Response to reviewer 2:

We would like to thank the reviewer for their valuable comments and suggestions.

Reviewers comments are quoted in italic, our answers in roman, changes in the manuscript are in courier font. Page and line numbers refer to the original AMTD publication.

Main comments:

1. Motivation and novelty

Several studies dedicated to assessment of the combined (A)MSU climate temperature record with radio-occultation (RO) data have been performed in the past (Ho et al., 2009; Ladstädter et al., 2011; Steiner et al., 2011). Since RO data dominate in the new created dataset, it should be clarified why new dataset suits better for the study (see also comments below).

Regarding the choice of the data sets, this is answered below under point 2. The goal of our analysis is not to validate the (A)MSU climate temperature record. Had it been, then quite probably, as the reviewer suggests, the use of only RO data may have provided a better comparison standard. Rather, we combined temperatures from different data sources to create a new, multi-instrument temperature climate data record (CDR), and as one possible application used this new CDR to validate the robustness of the merging of the MSU4 and AMSU9 records.

P. 238, l. 14, insert:

The purpose of this study is to work towards the inclusion of ESA and ESA-TPM data in stratospheric climate data records (CDRs). As a number of key US-based CDRs ended in 2005/2006 while ESA and ESA-TPM vertical profile observation records begin a few years before 2005/2006 and continue to the present, ESA and ESA-TPM are potential candidates to extend existing CDRs in time. While similar comparisons of RO data only with (A)MSU have been performed (Ho et al. (2009), Ladstädter et al. (2011), Steiner et al. (2011)) the motivation to use temperatures from different data sources, not only RO, was to create a new, multi-instrument temperature CDR with smaller uncertainties. The temporal and spatial overlap with the RO data should be sufficient to account for systematic biases between the different instruments.

2. The choice of datasets

Datasets for merging

In general, to get the best climate data record, it is not necessarily to use as many data as possible. Instead, the best solution would be the measurements of the same type, which are coherent with each other and do not require calibration. For the purpose of the paper, this would be more logical to use the radio-occultation measurements, which can be used together without any correction. However, such analysis has been already performed (Ladstädter et al., 2011).

The dataset created by the authors is also "RO-dominated", but it has additional difficulties when merging measurements from other-type instruments (and thus differences in measured parameters (note that RO give "dry temperature") and uncertainties due to bias and drift correction, see also below). I think that the collection of datasets for merging not optimal. The motivation for using so diverse data (some of them are drifting due to instrument aging) should be explained.

The research and subsequent analyses presented in this paper was shaped in large part by the requirements of the ESA SPARC Initiative (SPIN) under which the work was funded. One of the main

objectives of SPIN was to combine ESA and ESA Third Party Missions (TPMs) data sets to improve the characterization of existing upper-air CDRs, and produce new CDRs. As such, the choice of data sets used was dictated primarily by the scope of the SPIN project.

Datasets for validation

- *Why do you prefer using the RATPAC-A dataset with relatively low sampling (85 stations) and insufficient information about the data averaging instead of larger radiosonde databases such as RAOBCORE (~1000 stations) or IGRA (~1500 stations)?*

IGRA appears to only supply monthly mean temperature series for individual stations, and not averaged over certain latitude zones. Furthermore, IGRA data are not homogeneity-adjusted. In contrast, RATPAC supplies homogeneity-adjusted data for a subset of IGRA stations, averaged over large latitudinal zones. For that reason, we considered RATPAC to be a better choice for comparison with our data set than IGRA.

While RAOBCORE provides gridded data, and could have been used for validation of our merged data set, the RAOBCORE data have not been through the thorough data quality screening that has been applied to the RATPAC-A data. Our intent was to use a smaller number of higher quality data rather than a larger number of unscreened data.

- *The differences with respect to the NCEPCFSR dataset are large. Should the readers make the conclusion that NCEPCFSR is wrong? If yes, why this dataset is used for validation of VRT?*

The non-linear scale of the temperature differences in Fig. 4 might make the differences appear larger than they are. As described in the text, differences are typically less than 2K for most pressure levels, which would not be considered large for limb sounding instruments. We did not imply that the reader should consider NCEPCFSR to be wrong but simply stated the differences found.

3. Merging method and merged dataset

1) For nearly all instruments used in this paper, the "native" vertical coordinate is altitude (the only exception is SMR). The effect of altitude-pressure conversion using the meteorological models should be mentioned and evaluated (long-term temperature records from meteorological models do not suit for trend analyses, as they have artificial jumps due to different assimilated data).

For our analyses, we used monthly mean zonal mean temperature climatologies that were created within the SPIN initiative as initial input for our merging process. All SPIN climatologies were compiled using the guidelines for creating monthly mean zonal mean climatologies as specified by the SPARC Data Initiative. Therefore, the same pressure levels were used as for the trace gas and aerosol climatologies of the SPARC Data Initiative. An evaluation of the effect of the altitude-pressure conversion conducted within the SPARC Data Initiative which was the source for much of the data used in our analyses, is beyond the scope of this paper.

