Author Response in boldface
Changes in the paper in italic

Referee 1

General Comments The manuscript entitled a simple and versatile could-screening method for MAX-DOAS retrievals by Gielen et al. presents a method for the detection of clouds using the observed radiance and/or O4 slant column density. Such a cloud flag can be used to filter out cloud-contaminated observations, but is also a valuable information by itself. Therefore, the topic of this manuscript fits well within the scope of AMT. In general, the applied methods appear to be valid and the paper is well written. However, as outlined in the specific comments there are several aspects that lack conciseness and require some more detailed explanation.

It is mentioned several times (e.g., at the beginning of the abstract and in the intro duction) that the method proposed here is based on zenith sky measurements only. This statement is a bit misleading as it raises the expectation that the method can be applied to zenith-sky instruments without any restrictions. However, the multiple-scattering flag uses O4 dSCDs which are based on measurements at 30deg elevation angle and require either a MAX-DOAS instrument or a significant modification of the method proposed here. It should be mentioned that the method can be applied to zenith-sky measurements only with certain restrictions.
We have adapted these statements at several instances in the text: Abstract: ***We present a cloud-screening method based on differential optical absorption spectroscopy (DOAS) measurements, more specifically using intensity measurements and O_4 differential slant-column densities (DSCDs).***
Introduction: ***We focus on 90degree elevation observations for the colour index as our simulations show these are the most sensitive to the sky conditions (see Sect. 3). Moreover, they are independent of the azimuth angle, and are very sensitive to the temporal variability of clouds above the instrument site. The use of the zenith measurements means that the cloud-screening method is not only limited to MAX-DOAS but can also be applied to similar instruments working in the zenith mode only. For the O_4 measurements we also use the 30-90degree elevation measurements, but the method can also be applied if only zenith measurements are available (see Sect. 4.3).***
Conclusions: ***We present a cloud-screening method for MAX-DOAS measurements to qualify the sky and cloud conditions. The method is based on the colour index (CI) and O_4 DSCD retrievals. We focus on colour-index observations made at zenith elevation, whereas for the O_4 DSCDs we use both the zenith and 30degree data, but the method can be adapted to work only with zenith measurements. This means that the method is not only limited to MAX-DOAS instruments, but can also be applied to traditional zenith-sky DOAS measurements***

Apart from the measurements in Brussels, the improvement of the aerosol retrieval using the different cloud flags is not very convincing. For Xianghe, there is hardly any improvement of the agreement between AOD form MAX-DOAS and form sun photometer when applying a cloud filter. A statement on the level of improvement for the Jungfraujoch data cannot be made at all since, due to the small AOD at this site, there is virtually no correlation between MAX-DOAS and sun photometer. Here perhaps histograms of the difference between MAX-DOAS and sun photometer data would be more useful. In summary, I feel that the conclusions should be formulated in a more balanced way regarding the capability of the algorithm to improve the retrieval of aerosol properties from MAX-DOAS measurements.
We have redone the analysis using non-cloud screened co-located measurements, to better compare our method. This better shows the total effect of our method. We have changed parts of the text of Sect.6 and conclusions based on this new analysis. We have also added histogram plots to compare the AOD distribution between retrievals and measurements.
To study the effect of our cloud-screening method, AOD values retrieved by MAX-DOAS are compared to co-located AOD measurements. For Xianghe and Brussels we use AERONET Level 1.0 (unscreened) (and 1.5 (cloud-screened)) data, and for the Brussels site we ex-

tend the comparison with co-located Brewer spectrophotometer measurements at 320 nm (Brewer instruments #16 and #178). A detailed description of the co-located instruments and measurements can be found in Cheymol2003,DeBock2010, and Holben2001.

