Dear editor,

We have prepared a revision of the paper taking the referee comments into account.

We also discovered that we had incorrectly implemented the Piece Wise Linear Fitting procedure, which is non-trivial for multi-variate regressions.

This has some consequences for the paper – outlined below – but the important conclusions remain the same.

We are open to considering the revised version of the paper as a resubmission (thus re-entering the discussion phase), but with the same referees.

We leave it up to you as editor to decide.

Best regards, also on behalf of my co-authors,

Jos de Laat.

## General comments

### Incorrect implementation of PWLT fit

We discovered that we had incorrectly implemented the PWLT fit in the regression (implementation is not trivial). We apologize for this. Our misunderstanding is partly related to a lack of detailed description in most relevant papers on how to implement the PWLT in the regression. This has a number of consequences for our paper:

**What has changed:**

- For a few scenario combinations the post-break trends up to 2010 now become statistically significant, better in line with Kuttippurath et al. [2013].

- The number of ensemble-wide significant trends for the period up to 2010 now is of the order of 30-60% (was 0-20%). That means that still it still cannot be argued that the recovery of Antarctic ozone is statistically significant.

- A majority of post-break trends up to 2012 are statistically highly significant – consistent with, and evidence of, the notion that the longer the post-break period, the better the statistical significance.

- PWLT and EESC trend distributions are more similar, which is actually what one would expect and should have triggered some suspicion on our side.

- There is a bit more consistency on which ozone and EP-flux scenarios results in the best explanatory power (but still with caveats)

- Separate description of PWLT and LINT trends (thus with/without connecting the separate piece wise trends) is no longer included.

**What has remained the same:**

- The EESC regression results have not changed

- Our advice thus remains to avoid using the EESC as it is unclear what is its best description

- How to define the best ozone and EP flux time periods over which to average remains unclear

- It is still argued that this may lead to introduction of non-chemical variations in the ozone record

- We still find that a longer post-break period results in a larger number of statistically significant trends

- Similarly, a longer post-break period still does not necessarilly always result in improved statistics

- Volcanic ash, the QBO and the solar flux still do not improve the regressions

**How this affects conclusions:**

- results now better show that a multi-variate regression help in reducing deterministic non-chemistry variations in average ozone.

- there is general tendency for more significant trends with increasing time period. This is in a way similar to what we already concluded, but the number of significant trends to start with (from) is larger, suggesting we are closer to detection of recovery. This bodes well for the near future and expectations are that a high confidence of the recovery occurring may be reached before 2020.

- Nevertheless, one remains to be careful with the interpretation of regression results, as it remains unclear what the best ozone and regressor records are for this type of study (area and in particular the time period for which to average are ill defined).

- the abstract and conclusions were modified accordingly.

**How to advance**

We are open to considering the revised version of the paper as a resubmission (thus re-entering the discussion phase), but with the same referees.

Reason is that there have been major revisions and modification of the text, the tables and the figures, even though what is presented in the figures and tables has remained the same.

Similarly, the setup of the paper has also remained the same, but the discussion of the (optimal) regression model has been reduced in favor of more emphasis on the statistical analysis and trends.

However, we leave it up to the editor to decide on how to proceed exactly and are willing to accept other options (the incorrect implementation of the PWLT trend was our mistake).

## Response to referees

Below follows a detailed response to the referee comments.

**Referee #1**

Major issues

1) the first point raised by the referee in essence comes down to the question if the uncertainty in proxies as defined in our paper can be reduced, possibly by weighting the proxies in a weighted multivariate regression analysis.

First of all, this is a valid and highly relevant question. As pointed out in the paper, the topic of uncertainties in variables used in multivariate regression analyses of ozone has not drawn much attention. Our paper is thus a first attempt to try to quantify the impact of proxy uncertainties on the regression analysis. We are quite certain that there are possibilities to improve on our analysis, and we hope that our paper provides some incentive for others to have a closer look.

