Reply to the interactive comment by the Anonymous Referee #1 on "Radiosounding HARMonization (RHARM): a new homogenized dataset of radiosounding temperature, humidity and wind profiles with uncertainty" by Fabio Madonna et al.

The authors of the manuscript by F. Madonna et al examined in detail the comments provided by the anonymous referee #1 and acknowledge his/her effort in providing comments aimed at improving the robustness of the manuscript.
Below, the authors report in bold their replies to the reviewer's comments

The authors tried to obtain a homogeneous data set of temperature, humidity and wind profile through developing a Radiosounding HARMonization (RHARM) approach. This approach first post-processes the data since 2004, and then detect and adjust systematic errors in the historical observations with the help of the documented metadata. This involves one important issue, that is, the stations used in the study have complete metadata records. How complete are the metadata records for the stations used? Incomplete metadata records will lead to many unadjusted jumps in the final data set, which can't be used for climate research and other applications, similar to the raw dataset.

**As described at lines 115-129 of the manuscript discussion version "*In the frame of the Copernicus Climate Change Service (C3S) activities, a novel algorithm has been designed and implemented for homogenization of historical radiosounding data records available since 1978 (earlier records are not assessed due to the more heterogeneous data availability at mandatory levels before). The approach is named RHARM (Radiosounding HARMonization) and it is based on two main steps:*

a. Adjustment of systematic effects and quantification of uncertainties by adjusting the radiosounding observations of temperature, humidity and wind from 2004 to present using the GRUAN data and algorithms as well as the 2010 WMO/CIMO radiosonde intercomparison dataset [hereinafter ID2010, Nash et al. 2011], made available upon agreement with WMO;

b. Identification of change-points in the time series and adjustment of non-climatic (systematic) effects using statistical methods with related quantification of uncertainties in the historical observations. "

Therefore, in the present manuscript we are discussing the section of the RHARM algorithm obtained by adjusting the RS92 Vaisala radiosondes and all the other radiosonde types involved in WMO-CIMO 2010 radiosonde intercomparison. The identification of these sonde types within IGRA is based on the metadata collected by IGRA (i.e. the radiosonde code related to the WMO 3685 table) and on the use of the TAC code made available in the high-resolution BUFR files provided by an increasing number of stations since 2016 (about 1-2 hundred at present) and supplied for RHARM by ECWMF; these BUFR files include extensive metadata and a larger number of pressure levels for each reported radiosounding launch.

The selected radiosondes types are generally available since 2004 depending on the station. Each of these radiosondes has been post-processed using a combination of physical and statistical adjustments to remove various effects, e.g. the solar radiation bias or the bias due to the factory calibration. The post-processing applied to the radiosonde profiles allows the RHARM dataset to be a high-quality dataset which may be potentially used as a benchmark for intercomparisons and

validation purposes, including the validation of time series homogenization algorithms. The goal of the RHARM post-processed data discussed in the manuscript is to bridge the gap to a reference-quality data processing like that provided by the GCOS Reference Upper-Air Network (GRUAN). We have improved the text wording in a few sections to further clarify these conceptual aspects.

