Final Response Anonymous Referee #2

We would like to thank the referee for his comments. Most have been incorporated in the newest version of the manuscript. Below our response to each comment.

I have tried, somewhat unsuccessfully, to keep the discussion in this review on what was done in this study as opposed to what could be done in three or four other studies.

Given concerns about iteration, convergence, and processing time, it is not clear why the A Priori is used as the first guess. (The only benefits would be in the interpretation of the initial measurement residuals and a standardized set of starting points for convergence and counting the number of iterations. Since both of these change with the choice of a climatology, even that is lost.) One could use previously retrieved profile for adjacent FOVs to reduce the expected number of iterations and to improve convergence. It would also be interesting to see how the non-convergent cases for one climatology perform when given the retrieved profiles for a better performing climatology as their first guesses.

> Using the *a priori* as a standardized starting point in the experiments of Section 4-6 (which are done with the same ozone climatology) guarantees that the improvements in convergence statistics are entirely due to the algorithm adaptations.
> As the reviewer correctly remarks, this argument does not hold for the comparison study of different climatologies in Section 7. We have added a new section about the influence of the first guess on the convergence behaviour. The intercomparison of ozone climatologies has been done again, but taking the result of the previous retrieval as a first guess instead.

The convergence criteria on Page 1168 line 10 use the A Priori covariance as part of the measure. This means that the later tests with different covariance matrices will have different convergence criteria. An alternative is to use a measure of the size of the measurement residuals as the stopping criteria.

> In this study we focus on convergence behaviour as diagnostic tool to isolate different problem areas of the algorithm. We do not feel that switching to other criteria will change our main results. Besides from the ozone climatology comparison, in each retrieval experiment the same covariance matrix (and therefore the same convergence criteria) is used.

The calculated DFS for the retrievals are a useful measure, but as noted they are sensitive to the choices of the two covariance matrices. (If one somehow had an A Priori set that was close to the truth, then the measurements would not add much, but the results would be good.

> In the next version of our manuscript, we replace the TOMS climatologies at three different fixed relative errors by one with a more realistic error, based on the variability of the climatology with respect to ozone sonde measurements from the WOUDC database. Comparing retrievals with FK, MLL and this TOMS, one can see that a higher DFS is coupled to worse convergence. Here, a high DFS is not an indication of better retrieval quality, but of weaker regularization of the climatology. In the new text this has been explained in more detail.

It is also not clear how one includes the information in a total ozone estimate as used with TOMS climatology in the DFS calculation and whether measurements are used twice; once in obtaining the total ozone estimate and again the profile retrieval.)

Obtaining a total ozone column with differential spectral absorption techniques at two wavelengths is independent of the technique to find an ozone profile by a spectral fit over a large spectral wavelength range. Hence the total ozone column and its corresponding profile from the TOMS climatology can be considered as independent *a priori* information.

The retrieved profiles should be examined to see how their covariance about the A Priori profiles compares to the assumed profile covariance, and the final measurement residuals should be examined to see how they compare to the assumed measurement error covariance.

The first part of this comment implies an extensive validation study, which is considered out of scope for this paper. Regarding the measurement residuals, we have added results and interpretation of the goodness of fit (tested with the reduced chi-square) for the different retrieval experiments, showing the improvement of the SAA filter (χ^2_{red} drops from 52 to 6.5), the insensitivity to the cloud-workaround (χ^2_{red} remains 1.3), and the improvement by switching from Bass-Paur to Brion ozone cross-sections (χ^2_{red} drops from 1.3 to 1.2).

Biases in the initial and final residuals can provide indications of calibration or model biases. These should be examined for the cross-section studies. (It is not clear how possible measurement calibration biases are addressed in this study. More information on the process used to derive them would be useful, e.g., is it recalculated when the ozone cross sections are switched? The determination of a radiometric offset could remove information. The results for these offsets should be communicated to the Level 1 processing team.)

Despite the improving quality of GOME 0-1 processors, we see an improvement of the spectral fit by the forward model if for each retrieval a radiometric offset C_0 is fitted in the band 1a window:

$$R' = R_{sim} + 10^9.C_0 \quad \text{, in [photons/s.sr.cm}^2\text{.nm] ,}$$

in which R_{sim} is the simulated radiance and R' the corrected radiance used in the calculation of the simulated reflectance. Typical values of C_0 are 0.08 ±0.06, meaning that typical corrections for the shortest wavelengths are around 1%.
In the final version we will include the results of the goodness of fit for both cross sections. The reduced chi-square drops from 1.3 to 1.2 when switching from Bass-Paur to Brion ozone cross-sections, indicating a better performance of the forward model with the latter cross-section. Detailed study of biases to find measurement calibration issues are described in Van der A et al. (2002) and Krijger et al. (2005) and are referenced in the manuscript.

