Reply to Mark Wenig:

First of all we want to thank Mark Wenig for his positive assessment of our manuscript and the contructive and helpful suggestions.
Before we respond to the reviewer comments in detail, we want to give a brief overview on the most important changes of the revised version:

A) We made the main focus of our manuscript more clear.
The main focus of this manuscript is not the further improvement of our original cloud classificaton scheme (Wagner et al., 2014). Actually, the original scheme was almost unchanged. The main aim of our study is the comparison to independent data sets.
Such a comparison to independent data sets was so far missing and is of high importance for the validation of the MAX-DOAS cloud classification scheme. We made this more clear in the revised version (in the abstract, introduction and conclusions).

B) We added the information that no direct and quantitative comparison to the independent data sets is possible.
From the MAX-DOAS observations and the satellite observations not the same quantities are derived: the MAX-DOAS classification results are complex quantities (e.g. categories of broken clouds or continuous clouds), which cannot directly be compared to the independent data sets. Thus only quanlitative conclusions can be drawn, e.g. that the probability for the detection of continuous clouds from MAX-DOAS are found increases if satellite observations indicate high cloud fractions.
Also the temporal and spatial resolution is different: the satellite observations typically cover an extended area, but are valid only for the time of the satellite overpass. Also the AERONET observations are not made at exactly the same location but about 18 km south-west of the Wuxi MAX-DOAS site. And the visibility meter covers only the layers close to the surface.
In spite of these fundamental difficulties, the comparison to the independent data sets is still very important, because from the general dependencies derived from the comparisons it can be concluded whether the classification scheme yields reasonable results or not. Here it should also be noted that to our knowledge, no data set exists, which could be used for a direct and quantitative validation of our classification scheme.
We added this information to the introduction (and also to the abstract and conclusions).

C) We added sensitivity studies with respect to the selected threshold values.
From the comparison to the independent data sets no quantitative conclusions on the accuracy of the cloud classification results can be drawn (see point B above). As a consequence, it is difficult to qantify the accuracy of the cloud classification results. Thus we chose a different way to assess the uncertainties of our classification method: we varied the different threshold values (by ±10%) and studied the corresponding changes of the classification results. Fortunatly, most classification results were only very weakly affected by these variations. This finding indicates that for these quantities the exact choice of the threshold values is not critical. In contrast, for the CI the exact value of the threshold value was found to have a significant impact on the classification results. In particular, it affects the assignment to clear sky conditions with either low or high aerosol. Fortunately, from the compariosn with smultaneously measured AERONET AOD we find that the chosen threshold value was a very reasonable choice. In particular we found that the transition from the categories clear sky with low aerosol to clear sky with high aerosols corresponds to an AOD of about 0.5.

Also the assignment to either broken clouds or cloud holes is strongly affected by the chosen threshold for the CI. But this ambiguity is of minor importance, because both classifications results basically belong to the same cloud category.
We added this informatuon to the new section 3.1.1.

D) We added sensitivity studies of the effect of temporal and spatial averaging
For the comparison of our cloud classification results we averaged the MAX-DOAS results over a period of one hour around the satellite overpass. Also the satellite observations represent averages over extended areas. We investigated the effect of the chosen ranges for the temporal and spatial averaging on the comparison results. Interestingly, the results hardly depend on the selected temporal and spatial averages. In particular the results for different time averages are almost identical. In contrast, a small effect of the spatial averaging on the classification results is found. If the satellite data are averaged over a larger spatial range a more clear assignment of the extreme values is found: The fraction of clear sky scenarios (with either low or high aerosol load) and of cloudy scenarios increases for MODIS observations of small (0 to 10%) and high (90 to 100%) cloud fractions, respectively. These findings indicate that if the requirement of either completely clear or cloudy sky are applied for a larger area the probability that the MAX-DOAS results around the time of the satellite overpass were really clear or cloudy, respectively, increases. The overall conclusion from these sensitivity studies is that the selected standard averaging criteria (temporal averaging: ±30 min around the overpass time; spatial averaging: 0.1 ° x 0.1 °) are well suited for the comparison of the MAX-DOAS results with satellite data. We added this information to the new section 3.2.4.

E) Most of the description of the O_4 analysis (section 2.2.2) was shifted to the supplement.
The old section 2.2.2 contained a lot of very detailed information on the O_4 analysis. To shorten the main part of the manuscript, we shifted most of these technical details to the supplement (including Fig. 5). The new section 2.2.2 contains only a brief summary of the O_4 analysis, which is necessary to understand the remaining part of our study.