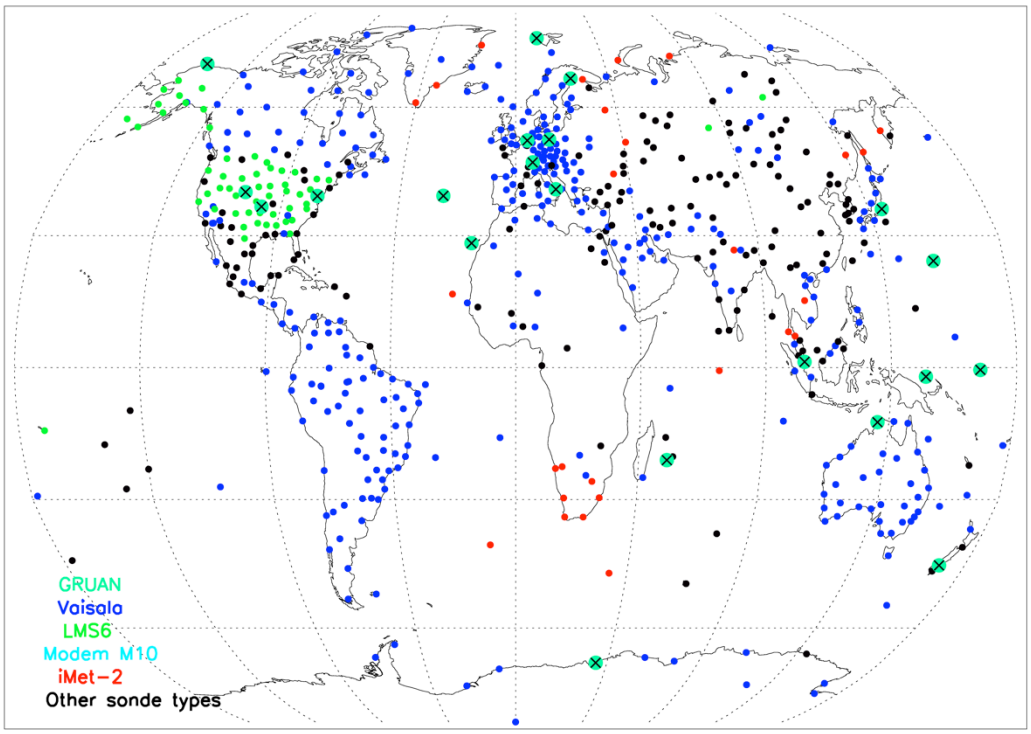
The homogenization of earlier data is mentioned in this paper to create a link with a second paper, currently under preparation, describing the RHARM approach used for the radiosounding data before 2004, which is different and based on the use of a statistical method only and not with the help of the documented metadata. This is clarified in the introduction at the lines 127-129:

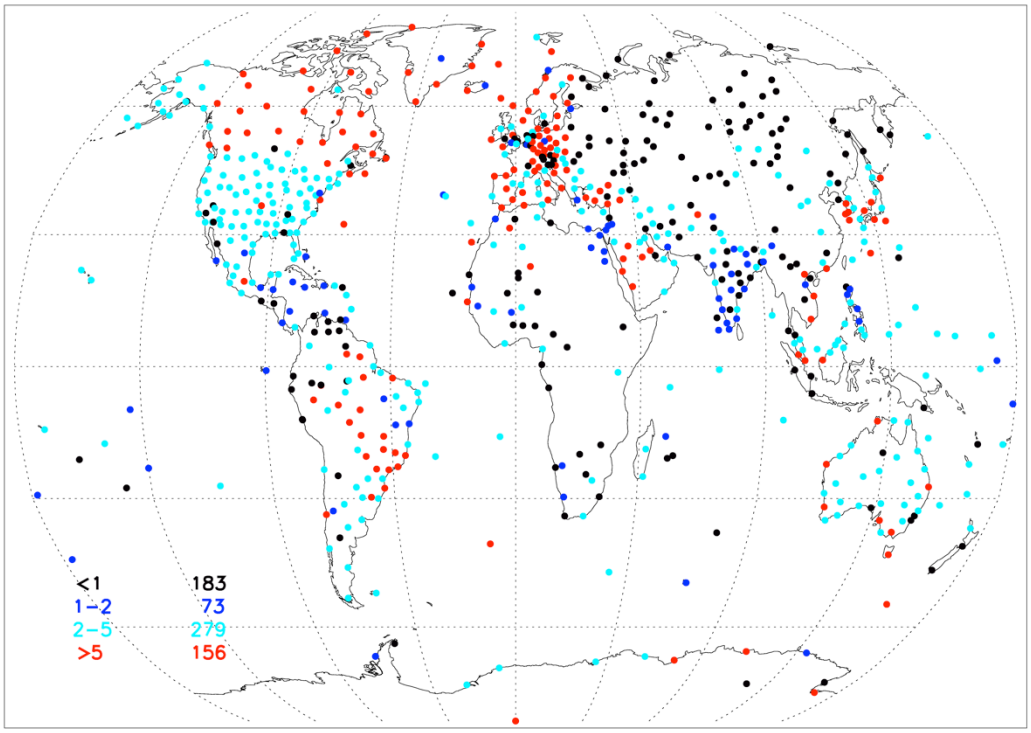
> **"b. Identification of change-points in the time series and adjustment of non-climatic (systematic) effects using statistical methods with related quantification of uncertainties in the historical observations.".**

At very few stations only, where a full reconstruction of the utilized radiosonde types was possible, metadata are used for validating the identification of change-points in the time series. As a consequence, to have a complete metadataset at each station is not a mandatory requirement to process historical data using the RHARM approach, but surely provides value and increases the volume of information available for the validation of the results. Nevertheless, we clarify that all the selected RHARM stations have complete metadata since 2000 (i.e. stations reporting a radiosonde code, see also the WMO table 3685).