With regard to filtering the data ensemble, there are many textbooks providing methodologies how to optimally select the independent variables in your regression. There are rules of thumbs (make sure to have considerable fewer independent variables than degrees of freedom), pre-processing analyses (check the cross-correlation of the independent variables); check the probability distribution of the independent variable, as the presence of outliers (think about the year 2002, which we will address later) can have deteriorating effects on the regression), post-processing analyses (check to what extent residuals are randomly distributed; look at the p-values of the regression coefficients) and in-processing analyses (out-of-sample testing)

In addition, although we have chosen a probabilistic approach to estimate the impact of “errors in variables” on the regression, there are other methods to address this “errors in variables”. However, these methods require well characterized variable errors. And, as shown in our paper, for most variables errors were unavailable or difficult to properly characterize.

Because of a lack of properly characterized variable errors, some methods for estimating the optimal set of independent variables also become “tricky”. A frequently used approach is to select the set of variables that results in the best post-regression statistics (random residuals, smallest residuals, best correlation). However, such a selection method does not take variable errors into account (this becomes particularly problematic if errors are systematic and/or non-linear).

We thus have reservations to the post-processing selection of an optimal set of variables – even though we do apply that method in our paper as well, *i.e.* our results indicate that several variables do not contribute to the regression.

So, is a regression that explains a small fraction of the observed ozone variations not as “good” as a regression that explains most of the observed ozone variations? Well, not necessarily: if the variations in the independent variables are dominated by errors than much of the “goodness” of the regression may be artificial. The answer to this question still comes down to the question on how well one understands the variable errors.

Based on our analysis, one could argue that the understanding of variable errors is not very large: results can be all over the place depending the estimated variable errors. Worse yet, even the choice of optimal time period for determining ozone is not well motivated, adding an additional layer of uncertainty.

On the other hand, our analysis does provide some guidelines on what and what not to include. Do use the EP-flux, preferably including Aug+Sep. Include Sep+half Oct for ozone, but nut much more. Do not use the EESC. Do use the PWLT and the SAM. Exclude volcanic ash and the combinedQBO-solar flux index.

Furthermore, the length of the period over which the trends are calculated is also quite important.

With that knowledge now established with this paper, one can start thinking about how to improve the process, or alternatively, to start thinking about other methods than a multivariate regression on total ozone to determine stratospheric ozone recovery. However, we argue that both aspects beyond the scope of our current paper, which we view as a starting point for future research.

2) What about the solar-QBO index.

First of all, in section 2.3 we state that we use the 30 hPa winds - consistent with Kuttippurath et al. [2013] and Haigh and Roscoe [2006] and Roscoe and Haigh [2007]. Reason for only testing the combined solar-QBO index is that we want to remain consistent with Kuttippurath et al. [2013].

The referee has a number of questions about the use of the combined solar-QBO index. It is suggested that maybe both parameters should be included in the regression analysis separately, and provide uncertainty estimates of both (including uncertainties in the QBO phase).

However, as explained in Holton and Tan [1990], the QBO and solar effects cannot be considered separately. Whereas the solar influence modifies tropical stratospheric ozone and dynamics, the transport of the solar signal to higher/polar latitudes depends on the phase of the QBO. Or in other words: the QBO modulates the solar signal [Labitzke, 2004]. It is thus necessary to somehow combine both proxies when studying solar and QBO effects on stratospheric ozone, for which Haigh and Rosco [2006] and Roscoe and Haigh [2007] provide a methodology. In addition, Labitzke [2004], as well as Roscoe and Haigh [2007], show that there is little influence of the phase of the QBO and the solar cycle on Antarctic vortex dynamics during Antarctic spring (contrary to Antarctic summer and autumn as well as the Arctic regions, where there are clear QBO and solar effects).

In summary, it is well established that the QBO and solar effects on stratospheric ozone should not be separated, and there were already indications that their effect on Antarctic springtime ozone is small. Whether then to include them in the regression altogether is could be discussed, but since our analysis is based on a study that does include the combined solar-QBO index, we decided for the purpose of comparing our results to start with the same proxies as previous studies.

Nevertheless, our findings confirm the lack of QBO-solar signal in Antarctic springtime ozone, which we interpret as the Antarctic vortex being too strong for the fairly modest dynamical signals to penetrate.

*We included a summary of this discussion in the section on the combined solar-QBO index and in the conclusions.*

Note that a similar argument applies for stratospheric volcanic ash: literature exists suggesting that the influence of volcanic ash on Antarctic ozone is small or absent, also something opposite to stratospheric ozone elsewhere, and thus that there appears no need to include volcanic ash in studying Antarctic springtime ozone. Similarly, we interpret that as the influence of volcanic ash – either via dynamics/QBO or chemistry – is too small to affect the very strong Antarctic vortex and changes via chemistry are small compared to the very large effect of ozone depleting substances on Antarctic springtime ozone. Consistence of our findings on volcanic ash with other published results was already included in the paper.

