Response to Reviewer No. 1

We thank the reviewer for his/her helpful comments. The reviewer's comments are repeated below in italics.

Summary:

This paper proposed an approach to merge the global mean SSU and AMSU observations using MIPAS as a bridge, accounting for residual biases between the two observations which were left over from directly merging the weighting functions of the two observations. The merging results were reasonably well, given by near zero difference time series between the SSU observations and its equivalent merged AMSU observations during their entire overlapping period from 2001-2006. The paper then compared the merged SSU/AMSU observations with MLS observations and chemistry-climate model simulations. Favorable results were obtained for both of the comparisons. The approach is potentially useful for other nadir-sounding observations even when no overlaps exist between them as long as a bridge exists from the limb-soundings to link the nadir-sounding observations to be merged. Although the merging approach appeared to be useful and the results were interesting, I have a major concern on the conclusion and some specific comments relate to it. I recommend publishing the paper only after these issues were suitably addressed. These comments are listed below.

Major Comments:

In the abstract, the authors stated that the 'The extended SSU data set also compares well with chemistry-climate model simulations over its entire record' and Figure 7 gave the comparison results. However, the chemistry climate model simulations in Figure 7 extended only from 1979 to 2006 where the SSU-only observations were available for the entire period. It is desirable to understand how the SSU-only observations compare with climate model simulations. In my view, this conclusion has nothing to do with the merging and extension. Good agreement between the merged SSU/AMSU observations and chemistry climate model simulations should be due to good agreement between the NOAA SSU Version 2 data and model simulations. This is because only if the SSU observations agree well with model simulations, good merging between SSU and AMSU could also agree with model simulations. Otherwise, good agreement between the merged observations and the model simulations cannot be achieved unless the merging is poor. To my knowledge, independent studies by other groups (not published yet) also suggested that the NOAA Version 2 SSU data agreed very well with model simulations. It appears that the conclusion of this paper claims a credit or achievement that does not belong to the proposed merging approach. This should be addressed in the revised manuscript.

The reviewer seems to be saying that in the abstract we suggested that the good agreement between the extended SSU data set and the models was somehow due to our merging procedure. Rather we were simply describing the characteristics of the extended SSU data set, and not using the comparison to models as a validation of our method. Our meaning was clearly stated in the body of the paper: On lines 5 and 6 of page 10102 of the Discussion paper we clearly stated that the CCMVal2 models cannot be used to compare to the extended SSU results post 2006, since the CCMVal2 simulations end in 2006. In the Conclusions (lines 10-13 of p. 10106 of the Discussion paper) we stated that the good agreement between SSU and the models is because we were using the Version 2 data, not Version 1 in Thompson et al. (2012).

In the Conclusions we also discussed the agreement between our extended SSU data and the CMIP5 model results shown in Thompson et al. Unfortunately that comparison had to be made by visually comparing Fig. 7 from our paper to Fig. 1 of theirs, which we now realize was asking too much of the reader. We have therefore included the CMIP5 data in our paper, the results of which are now shown in Figs. 7 and 10. This allows us to discuss the comparison with models over the extended period 1980-2012 in the revised abstract, which should help alleviate any misunderstanding.

To further avoid any misunderstanding, we have also added a statement in the third last paragraph of the Introduction that "the merged data set can only be as good as the component data sets going in, and relies on the extensive efforts spent on homogenizing the SSU and AMSU data records themselves."

Specific comments:

1. Page 4, lines 10-15: It should be noted that the NOAA Version 2 SSU data and its paper by Zou et al. (2014) intended to address the differences between climate model simulations and observations found in Thompson et al. (2012) paper. The NOAA SSU Version 2 data had big improvement over its Version 1 data and how this improvement resolves the mystery of the stratospheric temperature trends need to be addressed here since, as mentioned in my major comment, this is a key issue that will affect the SSU/AMSU merging and its comparisons with climate model simulations.

In the revised manuscript we have added a paragraph in the Introduction in which we discuss the Zou et al. (2012) results and make clear the reason for the development of NOAA SSU Version 2.