Secondly, the data set derived from this study is mostly only since 2014 and lacks lots of stations in USA, Russia and China (Fig. 1) so as to prevent its applications in the future.

First of all, it is important to clarify that there was a mistake in the maps of Figure 1 and 2, which have been now modified as follows.





We apologize for this inconsistency with the real RHARM dataset due to a mistake in the decoding of the traditional alphanumeric codes (TAC) codes reported in the BUFR files used as one of the

metadata sources for the identification of the radiosonde types used at IGRA stations after 2013 (IGRA does not provide radiosonde codes after 2013). TAC codes are an essential metadata source indicating the radiosonde type used at each station at any given point in time.

The new maps clearly show that there is a large number of stations and radiosounding launches adjusted in the USA, while the number of stations in Siberia remains limited. This is due to the fact that Russian sondes (i.e. AVZ mainly) have been widely used in Siberia but these sondes use a poorly known data processing producing widely-acknowledged and poorly quantified random and systematic data issues (Ingleby, 2017, available at https://www.ecmwf.int/sites/default/files/elibrary/2017/17551-assessment-different-radiosonde-types-20152016.pdf). The adjustment of profiles provided by Russian radiosondes is challenging without accessing detailed documentation, performing laboratory tests or using radiosonde intercomparison data. As remarked above the main goal of the RHARM approach is to provide a high-quality benchmark radiosounding dataset with the estimation of uncertainties. The Russian radiosonde, based upon present knowledge cannot guarantee such quality and their post-processing using the RHARM approach is not feasible. One of the main future objectives will be to extend the number of radiosonde types adjustable using RHARM and, as a consequence, also the number of historical time series which can be completely homogenized (i.e since 1978).

We would underline that for the purpose of climate studies, as also mentioned in the manuscript it is already known that the quantity of measurements alone cannot address the value of a dataset for specific studies without a representativeness study (Weatherhead et al., 2017). We also point out that the existing homogenized temperature, relative humidity and wind datasets are often each characterized by specific geographical gaps, similar to RHARM. The use of an ensemble of homogenized datasets could be a powerful solution for climate studies in future, emulating what is done for reanalysis and climate models.

The manuscript has been updated according to the information reported above.

Third, the verification of the derived data set is far from enough. This manuscript only shows the mean differences and PDF comparisons (whose differences are not obvious), which is not easy to see improvements from your approach. The authors will seek other ways to show the improvements of the derived data set. Fig. 14 shows time series that are not significantly different from the IGRA raw dataset and your homogenized dataset. Does this imply that the RHARM approach does not make significant adjustments for the raw data set? This reminds me that whether this small difference is due to incomplete metadata records, so that some jumps have not been adjusted?

We already clarified in the comments above that the effect of the incomplete data/metadata record and the correlated occurrence of jumps is not an issue in the production of the RHARM dataset investigated in the manuscript. The post-processing of the selected IGRA data, using specific adjustments to remove biases in the radiosonde measurements, is not affected by the missing identification of jumps like in analysis of an entire time series with techniques like kriging, nearest neighbors, interpolations, etc.
Regarding the referee's comment that the RHARM approach does not make significant adjustments for the raw data set, the authors stated in the abstract (lines 37-40) that "The evaluation shows that the strongest benefit of RHARM compared to existing products is related to the substantive adjustments applied to relative humidity time series for values below 15% and

above 55% as well as to the provision of the uncertainties for all variables." and later in the text (lines 720-723): "For temperature, it appears evident that the applied adjustments minimally alter the pdf of IGRA original data: the small magnitude of the RHARM adjustments for temperature also indicates the enhanced quality of the data collected by most recent radiosonde types available on the market compared to the historical observations (Thorne et al., 2012)." The major added value brought by the RHARM dataset for the radiosonde data since 2004 is related to the estimation of the uncertainties for each single pressure level of each temperature, relative humidity and wind profiles. The availability of uncertainties is the only way in physics to quantitatively demonstrate if discrepancies between the radiosonde measurements and a second dataset used to validate or inter-compare are materially relevant or not and surely the most appropriate approach to generate a benchmark dataset.