For the Xianghe data set we find high correlation coefficients R, already for the non-cloud-screened data.This is due to the fact that this site has only little influence from clouds, especially in comparison to Brussels, as can be seen in Figure ??. For both 360 and 477 nm we have a correlation value of $\sim R = 0.86$, and also the linear regression slopes S are very close to $S = 1$. For both wavelengths the cloud screening based on the CI (green crosses) slightly increases the correlation, with correlation values changing from $R = 0.86$ to $R = 0.89$. We do see a difference between the two wavelengths: at 360 nm our model seems to overestimate the AOD in comparison to AERONET, whereas the opposite occurs at 477 nm. Applying the cloud screening does improve the slope at 477 nm (from $S = 1.21$ tot $S = 0.91$), but worsens the slope at 360 nm (from 0.95 to $S = 0.78$).

In the supplementary material we also show the correlation between our AOD retrievals and co-located measurements, but now using cloud-screened AERONET level 15 and Brewer data. For Jungfraujoch no such cloud-screened data are available. We find that the AERONET cloud-screening procedure Smirnov2000, based on the stability of a measured AOD triplet over a 30 second interval and temporal AOD hourly and diurnal variability, removes more data compared to our cloud screening, leaving around 28% for Xianghe and 10% for Brussels. This results in better correlation and slope values for both Xianghe and Brussels, compared to the correlation with the non-screened level 10 data, with improvement on average of the order of $0.05 - 0.1$ for both R and S. As the AERONET cloud screening is based only on temporal variability of the AOD, stable uniform clouds and aerosol plumes can be misidentified. This could account for differences between our cloud-screening method and the AERONET screening, as for example seen in the first plot of Fig. 1. For this day with a strong rise in aerosol load, the second half of the day is flagged as mainly cloudy by AERONET, whereas we do not.

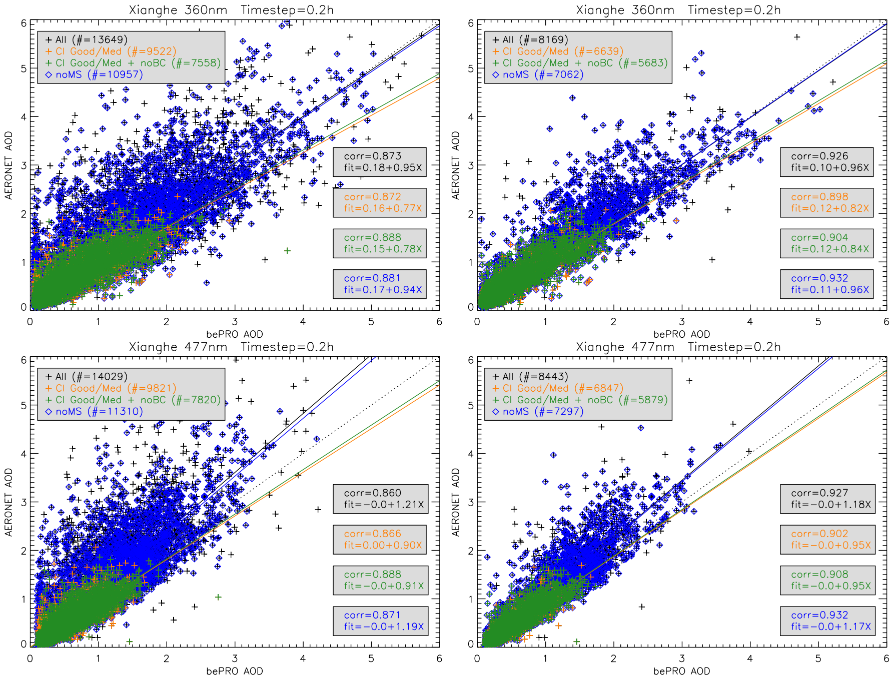

Figure 6: Continued below.