2. Page 5, lines 20-25: Again, the NOAA SSU V2 data are critical for interpretation of the merged SSU/AMSU data and its comparisons with climate model simulations. More information should be provided here on this dataset such as its improvement over V1 and spatial and temporal resolutions, etc.

In Section 2.1 we have added a brief discussion of the main differences between SSU V2 and V1. There we also now explain that the motivation behind the development of V2 was to address the reasons for the mismatch between V1 and SSU Met Office and between V1 and models discussed by Thompson et al (2012).

3. Page 6, line 5: I noticed that most of the results presented later on were from 2001. The authors should clarify here how starting dates of different channels affect the merging.

Without AMSU channel 14 (which is not available until 2001), the approximation of SSU channel 3 is hopelessly bad, and even that of channel 2 is compromised. Thus we can only consider merging from 2001 onwards when channel 14 is available. In any case the c_n's are computed with the use of MIPAS data, which starts only in 2002.

4. Page 8, lines 1-5: ERA-I and its adjustment were not fully justified for long-term trend investigations, especially for SSU channel 3. I don't see how including this dataset will help justify the conclusions of this study. I recommend removing ERA-I to focus on the SSU/AMSU merging and its related analysis in this paper.

We agree and have therefore removed the section on ERA-I as well as the SSU-weighted ERA-I results shown in Fig. 7.

5. Page 18, lines 1-5: As mentioned in my major comments, the good agreement in Fig. 7 has nothing to do with the SSU/AMSU merging. I recommend to only including the SSU V2 data in this comparison (Fig. 7), while adding another plot to compare the merged SSU/AMSU data with CMIP5 data where one can extend the time series to 2011 or present.

Figure 7 now shows only the CMIP5 models and the extended SSU data set for the 1979 to 2012 period.

6. Page 18, lines 5-15: See my comment #5. Since the authors already did this, I recommend to including the plot showing comparisons between CMIP5 and the merged SSU/AMSU data from 1979-2011.

Done (see above). We have also revised Fig. 10, which now shows extended SSU and CMIP5 trends for 1980-2012.

Response to Reviewer No. 2

We thank the reviewer for his/her helpful comments. The reviewer's comments are repeated below in italics.

Summary:

McLandress et al develop a method of intercalibrating distinct remote sensing measurements that have different vertical sampling. The method assumes a constant bias between different instrument measurements as well as a constant bias to account for errors in applying the weighting function to the data. The method is developed for generic datasets (though clearly with SSU/AMSU intercalibration in mind) and allows for the use of a transfer function with a third independent dataset to link two datasets (e.g. linking AMSU to SSU via MIPAS).This transfer function method is applicable in the case that the two datasets to be intercalibrated do not overlap in time. McLandress et al are able to continue SSU measurements that end in 2006 with AMSU measurements that begin in 2001 and continue through the present. Even though the two datasets have significant overlap, the authors use MIPAS data, which allows for a quantification of the biases incurred from the application and fitting of the weighting functions. The authors demonstrate that it can be misleading to compare vertical profiles with deep atmospheric layer measurements, that SSU/AMSU/MLS data are in good agreement during overlap periods, decadal variability is strongly modulated by anthropogenic and natural forcing, and that models and SSU observations largely agree. This work is valuable in that it extends the observational record to allow comparison with model simulations of the stratosphere, extends a long-term stratospheric climate dataset to the present, and develops a simple framework for connecting various remote sensing measurements that have different vertical sampling. Despite these strengths, the model-observational comparison is not sufficiently novel, the manuscript should provide more background on the datasets in use, the use of a third dataset (MIPAS) to calibrate SSU with AMSU is not sufficiently motivated since these datasets have significant overlap and because the MIPAS data is unstable, and there is no uncertainty analysis for the derived constant biases and its effect on the resulting time series.

Recommendation: I recommend that this work be considered for publication after the following major and specific comments are addressed.

Major Comments:

For clarity the reviewer's single long paragraph has been broken into four smaller ones which we have numbered.