For temperature, we clearly stated that the bias adjustment is small (for example 0.15 K at the tropics for the mean values). This result is also consistent with the content of the paper by Dirksen et al. 2014 (quoted in the manuscript) where the GRUAN Data Processing (GDP) is compared with the manufacturer data processing at the Lindenberg station, Germany. By construction, the performance of the RHARM approach is expected to be similar on average to the GDP. For historical data, we find that the applied adjustment for the temperatures recorded by radiosoundings from 1978 to 2004 are larger (paper in prep).

For relative humidity, the difference shown in pdfs of Figure 11, 12 and in Figure 13 of the old manuscript version where the comparison with the GDP is shown, is readily apparent and the correction of the well known radiosonde dry-bias in the upper troposphere appears to be largely reduced in the RHARM dataset. This result is also corroborated by the profile comparison shown in Figure 9.

To meet the reviewer's request to more clearly show the improvements of the derived data set, two further comparison/use cases have been added: to show the improvement derived from using the RHARM temperature data, the episodes of 2017 and 2018 Sudden Stratospheric Warming (SSW) events are shown using IGRA, RHARM and ERA5 data, while for relative humidity a comparison with MLS/AURA data at 300 hPa is discussed. These are reported in the Section 4.3 of the new version of the manuscript and summarized below for convenience.

The comparison in Figure R1 shows the time series of the daily averages obtained from ERA5 subsampled at the IGRA stations in the Polar European domain at 100 hPa and the corresponding times series obtained from IGRA and using the RHARM dataset in the period 2017-2018. During this period, in both Februaries two SSW events were observed and their effects in the lower stratosphere are pretty evident in Figure R1, i.e. a significant increase of the temperature with respect to the seasonal behavior. The comparison has been carried out at 100 hPa in the lower stratosphere to benefit from the larger number of radiosounding data available in the polar region than at greater heights.
For both the events, ERA5, IGRA and RHARM are in good agreement and within the RHARM combined uncertainty (vertical light gray lines) shown using a k = 3, where k is the k factor and represents an uncertainty of 3 standard deviations and approximately a 99% confidence level. Nevertheless, both the events are preceded by a strong cooling of the lower stratosphere which in 2017 is not well captured by IGRA, due to the warm bias affecting the unadjusted radiosoundings temperature profiles: in 2018, the discrepancy between RHARM and IGRA is similar to 2017 and due to the RHARM warm bias correction, while ERA5 is much warmer than

both IGRA and RHARM. In this case, the estimation of the uncertainty is the only way to quantitatively address the significance of any differences between the datasets. Other relevant differences among the three datasets can be pointed out in other months of the time series.