One could claim that the SAA difficulties were caused by an underestimation of the Band 1a measurement noise for that region. (The local wavelength to wavelength variability could provide one measure of the measurement noise.) The processing of the data to remove the outliers is a good idea and is shown to work well.

The additional measurements noise in the SAA is caused by spikes due to impacting high energy particles. Because of their non-Gaussian nature, these spikes can not be included in the optimal estimation method by just adapting the measurement error. Instead they should be discarded. We added information on the spectral fit: “Filtering the measurements improves

the goodness of fit of the forward model in the SAA region considerably: the reduced chi-square for converged retrievals drops from 52 to 6.5. Outside the SAA region the converged retrievals fit with $\chi^2_{red} = 1.3$."

The plot on the right side of Figure 3 shows some SAA effects producing large negative impacts on the radiances which are not filtered. Do the authors have any comments on the physical source of these or plans to improve the filters to identify them?

Negative reflectances are caused by negative radiances (solar irradiances are always positive, of course). At level 0 to 1 processing, the dark current is subtracted from the measured signal which can result in small negative values if the signal is below the noise level. Furthermore, the impact of energetic particles not only cause spikes, but also can cause a local discharge of the CCD detector, resulting in more negative values.
In the future, we will try to improve our filter and identify these negative spikes.

A good set of measurements combined with a good forward model should always have a reasonable retrieved profile. One can check non-convergence results to see how the measurement residuals are varying; Are the measurements internally inconsistent, that is, is there small scale structure in the residuals as functions of wavelength that is symptomatic of measurement errors? (Relaxation techniques can be used to obtain convergence for nonlinear problems. For ozone profile retrievals, one can have each iteration move a fraction of the full profile change given by the linearized step result to avoid cycling between results. The fractional value can be selected to insure that the maximum likelihood cost function is decreasing at each iterative step.)

Checking spectral residuals of non-convergent results to detect and classify forward model or measurements biases, and using relaxation techniques to avoid oscillating solutions, are good ideas to tackle the last bit of non-convergence and certainly will be considered for future research.

It would be good to provide convergence statistics for the algorithm applied to other instruments to determine which problems are caused by GOME measurement or calibration idiosyncrasies and which are more general results.

Added to the discussion: "The presented results are based on GOME measurements, but the different causes of non-convergence are not unique for GOME instrument; they will also show up (though not necessarily at the same strength) when the algorithm is applied to other spaceborne UV backscatter spectrometers." We agree that the application of convergence diagnostics to other instruments and see how they compare would be a good idea for further research.

The GOME measurements should be able to provide good cloud pressure, cloud fraction, and absorbing aerosol estimates by using rotational Raman scattering, discrete reflectivity channels, and aerosol index methods in the 340 to 380 nm wavelength interval. These should be more consistent with the quantities needed for the 265 to 330 nm range used in the algorithm than those from a much longer wavelength region. (See Sneep, et al. (2008), Three-way comparison between OMI and PARASOL cloud pressure products, JGR, 113, D15S23, doi:10.1029/2007JD008694 and Vasilkov et al. (2008), "Evaluation of the OMI cloud pressures derived from rotational Raman scattering by comparisons with other satellite data and radiative transfer simulations," JGR, VOL. 113, D15S19, doi:10.1029/2007JD008689 for more information.)

A short overview of alternative cloud parameter retrieval methods has been included in Section 5. The referee is right that cloud parameters retrieved at 340-380 nm will represent more the cloudy situation as sensed in the 265-330 nm range. However, compared with the oxygen-A band the absorption in the O2-O2 band is very weak. It is unclear if the mentioned advantage will outweigh the introduced error by doing the cloud retrieval in this spectral domain. Unfortunately, the O2-O2 cloud retrieval algorithm is not implemented in OPERA and can not be tested.

How much do the retrievals change for the different choices of climatologies and covariances, and are they well predicted by the Averaging kernels and the portion of the A Priori profile differences that lies outside of the retrieval null spaces? (See Rodgers, C. D. (1990), Characterization and Error Analysis of Profiles Retrieved From Remote Sounding Measurements, J. Geophys. Res., 95(D5), 5587–5595, doi:10.1029/JD095iD05p05587.)

We agree with the referee that this would be useful information. This should be addressed in a thorough validation study, which is considered out of scope for the present work.

There are additional considerations for choices of climatologies and A Priori information related to the expected use of the retrievals in operational versus climate applications.

The conclusion and discussion of the climatology comparison has been adapted in the new version of the manuscript. “Good convergence not necessarily implies good retrieval quality; the ozone profile retrievals based on a certain climatology should always be validated against independent measurements. But when selecting an ozone climatology for a specific application it is recommended to take also its convergence behaviour into account (combined with the degrees of freedom), considering available computational time, the desired use of measurement information content, and the ratio of successful retrievals in problem areas.”