3) Sensitivity of trend analysis on including 2002.

We performed a test of the ensemble without including the year 2002. Indeed, trends in ozone without including 2002 are larger, but not by much (mean trend difference +2.9 %; 2-sigma spread in trend differences ranging from -12.5 to 17.9 %).

In absolute numbers:

without the year 2002, the 2-sigma spread in post-break trends is:

1.91 to 4.67 DU/year with a mean of 3.29 DU/year

with the year 2002, the 2-sigma spread in post- break trends:

1.80 to 4.12 DU/year with a mean of 2.96 DU/year

Note that for trends calculated based on ozone itself – without “correcting” the ozone record based on the regression results - the inclusion of 2002 does matter, and trends are significantly different:

with 2002 the post-break trend in ozone for all 8 ozone scenarios varies from 0.52 to 2.09 DU/year

without 2002, the post-break trend in ozone for all 8 ozone scenarios varies between 2.35 and 2.96 DU/year.

Hence, the numbers above actually indicate that the regression is very effective in removing the 2002 anomaly.

For volcanic years (1983, 1984, 1992, 1993), this matters less. Even for the incorrected trends in ozone the differences are small.

With all volcanic years, the pre-break ozone trend ranges from -4.72 to -6.82 DU/year

Without the volcanic years, the pre-break ozone trend ranges from -4.62 to -6.68 DU/year

In relative terms, the differences are less than 4% for pre-break ozone trends with and without inclusion of volcanic years.

*We added a remark about the sensitivity to 2002 in the discussion of the trends, as well as about the sensitivity to inclusion of "volcanic" years.*

**Minor comments**

Minor comments are addressed accordingly. Below only follow comments that require a more detailed response.

- Abstract: see general comments to the editor

- PWLT vs LINT. See discussion of our incorrect implementation of the PWLT in the regression, which is not trivial. After proper implementation the distinction between PWLT and LINT vanishes.

- Solar-QBO effects. See discussion above. We added a summary of the discussion above to the paper.

- What is the proper solar index? With the question about F10.7 vs. Lyman-α the referee confirms that it is very unclear what the proper solar proxy is. However, as we argue above, it is crucial to combine both QBO and solar index into one new index. Haigh and Rosco [2006] and Roscoe and Haigh [2007] provide a methodology for doing so, using particular QBO and solar proxies, so we used them here as well. As noted by other referees, the final goal of the paper is not to discuss the optimal set of proxies but trend uncertainty. Yet as this comment and several others show, it is not simple to discuss one without discussing the other.

- Figure 2 has been updated, now including a legend indicating the color corresponding to each QBO-solar scenario combination as described in the figure caption.

- Section 2.8, Vernier et al. [2011] reference added. Also added Solomon et al. [2011] rather than Trickl et al. [2013], as we argue Solomon et al. [2011] is more appropriate for discussing global changes in stratospheric aerosol.

- As explained above, EESC based trends we calculate from applying an Ordinary Linear Regression (OLR) to the EESC pre-break and post-break shape multiplied with the regression coefficient. In particular the pre-break EESC shape includes a levelling of the EESC near the break year (late 1990s). Hence, the linear fit error for the pre-break period will automatically be larger due to the non-linearity of the EESC shape. This turns out to be less of an issue for the post-break trend (whose trends is smaller and thus less affected by the levelling off than the pre-break trend.

