was why we used the nested GEOS-Chem model to simulate the profile in the first place.

6) The discussion throughout focuses far too much on correlation, R-squared, case number and other statistical metrics and too little on conceptual explanations that interpret these numbers for the reader. The first paragraph of the discussion is an example. Paragraphs such as this should be completely revised to explain the point to an educated non-specialist.

As shown in the general response, we have improved the writing substantially, reducing the use of statistical metrics and adding more conceptual analyses. We have overhauled the discussion section (now Sect. 5.4).

**Reviewer #2**

In this manuscript, Lin et al. present sensitivity studies on retrievals of tropospheric NO2 from OMI observations over 3 MAX-DOAS sites in China. They present a new retrieval set-up which improves on several aspects of the air mass factor calculations used in the operational DOMINO product, evaluate the sensitivity with respect to assumptions made on surface reflectance, aerosols, and NO2 profile and compare their results to validation measurements from the three MAX-DOAS instruments.

The paper reports on an interesting study with relevance for satellite data retrievals used in many atmospheric studies. From the content, it would have fitted much better to AMT instead of ACP (I would have recommended moving the paper to AMT if

asked before publication in ACPD). I'm also not convinced that the special issue on Asian Emissions is the right place to have this paper but this decision is up to the editors of that special issue.

Emissions of NOx, among other species, are still poorly understood over China. Currently, NO2 retrievals from OMI (and other instruments) are used extensively to constrain emissions of NOx through a 'top-down' approach. There are large gaps between bottom-up emission inventories and top-down estimates in NOx emissions over China, which are usually attributed to uncertainties in emission inventories. However, we find that satellite errors are critical for such top-down constraint.

The East Asia Emissions Assessment (EA2) project is designed to improve the understanding of emissions by integrating various bottom-up and top-down approaches. And a key element of the project is to address emission uncertainty. As part of the project, this paper is focused on analysis of errors in satellite retrievals. Therefore we think it fits the scope of the special issue very well.

My first reaction to this paper was enthusiasm, as it promised to address a number of important questions in satellite NO2 retrievals which have often been mentioned but never fully been evaluated. This includes surface reflectance impact, aerosols over China, and most importantly the interplay between aerosols and clouds in the retrieval process. I think the right questions have been asked in this manuscript and an important step has been taken to get a better understanding of these processes. Unfortunately, the paper turned out to be much less useful than expected for three reasons:

1. The data used is limited to the three MAX-DOAS stations as they are used for validation. What is a strength of the paper (validation with independent data) turns out to also be a serious weakness, as the relevance of the results found for the 30 independent measurement points over 3 polluted locations in China for all the other OMI observations is limited.

As explained in detail in the general response, the current MAX-DOAS dataset is not optimal but is still useful for understanding the satellite retrieval, particularly on the day-to-day variability. The current paper is our first step to improve the NO2 retrieval over China, focusing on locations with MAX-DOAS data currently available to us. A later paper will extend the retrieval and analysis to the entire country, hopefully with the help of MAX-DOAS data in places with a more homogeneous distribution of NO2.

2. The reference algorithm produces NO2 columns showing impressive correlation to the ground-based data but there is a factor of 2 difference in the absolute values. None of the (realistic) sensitivity tests performed makes a large change to either the good correlation or the poor agreement in absolute values with the exception of the attempt to reproduce the DOMINO retrieval. My interpretation of these results is that either the parameters tested all have quite limited impact on the retrievals or that the data set

used is not well suited to test the sensitivities of the retrieval. My guess is that the second possibility applies.

As explained before and stated in the manuscript, the current analysis is indeed limited by the spatial representativeness of MAX-DOAS data, particularly concerning the magnitude of retrieved NO2. Nonetheless, our sensitivity tests indeed reveal several important factors affecting the correlation and magnitude (e.g., aerosol SSA, vertical profiles, an apparent overestimate in DOMINO-2 surface pressure, pixel-specific radiative transfer calculation versus use of a look-up table with interpolation). We have summarized the discussion of correlation and magnitude in the discussion section (now Sect. 5.4). We have overhauled the section to improve the readability.