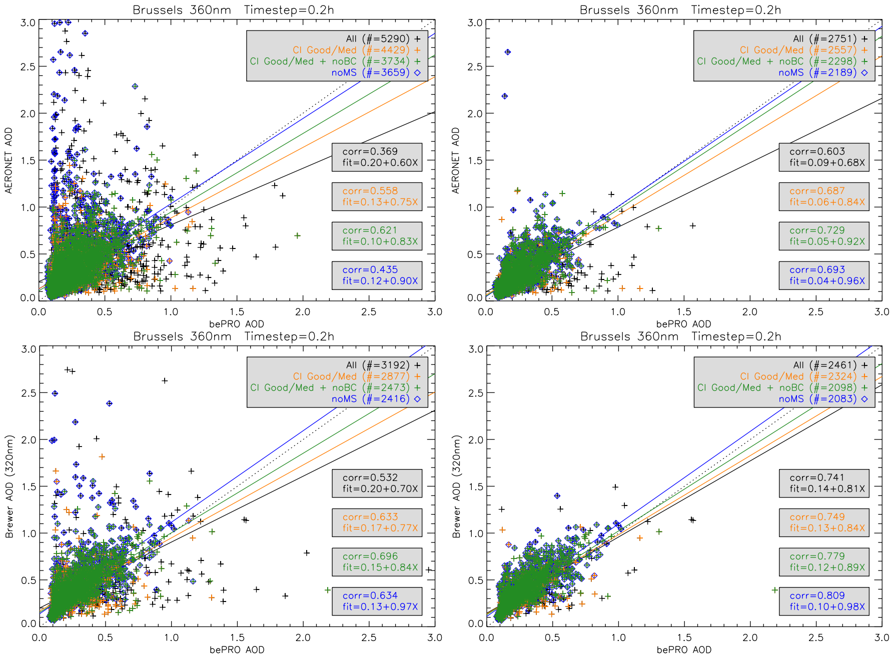

Figure 6: Correlation plots of our bePRO MAX-DOAS AOD retrievals and measured AOD values for the Xianghe and Jungfraujoch data set at 360 and 477 nm and for the Brussels at 360 nm, in time steps of 0.2 hour for Xianghe and Brussels and 1 hour for Jungfraujoch. The figures on the left use non-screened AERONET/Brewer data, whereas the figures on the right use cloud-screened measurements. The full non-cloud-screening data is given by black crosses. Cloud-screened data (based on the CI) with a 'good/mediocre' sky flag are marked in orange, data with 'good/mediocre' sky flag and no broken-cloud flag are marked in green crosses. Data with no multiple-scattering flag (based on the O_4 DSCDs) are marked with blue diamonds. For each sample set we also give the linear regression lines and correlation information.

…Overall, the effect of our cloud screening on the Xianghe data is only minimal, which is mainly due to the fact that the site is not highly affected by clouds ….. We conclude that our cloud screening has the largest influence on the Brussels data set, as expected due to it being the most cloudy site. For the Brussels and Jungfraujoch sites, it is sufficient to base the cloud screening on information from the colour index alone, whereas for Xianghe, additional information from O_4 DSCDs is invaluable for a correct cloud identification, as the colour index alone will result in a removal of non-cloudy data with high aerosol load.

It is mentioned that the CI-based broken-cloud flag would be more sensitive to aerosols than the multiple-scattering flag, which would render the latter more suitable for the discrimination between clouds and high aerosol load. However, both flags are based on the detection of a high temporal variability of the measurements, in case of the broken-cloud flag based on CI and in case of the multiple-scattering flag based on O4. Please explain why the sensitivity to aerosols should be higher for the CI- based flag than for the O4- based flag.
We do not say anything about the sensitivity of the broken-cloud flag to aerosols. We mention in the paper that the broken-cloud flag will be more sensitive to different types of clouds than the multiple-scattering flag, as the latter will be most sensitive to optically thick clouds and the former also to thinner clouds. However, the colour index is a very bad tracer to discriminate between clouds and aerosols (as they will have a similar effect on the sky colour), for this the multiple-scattering flag is better suited (as multiple scattering will occur more frequent in optically think clouds than aerosols).

Specific Comments

5884.2 and 5885.11: These sentences suggest that only zenith sky spectra are used for cloud detec-

tion, but later on you also use measurements at 30deg elevation.
Adjusted the text to avoid this confusion. Already discussed above.