- Table 1 trend uncertainties (now table 2): we calculated EESC-based trends by come and ORL to the pre-break and post-break EESC multiplied with the regression coefficient. Since we calculate EESC-based ozone trends using an OLR for both the pre-break and post-break period separately, there is no relation between pre-break and post-break EESC-based trend errors (see previous bullet). See also the previous bullet. We do see that the different EESC scenarios have different errors (2-sigma) for 5.5, 4 and 2.5 years Age of Air, respectively

    - Pre-break: -5.3 ± 0.1, -5.8 ± 0.4, -5.9 ± 0.7

- Post-break: 1.01 ± 0.12, 2.07 ± 0.03, 2.97 ± 0.16

The discrepancy between our EESC-based trend errors and those of Kuttippurath et al. [2013] suggest their error calculation differs from ours, but their paper does not discuss how their trend errors are calculated. Note that the regression based error in the EESC fit in our regression is much smaller than the error of the ORL fit to the pre-break and post-break EESC change.
Nevertheless, as we argue in our paper, a trend error calculation should include the residuals, *i.e.* the variations in ozone unexplained by the regression. The difference in PWLT trend errors – which does take the regression residuals into account - and the EESC trend errors therefore suggest that the EESC-based trend errors are overconfident.

- Figure 4: figure caption should indeed be “1979-BREAK” and “BREAK-2012” rather than “1979-1999” and “2000-2012”.

- A check was performed on the consistent use and description of BREAK periods in document.

- Auto-correlation: the main point of adding these references is that ozone time series generally are auto-correlated. If so, and in particular when calculating trends and trend uncertainties, one has to make sure that time series for which the trend is calculated does not show much – if any – auto-correlation. Fortunately this is the case, but otherwise the trend uncertainties should be corrected for auto-correlation (thereby increasing trend uncertainties, thus decreasing confidence in detection of recovery, which is thus relevant for this paper).

- Red/Blue outlines in figures 5 + 6. These lines indeed indicate the sum of all probability distributions of the scenarios. Reason to add them is that in the end it is these outlines on which the trend significance is based (even though we do make some refinements in the paper). Hence, we prefer to leave them in.

- Although we agree that aerosols have little effect, we feel that some discussion is needed as this finding is relatively new – there were a number of publications in the 1990s suggesting a significant influence of volcanic aerosol on Antarctic springtime ozone and only recently some new research (including this paper) combined with longer time series of ozone suggests no influence of volcanic aerosol on Antarctic ozone. Hence, we prefer to keep the figure as it presents a still rather new result that contrasts a number of older papers.

- Figure 6, lower panel: caption modified, it indeed shows the distribution of the regression value for the aerosol variable, including the distributions for the three different EESC scenarios. This was not properly explained.

- Section 3.6: optimal regression. As outlined in detailed above, we argue that even though this paper does not aim at providing the best set of regression parameters, choices in regression parameters do affect trend significance and the distribution of statistics from the regressions are related to the use of particular combinations of regressors. Some of the detailed questions and remarks by all referees also confirm a need of some understanding of the relevance of each variable in the regression. We nevertheless condensed section 3 in total, but a final word on what is and what is not an important variable in the regression and what appears to be better choices for time periods over which to average is really needed.

- Page 18516, lines 13-22, discussion of why not to use EESC. We agree, we removed most of the section but included the remark by the referee, which nicely summarizes the issue.

- Figure 9, table 4. Is now table 7, the table is modified to include percentage of significant changes also for all break years but ending either in 2010 or 2012, which are consistent with the red bars in figure 9. We also referred to table 6 in the caption of figure 9. The panels in figure 9 are swapped.

- With regard to tables 4 and 5 with significant trends for the ozone and EPflux scenarios: after proper implementation of the PWLT regression results are more consistent. For ozone including September and at least part of October results in the better trend significance (so the period should not be too short), while avoiding making the period too long. For the EPflux this means including September and August. This is explained in the text, including a warning that still a proper physical justification is lacking (for example, to include mainly full calendar months is quite arbitrary when you think of it).

- Figure 4: added the 2000-2012 trends and 2-sigma errors as presented in table 1 (now table 2). Added a brief discussion of them to section 3.3, nothing that maybe the uncertainties in the 1979-1999 ozone trends are larger than estimated from a single regression, but that 1979-1999 trends are nevertheless all statistically significant.

**Referee #2**

Major concerns

1) Try to better distinguish trend significance and explanatory power.

See also our answer to Referee #1 and point (3) below as well as our general comments to the editor. Although indeed we focus primarily on trend significance, the results of our study do have consequences for the optimal set of independent variables used for the regression. Furthermore, the two – significance and explanatory power – are not independent: the better the explanatory power of your independent parameters, the more likely it is your trends will become statistically significant as a better fit will remove some variations that otherwise would be considered noise in the trend calculation.