3. The presentation of the results is confusing in many places, focusing on correlations of the results from modified retrievals to the reference case (I do not see what the reader can learn from this number) and failing to addressing the obvious questions. For example, I haven't been able to find out

• what the main reason for the better correlation of the base case with MAXDOAS is compared to DOMINO (my suspicion is, that there is simply something wrong in the DOMINO product),

• which of the computational expensive steps of the base retrieval are important and which can be omitted in a future operational system, or

• whether or not the assumption made in DOMINO (cloud retrieval compensates for much of the aerosol impact) is reasonable.

As explained in the general response, we have improved the writing throughout the manuscript to enhance the readability. Our statistical analyses are now focused on comparisons with MAX-DOAS data (see Table 4).

As discussed in Sects. 4-5 and summarized in the new Sect. 5.4, the better correlation of our reference retrieval to MAX-DOAS is due to pixel-specific radiative transfer calculations instead of interpolating from a look-up table, inclusion of aerosols, inclusion of surface reflectance anisotropy, and possible interactions between the effects of these factors.

As explained in the general response, we are not concerned with the computational cost here. We would like to use an improved set of ancillary parameters to better retrieve NO2 from OMI, rather than aiming to a near-real-time retrieval that requires fast calculation. Moreover, our AMFv6 code has not been parallelized, therefore it is premature to determine the real computational cost of our retrieval approach.

As is now more obviously stated in the revised manuscript, aerosols should be included in the cloud and NO2 retrievals – this is important especially for applications

aiming at fine spatial and temporal scales that are more subject to errors in individual pixels. The errors by excluding aerosols are reduced by spatial and temporal averaging of the retrieved NO2 columns. For our dataset the average difference with versus without inclusion of aerosols is ~ 14%. (We have updated the number upon the original manuscript.)

Based on our study and previous works (e.g., Zhou et al., 2010 on surface reflectance anisotropy, and Russell et al., 2011 on high-resolution a priori NO2 profile), we think improving all these ancillary assumptions are important for the NO2 retrieval. We have improved the relevant concluding remark in the end of the original manuscript.

Considering these issues and the detailed points listed below, I think that the authors should make a real effort to improve on the description of the study and their results as well as the figures and to focus on the relevant conclusions before re-submitting the manuscript.

We thank the reviewer for comments and suggestions, which have been incorporated to substantially improve the manuscript.

Specific Points:

1. MAX-DOAS data. It is very important for the interpretation of the results to know how the data points used are distributed with season. Please include a table of number of data points per month.

The numbers are added in the new Sect. 2.4.4. The number of pixels (days) is 37 (10) in winter, 12 (4) in spring, 30 (7) in summer, and 48 (9) in fall, with relatively good seasonal representativeness.

2. All of the effects treated here (improved spatial resolution, surface reflectance, aerosols) have been discussed in the literature before, and in particular retrievals using WRF-chem, CMAQ or nested GEOS-chem for high resolution a priori NO2 profiles have been applied in several previous studies. It is therefore important to point out where your retrieval improves not only on DOMINO but also on what is in the literature.

We have revised the introduction. OMNO2 by Bucsela et al. (2012) is another NO2 product covering China. Surface reflectance anisotropy was analyzed for Europe by Zhou et al. (2010). High resolution a priori NO2 profile was applied over the U.S. by Russell et al. (2011). We had cited these papers in the original manuscript. We have further improved the analyses in the revised manuscript. We had also cited papers concerning aerosols. We have not found any previous retrievals explicitly accounting for the optical effects of aerosols, and we will be happy to cite the papers if the reviewer can provide some that we have missed. We have also emphasized that we improve various ancillary assumptions simultaneously, rather than enhancing just one particular parameter.

3. The “improved AMF formulation” doesn’t look new to me at all – I understand that AMFs are calculated without LUT which allows to include the effects of BRDF and aerosols in a detailed way. The reason why that’s currently not being done in the operational processors is a) computational limitations when not limiting the analysis to 127 points and b) lack of reliable global input data, for example for aerosol properties.

As explained before, we are not concerned with the computational cost as we are not doing a near-real-time retrieval. However, it is important to understand the effects of simplification in various ancillary parameters in producing the NRT product. Thus, we propose to conduct an ‘alternate’ (albeit more computational costly) retrieval that incorporates a set of improved ancillary parameters to enhance the AMF calculation.