(1) *The abstract insinuates that this extended-SSU time series compares well with chemistry climate models, but the extended time series has nothing to do with this agreement (since SSU is phased out after the time periods mentioned). Isn't this finding largely a result of the updated (e.g. v2.0) NOAA STAR data? Incorporating the few CMIP5 models that have data for MSU 4, SSU 1 and 2 could make this analysis more*

valuable (this would be historical + RCP scenario).

The revised paper now includes results from the CMIP5 models that have top data levels at or above 1 hPa. The changes to the manuscript consist of a new subsection in Section 2 briefly describing the models, a table listing the relevant characteristics of the models, a revised version of Figure 7 (and accompanying discussion) showing the CMIP5 results and extended SSU data, and a revised version of Figure 10 (and accompanying discussion) showing the CMIP5 and extended SSU trends. This allows us to discuss the comparison with models over the extended period 1980-2012 in the revised abstract, which should help alleviate any misunderstanding.

(2) *There should be more background provided on these datasets, especially the SSU. This includes: 1) a basic summary of the corrections made to the data (e.g. limb correction); 2) the type of measurement (e.g. microwave); 3) spatial coverage; 4) brief description of why NOAA SSU v1.0 and v2.0 are so different, since the agreement of SSU v2.0 and CCMVal2 is a major finding in this work; 5) The Wang and Zou, 2014 paper is for NOAA STAR AMSU version 2.0. Can this data be updated to the most recent version?*

Section 2 now includes a more detailed description of the data sets (primarily SSU and AMSU) in the manner suggested by the reviewer. The AMSU Version 2.0 data for channels 10-14 is not available on the NOAA STAR website – only data for merged MSU and AMSU (ch 5-9) are – hence we stick with what we have (Version 1.0).

(3) *The text does not explain the time averaging periods in Eqs. 6 – 9. Is a common reference period used? If not, isn't the large drifts in MIPAS an issue for this intercalibation? If so, why bother with MIPAS? Further, should we trust that these two versions of MIPAS data (presumably on two different platforms) are absolutely consistent? In general, cn is not important, but the stability of the constant offset is very important. The agreement between MLS and AMSU is impressive, if the authors desire a proof-of-concept, why not use MLS data? The inclusion of MIPAS detracts from what I see as the main value in this work (extending SSU). MIPAS is valuable if SSU and AMSU don't overlap (but they do overlap) and/or if it is stable (it isn't). The transfer function framework is theoretically valuable, but a convincing case is not made for it, and it makes the manuscript much less clean. Couldn't the weighting function error budget be computed using reanalysis (or some reference atmosphere)? It takes some effort to work through this framework, even though the idea is fairly intuitive. So I suggest the authors simply remove the transfer function framework, move it to an appendix, or potentially relabel A, B, C as AMSU, SSU, MIPAS; the analysis can be done merging AMSU directly to SSU. The offsets in merging different datasets can have a large effect on trends.*

The averaging period used is the overlap between MIPAS, SSU and AMSU. This is now mentioned in the text.

Since the period over which the time averages of the three bias terms in Eq. (6) are computed is the same, drifts are not an issue. (If the averaging periods used in Eq. (6) had been different, the drifts in the MIPAS data would indeed have been a problem.) For the same reason, the MIPAS temperature data is not actually needed in merging SSU and

AMSU. The reason we nevertheless include the MIPAS data is to demonstrate our methodology for using a limb-sounding data set as a transfer function between nadir-sounding data sets. Note that this furthermore allows an explicit calculation of the bias arising from the different weighting functions of SSU and AMSU (E_W in Table 3), which would not otherwise be obtainable (without the limb-sounding data, only the sum c_n is obtainable). The reason we do not use MLS as a transfer function is because there is only one year of overlap between it and SSU, and we were concerned that a reliable estimate of the bias term given by Eq. (7) could not be determined. We have now added a statement to this effect.