In the manuscript "Cloud and aerosol classification for 2 1/2 years of MAX-DOAS observations in Wuxi (China) and comparison to independent data sets" the authors Wang et al. provide a nice overview of classification features that can be derived from MAXDOAS measurements to detect clouds and aerosols. They also describe how those features could be improved compared to an earlier classification scheme by Wagner et al. 2014. Since the topic is relevant for ACP and could be of interest for other MAXDOAS projects, I recommend publications after some revisions. My concerns which should be addressed before the final publication are as follows:

The overall objective of this study is not very clear. I assume that the goal was to improve the classification scheme by Wagner et al. 2014, so that it can be used for the MAX-DOAS measurements as well, but no clear measure to quantify this improvement is presented.

Author reply:
We agree and made the overall objectives more clear (see point A above).

There are lots of open questions, e.g.: What is the percentage of incorrectly classified scenes (that you know of)? How did this value improve compared to Wagner et al. 2014? Do your new classification parameters also improve the Cabauw retrieval?

Author reply:
As pointed out above (see general point B) no direct and quantitative comparison between the results of the cloud calssification scheme and the independent data sets can be made. Thus it is not possible to derive a quantitative conclusion from these comparisons. However, we added a new section on the influence of variations of the threshold values to the revised versions (new section 3.1.1, see point C above).

The reviewer also suggested to apply the 'improved' cloud classification scheme to the Cabauw measurements. In our opinion, this suggested task will bring no new results, because the classification scheme is basically the same as the original scheme presented by Wagner et al. (2014). The main changes are a degradation correction and extension towards smaller SZA (which do not occur at Cabauw).
The determination of a new calibration (and new threshold values) for the radiance and the CI is a step, which is always necessary, when the classification scheme is applied to a new instrument (mainly because of the different instrument response function).
We made these aspects more clear in section 2.3.2.

Under which conditions can your classification be used for other measurement campaigns?

Author reply:
We added the following information at the end of section 3.2.3:
'In principle the classification scheme can be applied to any other MAX-DOAS measurements covering the same spectral range. For the calibration of the zenith radiance and CI as well as the determination of the threshold values, the measurement period should cover a variety of different sky conditions, including especially days with clear skies and low aerosol load. For these clear days independent also information on the AOD (e.g. from sun photometers or satellite observations) should be available. The comparison of the MAX-DOAS results with the AOD measurements allows in particular to adjust the threshold value for the CI to achieve a meaningful separation of clear sky conditions with either low or high aerosol loads (see section 3.1.1). '

You could use a classification scheme based only on AERONET and satellite data as ground truth and compare how often the MAX-DOAS algorithm leads to the same classification. Some scenarios might not fall clearly in the given categories, but you can filter those out, since you have a very long time series.

Author reply:
In principle this would be a good idea. However the fundamental problem of spatio-temporal mismatch remains (see point B above). Thus we did not follow this approach.

Another problem of your study is that you use additional data to improve the thresholds and add additional indicators (e.g. in Sec. 2.3.3 you describe that add an indicator to detect the presence of continuous clouds to match MODIS observations, and also in the supplement you mention that you use scenes with specific sky conditions as selected based on visual images from MODIS to select the thresholds), but then you compare your classification results to those additional data sets. If you use the comparison with AERONET and satellite data to improve the algorithm's thresholds, the better agreement between the MAX-DOAS classification algorithm and AERONET and satellite data is not the result of your study (as described in Sec. 4), the comparison is rather the tool to derive improved thresholds which are then your main results. However, then you need an independent way to assess the quality of the algorithm, e.g. a scientific explanation of the thresholds, or some equations for calculating the values depending on certain parameters (instrument characteristics, location dependent parameters, etc.). Then you could apply your improved classification to the Cabouw data set and check if it improves the classification for that data set as well. You describe that you adjusted the thresholds for different quantities according to properties of the MAX-DOAS instrument (p. 4664 l. 4), but I cannot find a description of that dependency (the supplement is referenced here, but there I can only find a description of how MODIS and AERONET data were used to select threshold values). The dependency on instrument characteristics might be of interest to readers who would like to adjust the classification scheme to their own instruments.