P. 241, l. 17, insert:

The monthly mean zonal mean temperature records were derived in a format following the specifications of the SPARC Data Initiative (SPARC-DI) (Hegglin and Tegtmeier (2015)).

2) What is the reason for selection of relatively narrow latitude zones of 5°? This results in increased sampling error.

The reviewer is correct in stating that increasing the latitudinal resolution of the data sets increases the sampling error. This is true irrespective of the analyses being performed. It is always a

judgement call. More highly resolved data are likely to have higher utility but, as the reviewer points out, are likely to suffer from increased sampling error. For our analyses we felt that 5° zones provided an appropriate balance. This choice was congruent with that made by the SPARC Data Initiative.

3) Applying the correction of sampling bias to radio-occultation data and not applying it to ACE-FTS and SMR looks very strange (erroneous). I think, the sampling uncertainty should be either corrected for all instruments or taken into account as additional uncertainty (also for all instruments).

The ACE-FTS, SMR and MIPAS data sets used in our analyses were supplied to us via the SPIN without a mean bias correction having been applied. In conducting our analyses we faced two choices, viz.:

1. Apply sampling bias corrections to the RO data sets because, where possible, such biases should be corrected for, but acknowledging that this introduces an inconsistency in the RO and non-RO data sets used in this analysis.
2. Be consistent and do not apply sampling bias corrections to any of the data sets.

In our analyses we settled on option 1. Now, having reconsidered our choice in the light of the reviewer's comment we have decided to revise our analyses under option 2. We therefore repeated the analyses without any sampling bias correction being applied to any of the data sets.

All analyses had to be repeated. Consequently, all Figs. and Tables have been changed, and an additional Fig. and Table have been added. Please see the marked-up supplement of the full article for all the changes.

4) The dataset used for merging have different vertical resolution. This (substantial) difference in vertical resolution will not affect the mean value, but it will affect the estimates of uncertainty for the monthly zonal mean profiles by the standard error of the mean: the sample variance of the high-resolution profiles will be larger due to better resolved gravity-wave fluctuations.

The reviewer is correct and this has been accounted for in our analyses. Because in the merging process the temperatures are weighted by their uncertainty, data sets with higher uncertainty will be weighted less in the final temperature product. The uncertainty on the final product takes into account both the size and correlation between the uncertainties of the original data series.

5) From my visual perception, the fit by Eq.(2) does not agree with the differences shown in right panels of Fig.1 (especially for GRACE, TSX, ACE-FTS). This means that the bias-drift correction by the function (2) might introduce additional uncertainties. Look at the panel for SMR in Fig.1: uncorrected data agree better with the merged time series than the corrected ones in years 2002-2007!
Have you optimized the regression model? Are all parameters statistically significant?

We spent a considerable amount of time optimizing the statistical model to describe the differences. Not all coefficients are statistically significantly different from zero for every fit for each instrument at all pressure levels, but we believe overall that the model chosen is a good representation of the systematic differences. Once the model is chosen, even non-significant coefficients should be taken into account for the correction as they might bear some weight that in a different model would have been accounted for by a different (possibly significant) coefficient. As far as SMR is concerned, in this particular example, the fit reveals a large linear dependence over the time period of 2002 to 2012. While the corrected data in some years might appear worse, overall, the fit will minimize differences. Besides, as there are not many SMR data points, the effect on the final product will be very small.

6) Eq.(2) assumes that the drift is linear in time, which can be not realistic in reality. In general, the comparison with COSMIC presented in section 4.2 clearly indicates problems with the merged dataset.

We believe as a first order approximation a linear dependence is an adequate assumption for a drift term. We are not sure what particular problem the reviewer refers to with regards to COSMIC.

7) For the results, it would be interesting to see also the linear fit coefficients for the difference MSU4+AMSU9 and iVRT, as well as their comparison with analogous presented in (Ladstädter et al., 2011). Such information and comparison would be informative for both temperature trend estimates and demonstration of VRT capabilities.

We agree that such a comparison would be of interest. However, Ladstädter et al. (2011) use differences in monthly anomalies in relatively broad latitude zones (tropics, extra-tropics and 70° S to 70° N). In our comparison, we compare the differences in temperatures between the iVRT and MSU4+AMSU9 temperature records in each 5° latitude zone, directly (we do not calculate anomalies). As such, the linear fit coefficients are not directly comparable to the analysis in Ladstädter et al. (2011), and we do not think it would be helpful to list all 36 linear fit coefficients for both the RSS and UAH database.

Detailed comments

1) P.242, l. 1-2: "Additionally, the RO data were screened such that temperatures below 150 K were omitted as were temperatures above 330 K."
How often this occurs; what is the percentage of screened data? Do you remove only the value at some layer of the whole profile containing temperatures below 150K?