Section 3: Very different wavelength pairs are used for the three sites. I understand that this originates from the different wavelength ranges of the instrument. However, wouldnt it be useful to homogenise the definition of the CI for different instruments in the framework of this study for the sake of better comparability?
We recommend to use the largest wavelength range possible for the calculation of the colour index, to increase the sensitivity. We therefore opted to adapt the CI wavelengths for each instrument. Also, due to the very different ranges of the different instruments, with little overlap, a single wavelength pair that works for all sites would not have a broad enough range.

5889.10: The CI for Jungfraujoch is defined as I405/I560, but you state above (5887.25) that the instrument only covers a wavelength range up to 550 nm. How can this be?
This is due to a typo, the wavelength goes to 560 and the CI is calculated using I550. Fixed in the text.

5889.23ff: Here it should be explained in more detail how the simulations with DAK were performed, including a description of the aerosol profile shape, cloud thickness and cloud and aerosol optical properties. I could imagine that the cloud base height has a significant impact on the modelled quantities, in particular on O4. It would therefore be important to vary also the cloud base height in the model simulations.
We tested the effect of a varying cloud height on the CI simuations, which proved to be only minor. No O4 simulations where used in this paper. Additional information on the DAK parameters was added to the text: ***Simulations of the CI corroborate the observed decrease of the CI in the presence of clouds and aerosols, as can be seen in Fig. 4. These simulations were made with the DAK (doubling-adding KNMI code) radiative transfer model (Stammes1989,Stammes2001) under varying aerosol and cloud optical depths, and varying parameters such as wavelength, elevation, SZA and azimuth angle. For the aerosols a homogeneous layer up to 1 km with a single scattering albedo of 0.9 and asymmetry parameter of 0.7 was used, for the clouds these values are respectively 1.0 and 0.85. The cloud base height was set at 1 km, with a total thickness of 1 km, a surface albedo of 0.05 was used, and atmospheric Rayleigh scattering and ozone absorption were included. We also tested the effect of varying the cloud base height, ranging from 1 km to 8 km, but found very little influence on the derived CI values, especially for higher elevation angles.***

5891.1: I do not understand the concept of broken or scattered clouds in the line of sight as this is a property of the entire sky and not only of a particular viewing direction.
We mean a rapid variation between cloudy/non-cloudy in the line-of-sight. We have adapted the text to avoid confusion: ***To determine the presence of broken (semi-continuous cloud cover) or scattered clouds (predominantly clear sky) in the line-of-sight of measurement, the temporal variability of the CI is studied.***

5892.3: The definition of the bad region is quite vague. At which value exactly do you cut off the peak? Do you use a particular percentile?
We extended the definition on the limit value, using the FWHM of the peak of the frequency histogram. More specifically, we place the limit at a value of 1xFWHM from the peak position of the histogram: ***To separate between the 'mediocre' and 'bad' regions we define a horizontal line in such a way that the peak of the frequency distribution falls in the 'bad' region. More specifically, we place the line at a distance of FWHM (full width at half maximum) from the peak position of the histogram. If x and y respectively denote the CI values and the frequency distribution, then the limit is $L_{bad} = x(y_{max})+FWHM(y)$. Note that this is of course only valid if the peak of measured CI values is associated with cloudy conditions. For sites with very clear skies and only little cloudy measurements a reverse approach could be taken. In this case a similar definition using the peak distribution could be used to define the 'good' regime, and the 'bad' regime by comparing with simulations.***

5892.20 Please specify what x means in this equation. If it is time, please replace x by t. Please avoid having the same symbol f both as function name on the left side and as parameter on the right side.
Adjusted: ***To quantify this we model the observed CI values over time*** t ***for each day with a double-sine function of the form*** $f(t) = A + B\sin(Ct - D) + E\sin(Ft - G)$***. Outliers are then identified as those data points with*** $|(CI(t) - f(t))/f(t)| > 0.1$***.***

5892.21: it is not clear what you mean with —CI model— ¿ C. I suggest to replace this equation by —CI(t) f(t)— ¿ C. Why is C different for the different sites and, more importantly, how exactly did you determine the different values for C? For the multiple- scattering flag, you used the relative change of the signal. Would this also be useful for a more general definition of the broken-cloud flag?
The limit value is different for the different sides as the observed CI range is different for the different sites, depending on instrument characteristics and adopted wavelength range.

