NOAA Climate and Global Change Program
Semi-annual Progress Report for
June 1, 2011 through November 30, 2011

Generating consistent radiance SDRs and deep-layer atmospheric temperature TCDRs from the MSU/AMSU/SSU temperature-sounding channels

NOAA Award No. NA09NES4400015

Science Team Lead: *Cheng-Zhi Zou*
Center for Satellite Applications and Research (StAR)
NOAA/NESDIS
Camp Springs, MD 20746

Science Team Members: *Sid Boukabara, NESDIS/StAR*
Lidia Cucurull, NESDIS/StAR
Dick Dee, ECMWF
Qiang Fu, University of Washington
Thomas Kleespies, NESDIS/StAR
Carl Mears, Remote Sensing Systems
John Woollen, NWS/NCEP

Collaborators: *Ralph Ferraro/NESDIS/StAR*
Ben Ho/UCAR

Introduction

This project is a complex, multi-institutional effort aimed at generating climate quality temperature datasets from satellite-borne atmospheric sounders. This report only covers the part of this work assigned to **Remote Sensing Systems**, which is to perform comparisons of various satellite products with each other, and with other sources of atmospheric temperature measurements. The purpose of this part of the work is to use such comparisons to help determine the reasons for, and significance of, any discrepancies between satellite datasets. This report covers 6 months of work, from June 1, 2011 to November 30, 2011.

One of the most important goals of this part of the work is to intercompare results from the satellite-based datasets with those from other data sources. In this report, we report results we have obtained by comparing MSU/AMSU temperatures with output from recent reanalysis efforts. These results, in contrast to those presented for the previous six months, focus on the spatial patterns in trends over the satellite period (1979-present). We have also extended our analysis to include results from the JRA-25 reanalysis, and results from the upper troposphere/tropopause region.

We have also begun an analysis that compares RSS and STAR results for individual satellites. By analyzing individual satellite results, we can develop a better understanding of the reasons for the differences between RSS and STAR results.

Our report is divided into 3 sections.

1. Description of the reanalysis data studied.
2. Comparisons of spatial patterns in long-term trends.
3. Comparisons of global time series derived from single satellite data.

In addition, we have prepared a manuscript for publication from the short-time trend analysis we discussed in our last report, in collaboration with Peter Thorne from the Cooperative Institute for Climate and Satellites (North Carolina). We anticipate submitting this manuscript within 1 month.

Description of reanalysis datasets.

We have now completed our analysis using 3 reanalysis datasets, the European Centre for Medium Range Forecasting (ECMWF) Reanalysis-Interim (ERA-Interim) (Dee et al. 2011), the Modern ERA Retrospective Analysis for Research and Applications (MERRA) (Rienecker et al. 2011), and the Japanese 25-year Reanalysis Project (JRA-25) (Onagi et al., 2007). Output from all three reanalyses now begin in January 1979 and continues through the present. All three reanalyses ingest satellite data in addition to *in situ* data. This satellite data includes data from the microwave sounders studied here, so there is some question whether the reanalyses and the MSU/AMSU datasets are truly independent datasets – they are probably not fully independent, but exhibit different long-term behavior than any of the satellite datasets. Despite these questions, we have concluded that such comparisons are instructive. Because modern analysis systems use some type of 4-d var assimilation system, i.e. they assimilate data using the actual time of the measurement, the effects of changing measurement times in the satellite systems (the so-called diurnal drift (Mears and Wentz 2005)) are thought to be automatically addressed. And, since the reanalysis is thought to be a synthesis of all available data using a fixed analysis system, in the best case scenario the reanalysis results can be a very good representation of the climate system. Unfortunately, the reanalyses are not free from problems that arise from changes in the observing system, and in particular when new or different types of satellite data are suddenly added, and the new measurements are biased relative to the model expectations. In the future, we also plan to evaluate data from the NCEP Climate Forecast System Reanalysis (CSFR). Processing of these data into synthetic MSU/AMSU brightness temperatures is still underway.

Progress in comparison with modern reanalysis products.

Since the previous progress report, the temporal coverage of ERA-Interim has been extended back to the beginning of the satellite era (1979), and we have constructed MSU/AMSU equivalent brightness temperatures from both the ERA-Interim and from the JRA-25 reanalysis using our MSU/AMSU specific fast radiative transfer model. In Figure 1, we plot global time series for middle tropospheric temperature (TMT) anomalies from STAR, RSS, and UAH satellite datasets, and from the MERRA, ERA-Interim, and JRA-25 reanalyses. It is interesting to note that the spread in long term trends is slightly larger for the reanalysis output (0.063 to 0.160 K/decade, a spread of 0.97 K/decade) than it is for the satellite-derived trends (0.56 to 0.141 K/decade, a spread of 0.85 K/decade), despite the diverse data sets that are ingested by the reanalysis systems which should constrain the results.