We strongly disagree with the suggestion to move the transfer function methodology, and the comparison with MIPAS, into an appendix. (Anyway such restructuring of a paper is largely a matter of taste.) In our view, the transfer function methodology is a significant novelty, which can be applied to other nadir measurements, not just of temperature. Using MIPAS to illustrate how the different bias terms can be obtained is consequently something that belongs in the body of the paper.

(4) *Uncertainty analysis should be incorporated into the offsets and this should be propagated into the derived trends/timeseries*.

We have now done this, estimating the impact on the trends by recomputing the trends with errors added to the offsets, and combining those uncertainties (at 95% confidence level) with the raw trend uncertainties. The resulting trend uncertainties for 1980-2012 are increased slightly (e.g., from 0.12 to 0.13 K decade^{-1} for channel 1); those for 2004-2012 are unchanged within the stated precision.

Specific Comments:

10087 13-15: Provide a reference for the discussion of GPS occultation discussion

We have added a reference to Wickert et al. (GRL, 2001).

10087 20: The MSU Channel 4 record is continued by AMSU channel 9 (e.g. Christy et al. 2003, Mears et al. 2009)

These two papers are now referred to.

10091 22 – 25: More detail about the exclusion criteria is needed. What was the top boundary condition required and how many models were eliminated because of this? Ideally, you could just ignore/mask temperatures if they are extrapolated into the Antarctic land mass, but you should at least tell the reader how many models were thrown out because of this? What do you mean the models were outliers, how many models does this include, and why should they be excluded?

The revised manuscript now uses 16 CCMVal2 models, all with tops above 1 hPa (the criterion used by Thompson et al for deriving all three SSU channels from the CMIP5 models). Three CCMVal2 models were excluded: two had tops at 3.5 hPa (CAM3.5) and

10 hPa (E39CA), the third (UMSLIMCAT) because the netCDF file containing the zonal mean temperatures did not come with a latitude array. (We note that with additional detective work the issue with the UMSLIMCAT data could be resolved, but we decided that it was not worth the effort given that we have included more CCMVal2 models in our analysis than did Thompson et al. (2012)). For data sets with missing data flags at points below the Earth's surface, those missing data points are filled using the constant value from the first good data point above. Since these data points are at high latitudes (Antarctica) and far below the peak of the lowest SSU channel weighting function, the precise manner in which they are filled has no impact on the SSU-weighted near-global mean temperatures. All this is now mentioned in the revised manuscript.

10092 15 – 16: Why did you use near-global? AMSU measurements go to _82 N/S, does another dataset only go to 75 degrees N/S?

The reason we used 75 deg N/S was because that is the latitude range for MIPAS. Since the MIPAS data (or more precisely the MIPAS temporal sampling) was used in computing the offsets (c_n) we have used this latitude range for analyzing AMSU, SSU and the model data.

10093 7: Does choosing a single H for the globe affect the results? For instance, if you chose another reasonable H-value (e.g. 7.5) does the resulting weighted time series have a different trend or variability that could effect the results? Would H(latitude, season) make this more accurate? The concern is that the time series may be sensitive to the assumption of a global mean H for the weighting function. This modulates the weighting function height and the authors show that trends vary quite strongly as a function of height. The global mean weighting function applied to near-global data is borderline overly simplistic since the spatial structure may be important to the results and uncertainties in this work.

The scale height H does not come into play in the method (Eqs. 3 and 4) we propose for merging the two nadir temperature data sets. Equation 3 fits the weighting functions of instrument A to the weighting functions of instrument B, while Eq. 4 uses the coefficients from the weighting function fit to generate the T_n's for instrument A. In neither of these two equations or in Eqs. 6-9 for the offsets does H appear.

10094 4 – 6: I do not agree that the agreement between MIPAS and SSU is uniformly "quite good." They have a large offset, relatively large residuals, and a large inter-dataset trend.

We agree that the wording was too vague, and has been revised. The good agreement applies to the seasonal cycle; we had earlier mentioned the offset, and now also mention the inter-dataset trend.

Eq. 10: Is there a sign error in this equation?

Yes, there should be a minus sign between the two terms. This has been corrected.