Author reply:
With this comment, the reviewer addresses (at least) two points:
a) He claims that we first determine the threshold values of the cloud classification scheme using selected independent observations. We then use the same independent data sets for the comparison of the classification results. We agree that this procedure does not represent a fully independent validation approach. Nevertheless, we think our procedure is well justified, because the determination of the threshold values is based only on a limited set of scenarios and independent data sets. In contrast, the statistical comparison of our classification results with the independent data sets (section 3.2) covers a much larger variability of atmospheric situations and measurement geometries.
b) The reviewer suggests to add a more detailed description of the determination of the threshold values. We added the follwing text at the end of section 3.2.3:
'In principle the classification scheme can be applied to any other MAX-DOAS measurements covering the same spectral range. For the calibration of the zenith radiance and CI as well as the determination of the threshold values, the measurement period should cover a variety of different sky conditions, including especially days with clear skies and low aerosol load. For these clear days independent also information on the AOD (e.g. from sun photometers or satellite observations) should be available. The comparison of the MAX-DOAS results with the AOD measurements allows in particular to adjust the threshold value for the CI to achieve a meaningful separation of clear sky conditions with either low or high aerosol loads (see section 3.1.1). '

You plotted a lot of different combinations of classification results and several other parameters and describe the results, but what is the significance of the results? For example, when you write "Overall the results from MAX-DOAS are mostly consistent with the AOD data from MODIS." (p. 4675 l.16), that's very vague and more information is needed to allow the reader to see this consistency as well. As mentioned already, a clear criterion is needed to quantify the improvements compared to Wagner et al. 2014.

Author reply:
As stated earlier (see also general point B avove), no direct and quantitative comparison of the cloud classification results can be performed.
But in spite of these fundamental difficulties, the comparison to the independent data sets is still very important, because from the general dependencies derived from the comparisons it can be concluded whether the classification scheme yields reasonable results or not. Here it should be also noted that to our knowledge, no data set exists, which could be used for a direct and quantitative validation of our classification scheme.

This information was added to the introduction.

In the following I will list some comments about minor concerns:

In the abstract (p.4654 l.24) and in Sec. 3.2.4 (p.4678 l.20) you write that “the satellite cloud products contain valuable information on aerosols”, but it’s more an unwanted contamination, right? (it sounds a little like “it’s not a bug, it’s a feature” thing, but in fact it is a bug).

Author reply:
To clarify this point we added the following text to the conclusions:
‘While the miss-classification of aerosol as clouds describes a fundamental problem for cloud retrievals, this finding also indicates the potential to derive information on aerosol properties from satellite measurements of cloud fraction and cloud pressure.’

In the list of different MAX-DOAS retrievals (p.4655 l.12f) you could add Hartl, A. and Wenig, M. O.: Regularisation model study for the least-squares retrieval of aerosol extinction time series from UV/VIS MAX-DOAS observations for a ground layer profile parameterisation, Atmos. Meas. Tech., 6, 1959-1980, doi:10.5194/amt-6-1959-2013, 2013.

Author reply:
We included that reference.

In Sec. 2.2.2 one could first get the impression that you are measuring SCDs and dSCDs. You write that “the FRS used in our analysis is also taken from the MAX-DOAS measurements” (p. 4662 l.15), so you only have dSCDs but no SCDs, is that correct?
You write “Because of the systematic variations we did not use the O4 absorption measured in zenith for the identification of optically thick clouds”, but isn’t the zenith measurement used for the dSCD? That should be made clearer.

Author reply:
We made several changes to the section on the O_4 analysis:
First we moved most of the old section 2.2.2 into the supplement (see also general point E above).
Second, we made the determination of the O_4 SCDs and O_4 dSCDs more clear (in the supplement).
Third, we provided in the new section 2.2.2 only a brief summary of the deteils of the O_4 analysis, which are relevant for this study.

In Fig. S1, can you get rid of the strong temporal variation of the O4 dSCDs if you adjust the reference spectrum according to the FWHM from the high resolution solar spectrum fit?

Author reply:
In principle this would be a good idea. However, in practice it turned out that the O_4 analysis is quite sensitive to these variations of the instrumental properties. In particular the absolute values of the O_4 dSCD change systematically with time. From these findings we conclude that the changes of the spectromteter properties are so complex that they can not be sufficiently accurate corrected.

Fig. 10 b is not really needed, since it only adds information about the number of measurements per month and is difficult to interpret. It doesn’t show the number of sky conditions in each month but rather the number of detected sky conditions, so it depends on the running time of the instrument.

Author reply:
We agree with the reviewer that the main information from this figure is about the number of detected sky conditions (basically the running time of the instrument). But we think this basic information is valuable and we prefer to keep the figure in the paper.

Fig. 11-17 can be replaced by a more meaningful quality measure (see comment above) which can probably be shown in a single plot or table. This plot or table could also compare the classification results using thresholds from Wagner et al. 2014.

Author reply:
As mentioned earlier (see general point B above), no direct and quantitative comparison between the MAX-DOAS classification results and the independent data sets is possible. Thus it is in particular not possible to quantify and condense the information given in these figures into one table. Since also, the classification scheme used in this study is basically the same as in Wagner et al. (2014) (see general point A above), no direct comparison to the results from Wagner et al. (2014) can be made.
To investigate the influence of the choice of the threshold values, we added the new section 3.1.1 (see also general point C above).