*With the major revisions of the document we put less emphasis on the regressor selection and more emphasis of the trend significance. The main message should be clearer.*

2) Distinguish between trend significance and trend consistency with expected recovery

It is currently expected that the Antarctic ozone hole will start to recover – or HAS to. When looking at trend estimates from various studies it is clear that stratospheric ozone over Antarctic has slowly started to increase. However, until now this increase in not considered statistically significant – a requirement for the recovery of ozone as defined by WMO – even though some studies do claim a statistically significant increase. However, as we show, that may be a too confident statement, as some errors in the methodology for determining recovery haven not been taken into account. The WMO ozone assessment report 2014 is very careful in its wording on this, noting exactly what is described above: Antarctic ozone appears to be increasing, but the increase is not statistically significant – nor does it already has to be based on modelling studies.

We realize that the distinction between an increase and a statistically significant increase may lead to some confusion, but we want to avoid that it is concluded that – because of lack of a statistically significant increase – the ozone hole is not recovering. In particular because of the fact that we do not yet necessarily expect to already see a statistically significant increase [Newman et al., 2007; WMO, 2010; Hassler et al., 2011], we wanted to make sure that people understand that despite the lack of statistically significant increase what we see is still consistent with expections.

*We have added a brief discussion of the current insights vs what we expect, also in the light of the changed results after implementation of the proper PWLT regression, and we now suggest that based on our results it can be expected that more confidence in recovery based on multivariate regressions can be expected before 2020.*

3) Contradictory statements with regard to the EESC: EESC being a better fit parameter vs avoid EESC in the fitting

This is related to the question of how to define what the best set of independent variables is, and how to determine independent variable errors.

There are various ways to determine the best independent variables. One which is often used is to look at the post-regression statistics and define the best fit model based on the best correlations and smallest residuals.

However, that may be misleading, as it does not consider the independent variable errors.

Alternatively, one can study the independent variable errors – as done in this paper – and conclude that there are structural uncertainties in variables  (here: EESC).

That the EESC fit results in the best fit may be related to the fit models used: the EESC implicitly considers the possibility of saturation of ozone depletion around the EESC peak. A piece-wise linear fit does not, so the change before and after the break is rather abrupt, leading to potential fit issues around the maximum.

Obviously there would be other ways to for fitting linear fits (for example a 3-section linear fit rather than the 2-section fit applied here). This is part of thinking about how to improve the process, or alternatively, to start thinking about other methods than a multivariate regression on total ozone to determine stratospheric ozone recovery. However, we argue that both aspects beyond the scope of our current paper, which we view as a starting point for future research.

*We tried to clarify this in several parts of the revised paper.*

**Minor comments**

Minor comments are addressed accordingly. Below only follow comments that require a more detailed response.

- Abstract: shortened in line with the request by the referee.

- Added a table with the url's of the data sources

- Description of how trends are calculated from the EESC fit is added (see extended discussion in answer to referee #1)

- Solar flux, QBO and solar flux – QBO index. We moved sections 2.3 and 2.4, discussing the solar flux and QBO separately to the supplementary information. A sentences was added at the end of the section discussing the combined solar flux – QBO index noting that uncertainties in the individual solar flux and QBO index are considerably smaller (see SI) than the uncertainty of the combined solar flux – QBO index.

- Reference to scenarios "above". There is a little misunderstanding, as the scenarios referred here are the volcanic aerosol scenarios, not the ozone scenarios. Text was modified for clarificiation.

- Brief description of the MSR dataset added to the now section 2.7 (prev. 2.9).

- A table was added with online data sources (upgraded from the supplementary information)

- Trend errors in the Kuttippurath et al. [2013] study vs. our estimates. See discussion in response to referee #1. We have no proper explanation, other than Kuttippurath et al. [2013] apparently has a different method for determining trend errors. We simply took the EESC scenario multiplied with the regression value, and then applied an Ordinary Linear Regression to the pre-break and post-break periods separately. For the post-break period the trend errors are comparable, but for the pre-break period they are not. Kuttippurath et al. [2013] do not provide a description of how they derive at their trend errors.

- Figure 4: caption states 1979-1999 and 2000-2012. This is incorrect, also noted by referee #1. The figure was adjusted accordingly.

- Mean trends (EESC, PWLT) and 95% CI values are added to section 3.3

- Supplementary figures are not needed any more (no LINT trends after proper impelementation of PWLT in regression).