We chose to use a relative limit for the O4DSCD, as these have been in a way scaled by removing the 90deg elevation DSCD. This makes the sample much more homogeneous over time and removes the large diurnal trend. For the CI values this is not the case, which means the CI values can, especially for good days at Xianghe, reach both very low values (morning/evening) and quite high (noon). Using a relative change cutoff limit, this gives a predominance to flag broken clouds during morning/evening. Also, for overcast measurements, the CI values can be extremely low, again resulting in too many points flagged if a minor change in CI occurs.

However, since this problem typically occurs for very low CI values, these data will in any case already be flagged as 'bad' and removed from the sample. This means the influence on our overall results is only minimal. We have adapted our study to use also for the broken clouds a relative change limit. But, one needs to keep this in mind if he wants to use solely the broken-cloud flag for some statistics: ***Outliers are then identified as those data points with*** $|(CI(t) - f(t))/f(t)| > 0.1$***. This value was derived by investigating those days with rapid temporal variability in the CI. For these days it was found that the observed jumps in CI predominantly fall above these cut-off value. These outliers are flagged as observations made under scattered/broken-cloud conditions.***

5893.7: Not only clouds but also aerosols can have these effects on the O4 dSCD. How can you distinguish between changes in light path due to aerosols and due to clouds?
At this point we make the assumption that all rapid variations in observed O4 DSCDs are due to clouds, as aerosols will typically introduce a smoother variation, as mentioned in the text.

5894.1: It should be specified that outlier means flagged as affected by multiple scattering. Again, I suggest to replace the equation by $|(O4(t) - f(t))/f(t)|$ >C.
Adjusted in the text.

5897.17: I do not understand what you mean with RMS < 50perc. 50perc compared to what? The RMS is an absolute value quantifying the (error-weighted) difference between measurement and retrieval. In an ideal retrieval, it should be close the number of elements in the measurement vector (see, e.g., Rodgers, 2000).
By this we mean the percent root mean square difference between the measurements and the simulations, more specifically: RMSD $=\sqrt{\frac{\sum(meas-retr)^2}{\sum meas^2}}$**. We have added this to the text.**

5897.2: The statement that A removal of data with evidence for the presence of clouds, be it either based on the sky and broken-cloud flag or the multiple-scattering flag, results in a much better agreement with the AOD measurements and retrievals only applies to the measurements in Brussels, as the correlation analysis later on shows.
We have adapted our analysis to use non-screened AERONET/Brewer data, where it is more clear that also for Xianghe an improvement in correlation occurs. The effect of the cloud screening at Xianghe is indeed only minimal, but this is to be expected as the site is not very cloudy, especially compared to the very cloudy Brussels. We have adapted this

in the text as already discussed above.

Instead, it appears from the data shown in Fig. 9 that measurements flagged as bad sky have systematically higher AOD values than the other data, and are most of the time also flagged as cloudy by the sun photometer algorithm. It would be useful to investigate in more detail to what extent cloud flags from DOAS and from sun photometer coincide.
Cloudy data will indeed have higher retrieved AOD values, as our radiative transfer model will try to model the increased optical depth due to clouds with aerosols (as clouds are not present in the model). We have also added a small paragraph on the comparison between our flags and the presence/non-presence of sun photometer data. ***...From Fig. 12 it is clear that 'bad' data on average have higher AOD values. This is due to the fact that our bePRO model tries to model the observed optical depth increase caused by the clouds with aerosol optical depth, as clouds are not present in the model....***
For Xianghe, about 46% of points with coincident co-located measurements for the correlation study remain. For Brussels and Jungfraujoch this is around 20%. This large removal of data is not only due to direct-sun restrictions but also long-time inoperation of the AERONET/Brewer instruments. Another note of caution is that the MAX-DOAS and other AOD-measuring instruments have different viewing directions, and might thus trace regions with slightly different cloud and aerosol characteristics.