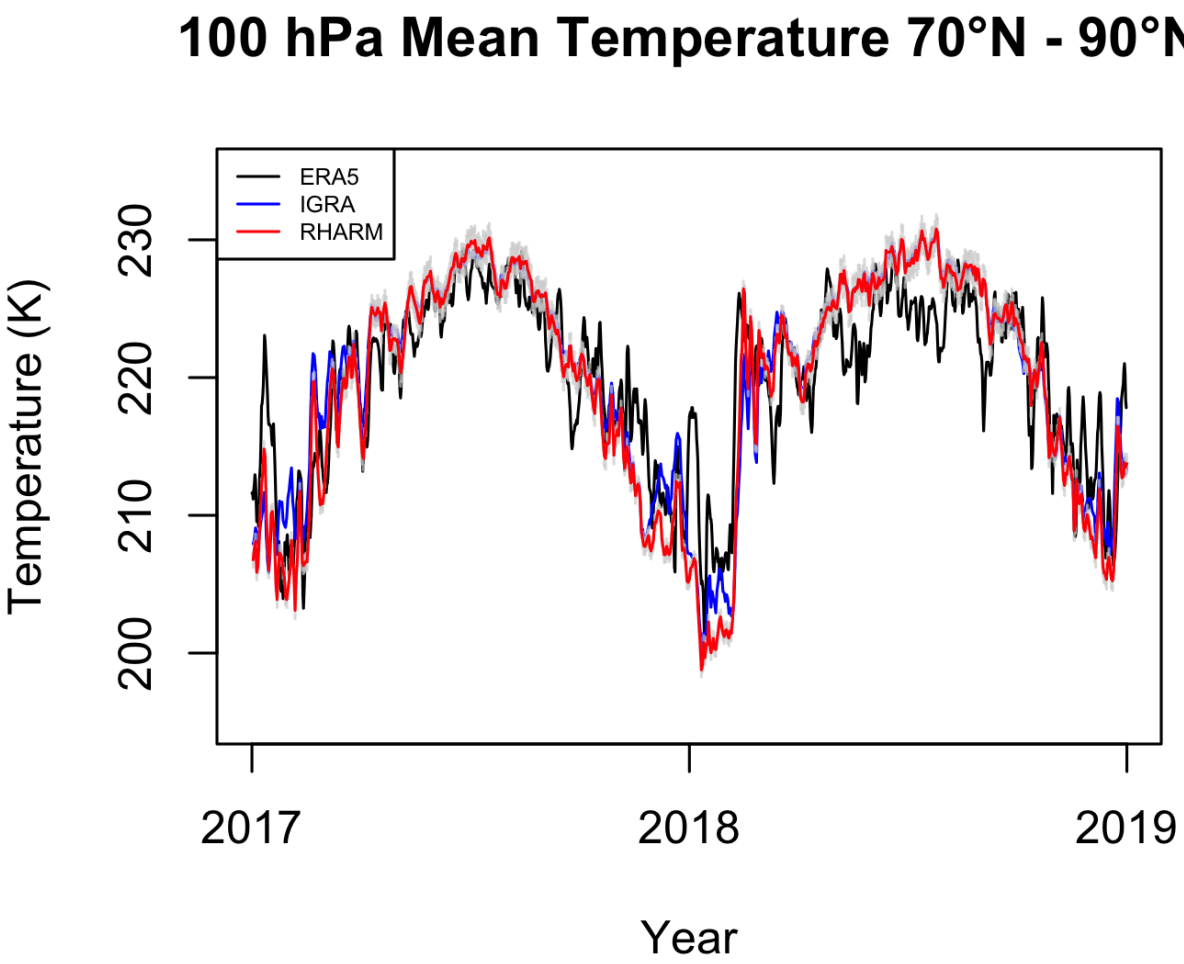


Figure R1: Comparison of time series of daily mean temperature at 100 hPa in the Polar European domain (70°N - 90°N; -10°W - 50°W) from 1-1-2017 to 31-12-2018. The black lines are the ERA5 daily averages obtained by subsampling the reanalysis data at the IGRA stations within the considered domain; the blues line shows the IGRA daily averages while the red line shows the corresponding RHARM averages. Grey shaded lines are the combined measurement uncertainties for RHARM. MLS data are from version 4, consisting of profiles reported on 12 pressure levels per decade between 1000 and 1 hPa, 6 pressure levels per decade between 1 and 0.1 hPa, and 3 pressure levels per decade between 0.1 and 0.01 hPa (Yan et al., 2016 ACP, doi:10.5194/amt-9-3547-2016).

Figure R2, shows comparison of the IGRA, RHARM, and MLS subsampled monthly time series of water vapour mixing ratio in the northern tropics from 2006 to 2019. For the comparison, the MLS time series at 316 hPa pressure level is chosen being close to the mandatory level at 300 hPa for the IGRA/RHARM radiosondes and also considering the good agreement of the MLS V4 data at this level with the Cryogenic Frostpoint Hygrometer (CFH) reported in the literature (Yan et al., 2016 ACP), which is still at present the only reference traceable instrument available for water vapour measurements in the UT/LS. The possible difference in the water vapour content between the two levels at 316 and 300 hPa must be considered as an additional colocation uncertainty contribution to the comparison shown in Figure R2.