- Section with reference to Chehade et al. [2014] was deleted for various reasons.

- The discussion section was completely rewritten (for reasons outlined in the general remarks to the editor).

- See discussion on persistent features in ozone and EP flux scenarios. We hypothesize that most likely the 21-30 period is so short that variability in vortex dynamics start to play a role, as well as exposure to solar radiation. A bit like just using one day of each year to discuss long term trends in ozone. A somewhat different shape of the vortex or how long the parts of the vortex have been exposed to solar radiation may considerably affect then can result in considerable differences in vortex mean ozone. (not added to the revised document, as this is merely an untested hypothesis).

- Figure 9: grey bars were removed.

**Referee #3**

Major concerns

(1) Confusion about what the goal of the paper is (NOT the best explanatory set of variables, but trend significance).

*In the revised paper and with implementation of the proper PWLT regression the discussion of optimal explanatory variables has shrunk in favor of discussion of trend significance. See also our answer to the other referees.*

(2) In the revision the main message is better worded and should be more consistent.

(3) Added a brief explanation about how the trend is calculated from the EESC. Note that there is an unresolved issue with EESC-based trend uncertainties in our paper and that of Kuttippurath et al [2013], the latter not providing sufficient information to resolve it.

(4) Better motivation as why to dismiss the EESC added. The EESC is a pre-defined function that does now allow much flexibility in ozone trends. PWLT is better suited to cope with that.

(5) Results now provide indications why certain ozone time series perform “better” than others and this is discussed. Keep in mind that we merely extend on what others have done. In a separate paper [Knibbe et al., 2014; ACP] we analyse the geographical distribution lon-lat of all local ozone time series for all months for a 30 year period. Such an analysis comes with its own intricacies which are not the scope of this paper, but does not solve the issue of trend significance and regression uncertainties. If anything it could even be worse, because then one would have to consider additional regressions as there may be other processes affecting ozone outside of the ozone hole season. Note that Knibbe et al. [2014] does not consider regression uncertainties.

**Minor comments**

Minor comments are addressed accordingly. Below only follow comments that require a more detailed response.

- Brief description of the MSR dataset is added to the now section 2.7 on ozone scenarios.

- The separate QBO and solar flux descriptions have been moved to the supplementary information ( see also referee #2) to avoid confusion. A brief remark is added discussing uncertainties of the individual QBO and solar flux variables, referring to the SI for additional information.

- A table was added with online data sources (upgraded from the supplementary information)

- Differences between pre-break trend uncertainties in Kuttippurath et al [2013] and this study are discussed in our response to referees #1 and #2. It is unclear where the differences come from as information is lacking in Kuttippurath et al [2013] on their calculation of EESC trend uncertainties, but it is clear that it differs from ours. See further our reponse to R#1 and R#2.

- Figure 4 is adjusted accordingly.

- The EESC regression is determined by the pre-BREAK period as the pre-BREAK trend distribution does not show a tri-modal shape as seen for the post-BREAK distribution.

- Auto-correlation. We made a small adjustment to the paragraph, also based on comments by the other referees. Basically, ozone records show a certain autocorrelation. The regression turns out to remove this, as the 1-year lag auto-correlation of the residuals is near zero for all regressions. If that would not have been the case then a correction would be needed for the PWLT trend uncertainties. Since this is not the case, a correction is not needed.

- The large sensitivity of the post-BREAK EESC-based trends to the exact EESC shape is the reason to not prefer EESC-based trends. This has been reworded.

- Volcanic aerosols. This section has been rewritten. In essence, it is unclear from the regressions what even the sign of the ozone-aerosol effect is. Hence, no reason to include volcanic aerosols.

- Added a remark about figure 9 being similar to figure 5 but with larger correlation bins for visualization purposes.

- Table 5. It is better explained in the text which time periods appear more relevant, and this is also summarized in the conclusions.

- Figure 2 now includes a legend with all 16 scenarios and a description which scenarios is what.

- Dark blue and red lines in figure 5 are explained in the figure caption.

## References

HOLTON, J., and H.T. Tan (1980), The influence of the equatorial Quasi-Biennial Oscillation on the global circulation at 50 mb*., J. Atmos. Sci. 37*, 2200–2208.