We do find a good agreement between our cloud flagging and the absence of AERONET/Brewer data. For Brussels $\sim 75\%$ ***of data without coincident measurements are flagged as cloudy, for Xianghe this number goes up to 80% and for Jungfraujoch around 65% of data with no co-located measurements are flagged as cloudy. A large percentage of the remaining data without co-located measurement but no cloud flag from our method can be attributed to instrumental inoperability.***

5900.1: Do you have an explanation for the poorer correlation between MAX-DOAS AOD and Sun photometer AOD in Brussels compared to Xianghe?
One explanation is the difference in instrumental quality between Brussels and Xianghe. The MAX-DOAS instrument at Brussels is only a mini-MAX-DOAS, which has a much lower signal-to-noise ratio. This results in a larger uncertainty on the DOAS O4 DSCD retrievals. These larger uncertainties will then result is a larger spread of retrieved AOD values. Also the very cloudy meteorology at Brussels results in very little stable measurements over time, which hinder our bePRO retrieval. One solution would be to perform the CI cloud-screening to the measurements before applying the retrieval model, but we have not yet tested this.

5900.20: I do not agree with your conclusions. For Brussels, there is clearly an improvement of the agreement between MAX-DOAS and sun photometer after a cloud screening has been applied. For Jungfraujoch, the correlation is very poor no matter if a cloud screening has been applied or not. And for Xianghe, the correlation is already very good for all data, and does not improve after a cloud screening has been applied. **We have adapted our analysis to use non-screened AERONET/Brewer data, where it is shown that also for Xianghe an improvement in correlation occurs. The effect of the cloud screening at Xianghe is indeed only minimal, but this is to be expected as the site is not very cloudy, especially compared to the very cloudy Brussels. We have adapted this in the text as already discussed above:** ***...When we apply the cloud filter to our aerosol retrievals we find an improvement in the agreement with other co-located measurements, such as from cimel and Brewer instruments, both in correlation and slope, which increases strongly for sites with the high cloud rates.***

In particular, I cannot support the conclusion that for Xianghe, additional information from O4 DSCDs is invaluable for a correct cloud identification, since this screening method leads to only very small changes in the correlation coefficient, and moreover leads to a slope of 1.16 for the measurements at 477 nm which is worse than when CI-based cloud flags are applied (slope of 1.01). **We did not say that information from the O4 DSCD will drastically improve the correlation/slope even further, but it is clear from Fig.9 that the O4 flags are needed to make a correct cloud identification possible and not remove data made under very high aerosol loads.**

5900.26: Again, you method is not only based on zenith observations but, in case of the O4-based flag, also on observations at 30deg SZA.
Adapted in the text.

5902.4: Given the results of your correlation analysis in Section 6.2, I do not agree with your statement that When we apply the cloud filter to our aerosol retrievals we find an improvement in the agreement with other co-located measurements, such as from cimel and Brewer instruments, both in correlation and slope. This is only true for the measurements in Brussels.
As stated before, we have redone our analysis using the non-screened co-located measurements, which are better suited to show the effect of our cloud screening. It is still true that the biggest effect is seen for the Brussels data set, but this is expected as this site is the most cloudy. We have adjusted the text to include these remarks as already discussed above.

Technical Corrections **All adjusted in the text**
5890.15: Junchfraujoch - Jungfraujoch
5890.25: 3 - three
5891.8: get rid of - remove
5893.13: effect - affect
5896.1: depends - depend
5897.18: points - point
5898.27: data for which no co-located AOD measurements are available
Figure 1: Please mark the different parts of the figure (a, b, . . .) to which you refer to on P. 5889. What are the units of the O4 dSCD?
Figure 6: In the figure caption, what do you mean with calculated CI values? Do you mean measured CI values?