Unfortunately, in the calculation of the monthly mean zonal mean temperature data sets, the data rejections rates during the screening were not saved. We can confirm, however, that when RO data fell outside of the range listed above, only the individual values were excluded rather than the whole profile.

2) P.243, l.5-10. What is the reason of GRACE bias with respect to CHAMP and TSX? This contradicts with the study by Foelsche et al (2011), which report high consistency of the temperature climate records from multiple radio occultation satellites.

To investigate this question in detail is beyond the scope of the paper. The input data sets used in our study were validated as part of the SPIN project against CFSR and ERA-Interim reanalyses. The SPIN Product Validation Report (http://www.esa-spin.org/index.php/documents) notes as an unresolved issue that both CFSR and ERA-Interim reanalyses are much warmer than the RO datasets in the troposphere (Sect 4.1.1).

We can only make assumptions about what might explain the differences between results in Foelsche et al. (2011) and the climatologies produced for SPIN. Two possibilities come to mind:

1) The high level of consistency displayed between the different satellite RO datasets in Foelsche et al. (2011) was achieved by applying the same retrieval algorithm (processing scheme) for all datasets. GRACE, COSMIC and TSX climatologies that were produced for SPIN used different underlying retrieval algorithms.
2) Foelsche et al. (2011) used ECMWF analyses to remove sampling biases, while our RO data sets used NCEPCFSR data to do the same. That could have introduced a difference in the resulting climatologies.

3) P.244-245: "The monthly mean data from an instrument are excluded from this merging process if there are fewer than 4 measurements in a particular month or less than 5% of measurements of the month with the highest number of measurements within that year" For good fitting by Eq.(2), all seasons should be covered. Is this satisfied in all latitude zones for ACE-FTS?

With this being an occultation instrument with a repetitive orbit, it will cover the same/similar latitudes for a given day of the year, which ultimately means it will not cover all latitude zones in a given month. The minimum amount of time to achieve data coverage in all bins is three months, but that does not necessarily coincide with the same three months that are usually referred to as a season. As all available years of data are used for the fits, we believe that the coverage is adequate, though by no means ideal.

4) P. 251, l.8: Why the number of harmonics in the regression model is increased compared to Eq.(2)?

Eq. (2) fits differences at each altitude level, separately, with some sparse source data like ACE-FTS and SMR. For the second fit, the data will not be anywhere near as sparse because all six data sets have been merged and therefore the seasonality in the regression model fit coefficients can be better resolved. Additionally, the VRT data set is integrated vertically over all altitude levels so that there is only one fit necessary for each latitude zone. This allows a good fit for a larger number of coefficients.

5) P. 254, l. 24-26: "It has been shown that the uncertainty on the monthly mean zonal mean temperatures decreases with an increased number of instruments used in the merging." This is true only if the bias and drift correction is perfect. Otherwise, this statement is generally not correct (and Eq.(7) is not valid: RHS should be changed in order to account for differences in the mean profiles). Your statement on page 244 is more accurate.

We are not sure what the reviewer means here. Of course, no correction is ever perfect, but our corrections account for bias and drift to the extent possible. How this would invalidate Eq. (7) is not clear.

References

Foelsche, U., Scherllin-Pirscher, B., Ladstädter, F., Steiner, A. K. and Kirchengast, G.: Refractivity and temperature climate records from multiple radio occultation satellites consistent within 0.05%, Atmos. Meas. Tech., 4(9), 2007–2018, doi:10.5194/amt-4-2007-2011, 2011.

Hegglin, M. I. and Tegtmeier, S. (Eds.): SPARC Data Initiative, SPARC report on the evaluation of trace gas and aerosol climatologies from satellite limb sounders, in preparation, 2015.

Ho, S.-P., Goldberg, M., Kuo, Y.-H., Zou, C.-Z. and Schreiner, W.: Calibration of Temperature in the Lower Stratosphere from Microwave Measurements Using COSMIC Radio Occultation Data: Preliminary Results, Terr. Atmos. Ocean. Sci., 20(1), 87--100, doi:10.3319/TAO.2007.12.06.01(F3C), 2009.

Ladstädter, F., Steiner, A. K., Foelsche, U., Haimberger, L., Tavolato, C., and Kirchengast, G.: An assessment of differences in lower stratospheric temperature records from (A)MSU, radiosondes, and GPS radio occultation, Atmos. Meas. Tech., 4, 1965-1977, doi:10.5194/amt-4-1965-2011, 2011.

Steiner, A. K., Lackner, B. C., Ladstädter, F., Scherllin-Pirscher, B., Foelsche, U. and Kirchengast, G.: GPS radio occultation for climate monitoring and change detection, Radio Sci., 46(6), doi10.1029/2010RS004614, 2011.