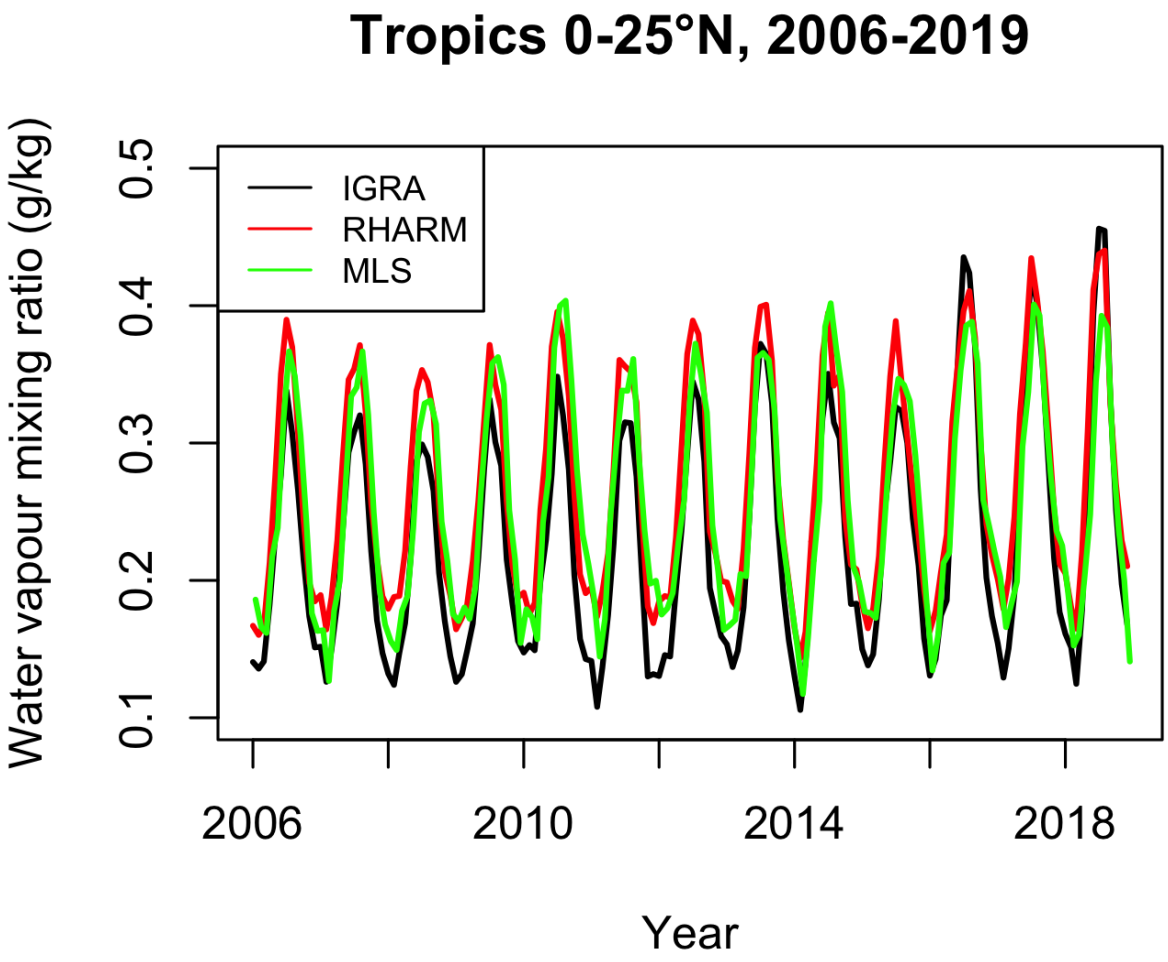


Figure R2: Comparison of monthly time series of the water vapour mixing ration zonal average in the northern tropics (0°N-25°N) from 1-2006 to 12-2019. The black lines show the IGRA time series, while the red and the green lines are the corresponding RHARM and MLS time series. MLS product has been subsampled at the IGRA stations within the considered domain.

Figure R2 reveals good agreement between RHARM and MLS and the efficacy of the RHARM dry bias correction, especially tangible at the lowest observed values of water vapour content. To show more quantitatively the progress achieved using the RHARM in the comparison with MLS, the mean difference, rms differences and Pearson's correlation coefficient for IGRA and MLS compared to those for RHARM and MLS are shown in the table below:

	mean difference (g/kg)	rms difference (g/kg)	correlation
IGRA-MLS	-0.03	0.03	0.95
RHARM-MLS	0.01	0.02	0.99

Lastly, the narrative structure of the piece and figure plots do not engender easy comprehension of claims, methods or main takeaways, which prevents a robust assessment. These may help improve the manuscript and requires extensive revisions.