4. The aerosol description is not really clear – are phase functions also taken from the model and if so how? – how are values transferred to other wavelengths? – how exactly was the scaling of model aerosol values done and how large were the factors applied?

We have clarified the aerosol analysis in Sect. 2.3, including how aerosol information is taken from the model, how aerosol optical effects are converted from one wavelength to another, how to scale model aerosols according to AOD measurements, and how large the effect of the scaling is.

5. Why can the DOMINO retrievals not be reproduced better? I would be concerned if the use of a LUT would be the main reason and would in general expect better agreement. Please comment.

As detailed in Appendix C, this is because we used pixel-specific radiative transfer calculations instead of a look-up table that requires projection/interpolation in several important aspects. Also, the LIDORT RTM used here differs by 1-5% from the DAK used to produce the look-up table. A fully quantitative analysis will be important, but is out of the scope of this study.

6. The discussion of geometric AMFs is an interesting one, pointing at an intrinsic problem of this study which tries to evaluate different retrieval settings by looking at the correlation with MAX-DOAS data. If – as the authors say – the majority of the variability resulting in the good correlation is achieved even when using a geometric AMF then this is not a good data set for the sensitivity study! In addition, it would be very interesting to see one more comparison, namely the one between satellite SC and MAX-DOAS VC (the geometric approximation is good for NO2 in the stratosphere but not so good for NO2 in the BL where a constant AMF is more appropriate). It would also be interesting to know the slope for these two simple comparisons.

As explained before, the MAX-DOAS dataset is limited by its spatial representativeness, but is still useful to compare with our OMI retrievals, especially on the day-to-day variability. We find both geometric VCDs and tropospheric SCDs

correlate similarly to MAX-DOAS, and our reference retrieval produces better correlation than both of them. The slopes for both geometric VCDs and SCDs are lower than our reference retrieval. Nonetheless, we have decided to remove the entire discussion as it does not seem to contribute significantly to the overall analysis. We have focused the revised manuscript more on processes and methods than on comparison with MAX-DOAS.

7. Discussion of difference in absolute values between base retrieval and MAXDOAS: In my opinion, this is a weak point - either the MAX-DOAS data are representative for the satellite data, then I would expect better agreement. Or they are not representative, then their use as a standard for decision on the quality of the retrieval is questionable. Relying on relative changes (correlation) helps but only if changes in NO2 are dominated by large scale features.

We admit the limitation in the spatial representativeness of MAX-DOAS data, especially for evaluating the absolute magnitude of OMI NO2. The day-to-day variability is obvious from Fig. 4c-f (where each data point represents a day). It is also well known that over the Beijing area (our study domain), the day to day variation in pollution is regulated greatly by meteorological conditions (polluted days with stagnant atmosphere or weak southerly, transitioned rapidly to clean days with the passage of background air from the northwest/north). And, our reference retrieval captures the variability much better than DOMINO-2.

8. Figure 4: It would be good to have the same x and y-axis in all three figures in a line for better comparability

Revised.

9. Figs 6, 7, 12: it is difficult to read these figures. Please add points to make orientation easier and if possible replace case numbers by more meaningful labels such as “reference”, “no aerosol”, etc. If that’s too difficult, at least add a legend giving the translation from case number to scenario for each figure

These figures are removed, and some of the relevant statistics are revised and moved to Table 4.

10. Fig. 6: I think a table would be more appropriate to display these results

The figure is removed, and relevant values are specified in the text.

11. Fig. 7: As mentioned above, I do not see what I can learn from the R2 of the NO2 column of a special scenario with the NO2 column from your reference case.

The figure is removed, and now the statistics refers to comparison to MAX-DOAS and is moved to Table 4.

12. Fig. 9: How is “change in xxx” defined – is that relative changes or absolute

changes?

They are absolute changes. We have clarified the caption.

13. Fig. 11: Again: What are “changes in xxx”? And why are changes in columns and AMF so different – I thought that VC = SC / AMF and thus any change in AMF leads to a corresponding inverse proportional change in VC?

They are absolute changes. We have added a third column showing the percentage changes in NO2 VCD. We have also clarified the caption.

VCD = SCD / AMF, so the absolute changes in AMF will differ from the absolute changes in VCD.