Section 3.1.3: MLS has excellent agreement with AMSU and SSU, did you consider using it as a transfer function? It appears to have a little more than a year of overlap with SSU and several years of overlap with AMSU and apparently has fewer known biases than MIPAS.

As explained above the reason we did not use MLS as a transfer function is because there is only one year of overlap between it and SSU.

10098 21: The individual biases are “less than 1.2 K,” less than “_1 K” is misleading.

We have changed this to “less than 1.2 K.”

10101 6-7: The trend profile should be described without using “(tilted) axes.”

This sentence has been reworded.

10101 19 – 21: I’m not sure how “given the similarity of the extended SSU and AMSU trends” tells us anything about whether the strong vertical structure of the MLS trend profile – can you explain?

This phrase has been deleted.

10101 23: Why did you calculate anomalies with respect to 1980 – 1985? This makes the later comparison (10102 21 – 25) more difficult to see.

The reference period of 1980-85 was used in order to compare our Fig. 7 to Fig. 1 of Thompson et al. (2012). Since we have decided to delete the discussion of ERA-I, which are the lines the reviewer is referring to here, the reviewer’s concern over our choice of the earlier reference period is no longer relevant. We also note that the revised version of Fig. 7 uses 1979-1981 for the reference period to avoid inclusion of data in 1982 when El Chichon erupted.

10101 25 – 29: Are the weighting function normalized if the model top is below the top of the weighting function?

Yes, this is now mentioned in the text where the CCMVal2 results are first discussed. We note that normalizing the weighting functions has a negligible impact on the weighted temperatures since the models all have tops above ~70 km.

10102: Why isn’t ERA-I extended through to the end of extended SSU? Can you say something about whether ERA-I should be considered a reasonable comparison dataset (my default assumption is that reanalysis is generally not trustworthy for longterm time series)?

Since the other reviewer commented on the unreliability of ERA-I data for long-term trends, we have decided to remove all discussion of the ERA-I data.

10103 20 – 24: Please list the major ideas for why there may be discrepancies in lower stratospheric cooling. Some relevant papers may include, Lu et al., 2014, Solomon et al., 2011, Shindell et al, 2013, Eyring et al. 2013, Hassler et al. (2013), and others.

Since the discrepancy in lower stratospheric cooling is within the stated uncertainties and is not seen in the CMIP5 results (Fig. 10a), we have deleted this discussion. The only reason for showing the CCMVal2 results in this figure is to demonstrate that weighted trends should not be compared to profile trends.

10104 17 – 18: It was not immediately obvious what you meant by the the correlation between channels is geophysical. Something like "interannual temperature variability is coherent with altitude" or something similar might be more intuitive.

We agree, and have replaced "is geophysical" with "reflects strong vertical relationships in the variability of global-mean temperature", which is the wording used earlier in the paper.

Discussion: Given the poor weighting function representation of SSU3 using AMSU and because trends vary as a function of height, some caution should be suggested regarding the results for SSU 3.

We have added some text to this effect in the Discussion.

Figure 9: Do the model trends look anything like this? It would be useful to try to shed light on this weird zig-zag structure being an observational issue or model issue.

Model trends (e.g. Fig. 10) do not exhibit the zig-zag structure in the vertical. It could be an observational issue, although we are not knowledgeable enough about the MLS data to make a more definitive statement.

Figure 10: Use 95% confidence intervals.

Figure 10 has been broken into two separate figures. New Fig. 10 now shows long-term linear trends instead of differences and 95% confidence levels instead of standard deviations. As in Fig. 9 the uncertainties are computed assuming serially correlated residuals. New Fig. 11 shows temperature differences for the two 10-year periods, and is identical to the two right panels of old Fig. 10 with the exception that the error bars now denote the 95% uncertainties which are computed under the (conservative) assumption that the standard deviation represents the standard error of the mean. Since Fig. 11 is used to contrast the temperature changes in the ozone depletion and recovery periods, comparing successive solar minima to avoid the effects of solar variability, it is perfectly fine to show temperature differences.