The narrative style of the paper has been kept as much "standard" as possible, with few sections (introduction, datasets, methodology, results, etc.). Two flow diagrams have been also provided to facilitate the reader in following the provided description: each block in the flow diagrams has a one-to-one correspondence with the equations presented in the "methodology section". In the same section, results from the application of each single adjustment considered in the RHARM algorithm are quantified and shown in several plots. In our opinion, the paper, as a dataset description paper, must be very analytic to guide the users toward the full exploitation of the dataset itself. This also means that the paper has to be sufficiently long to provide the reader with all the needed demonstration that can increase confidence in the RHARM data usage. Given this, we have tried to remove any redundancies or unneeded text in the new version of the manuscript in order to optimize the reading and the clarity of the contents.
Messages to takeaway are highlighted in each section, with several caveats and clarifications. To further clarify the results presented in the manuscript and summarize in a clear and efficient way advantages and drawback linked to the use of the RHARM datasets, in the new version of the manuscript, whenever valuable, a sentence has been added at the end of each section (from the section 3 on) with the main messages to take away. To avoid redundancies, a few paragraphs have been moved to these new closure paragraphs. We hope that by also addressing the reviewers' specific comments, the text of the manuscript will be clearer than the previous version.

Specific comments:

1. L21-25, do you mean the variables except relative humidity were harmonized on 16 standard pressure levels? Please clarify it.

The lines 21-26 have been modified as follows: "*The RHARM method has been applied to the daily (0000 and 1200 UTC) radiosonde data holdings (from 1000 to 10 hPa) from 1978 to present available in the Integrated Global Radiosonde Archive (IGRA). Relative humidity bias adjustment and data provision have been limited to 250 hPa owing to pervasive issues on sensors' performance in the upper troposphere and lower stratosphere. The applied adjustments are interpolated to all reported significant levels to retain information content provided within each individual ascent profile.".*

1. L27-34, it is not clear that how to post-process, adjust and homogenize in each step?

The abstract has been simplified indicating the time series obtained with RHARM as "homogenized time series" and using the expression "bias adjustment" to indicate the removal of the non-climatic signals.

L34, what is systematic effects? What's different from 'systematic biases or errors'?

The concept of "systematic effects" follows the definition reported in the Basic Concepts of the Guide for Uncertainty Measurements (GUM). "If a systematic error arises from a recognized effect of an influence quantity on a measurement result, hereafter termed a systematic effect, the effect can be quantified and, if it is significant in size relative to the required accuracy of the measurement, a correction or correction factor can be applied to compensate for the effect. It is assumed that, after correction, the expectation or expected value of the error arising from a systematic effect is zero."

To further clarify the GUM has been quoted the first time the expression "systematic effects" has been used (line 115).

1. L51, 'these climatic time series' is changed to 'these biased time series'?

The sentence at lines 53-54 has been removed because considered not redundant.

1. L81, please cite other literatures at the end of the sentence and move (Hersbach et al., 2020) to the back of the ERA5?

Lines 80-81 have been modified as follows: "A similar situation exists for the most recent ECMWF ERA5 reanalysis (Hersbach et al., 2020) as well as for other meteorological reanalyses (e.g. Kobayashi et al., 2015; Gelaro et al., 2017).".

1. L103, how many stations? instead of 'several sites'

L103 has been rephrased and extended as follows: *"….GRUAN is providing long-term, high-quality radiosounding data at 30 sites (12 sites are certified) around the world with characterized uncertainties, ensuring the traceability to SI units or accepted standards, providing extensive metadata and comprehensive documentation of measurements and algorithms.*"

1. L122, 'since 2004' is changed to 'from 2004'?

Fixed.

1. L132 and 179, pls revise 'a subset of 650 radiosounding stations'?

Fixed.

1. This study missed lots of stations in USA, Russia and China in Fig. 1?

Please see the comments provided above to the referee's general concerns which also clarifies this point.

1. Fig. 1 shows sonde types probably from different decades at different stations, which may mislead the readers. If there are several sonde types for one station, which type did you use?

Table 1, which precedes Figure 1, provides the list of the radiosonde types reported in Figure 1. As mentioned above, among the RHARM post-processed radiosounding profiles, 28 stations only used multiple radiosonde types in the period since 2004 to present. These stations have been reported in Figure 1 showing the most abundant radiosonde type used at the station in the same period.

1. L251, 'reasonably complete?' please see #8 12. L307, how did you handle surface data?

Assuming that with "surface data" the referee is indicating the first point of each radiosounding profile, these are corrected using the same approach followed for the entire profiles. Radiation correction is negligible while the bias due to the factory calibration and ground-check procedure

are removed using the discrepancies estimated between IGRA and GRUAN at the GRUAN sites selected for the intercomparison. The bias is estimated at mandatory levels and then interpolated at any higher pressure reported at the station, including the ground level.