In Fig 2, we plot the difference between each TMT time series from the mean of all time

series. Several features become apparent. First, the JRA-25 differences show a number of discontinuous jumps, with the most prominent jumps occurring in late 1998 and again in 2005. These may be due to the introduction of new kinds or sources of data into the reanalysis system at these times. From about 2000 onwards, the MERRA reanalysis trends upward relative to the mean, which the STAR dataset trends downward. This is despite the fact that MERRA ingests the STAR L1C brightness temperatures and serves as a counter argument to those that claim that the large trends seen in MERRA are due to its use of STAR brightness temperatures.

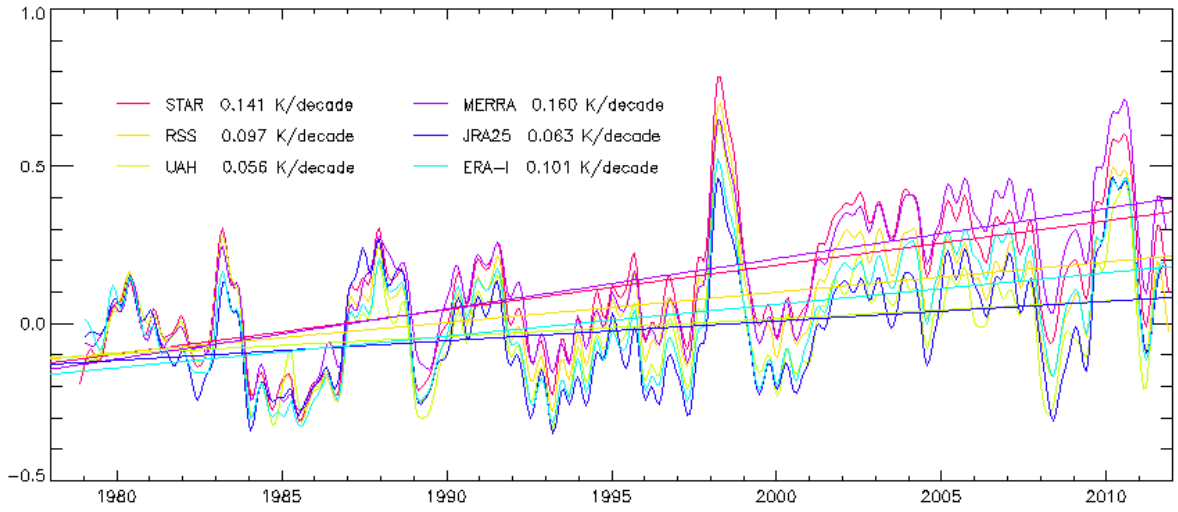


Figure 1. Globally averaged monthly time series of middle tropospheric temperature (TMT) for the three satellite datasets (STAR, RSS, and UAH), and 3 modern reanalysis products (MERRA, JRA-25, and ERA-Interim). The data have been smoothed to remove variability on time scales shorter than 6 months.

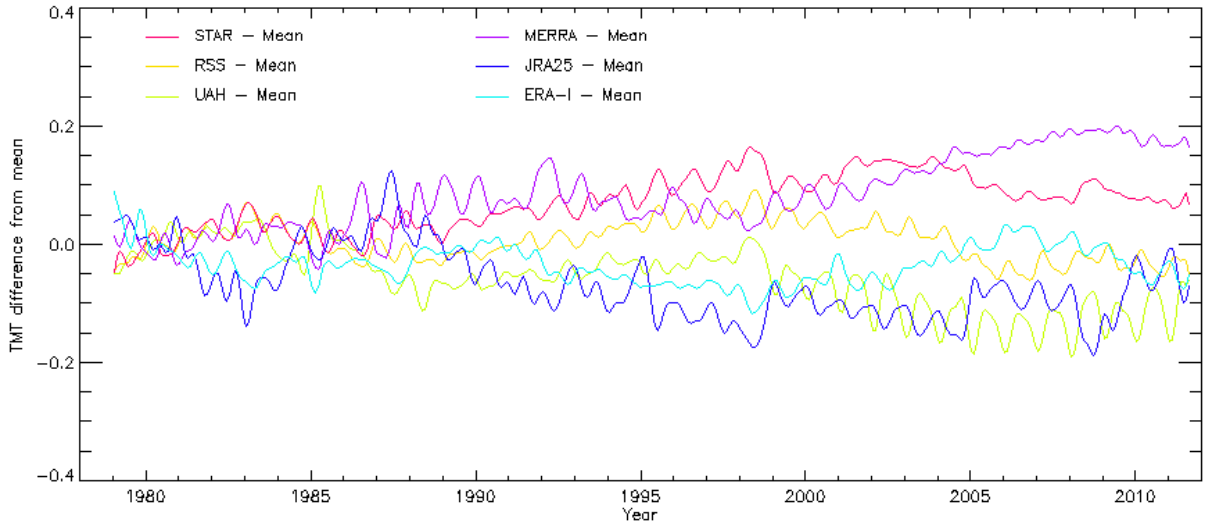


Figure 2. Globally averaged monthly time series of the difference of each time series plotted in Figure 1 from the mean of all six time series. Again, the data have been smoothed to remove variability on time scales shorter than 6 months.

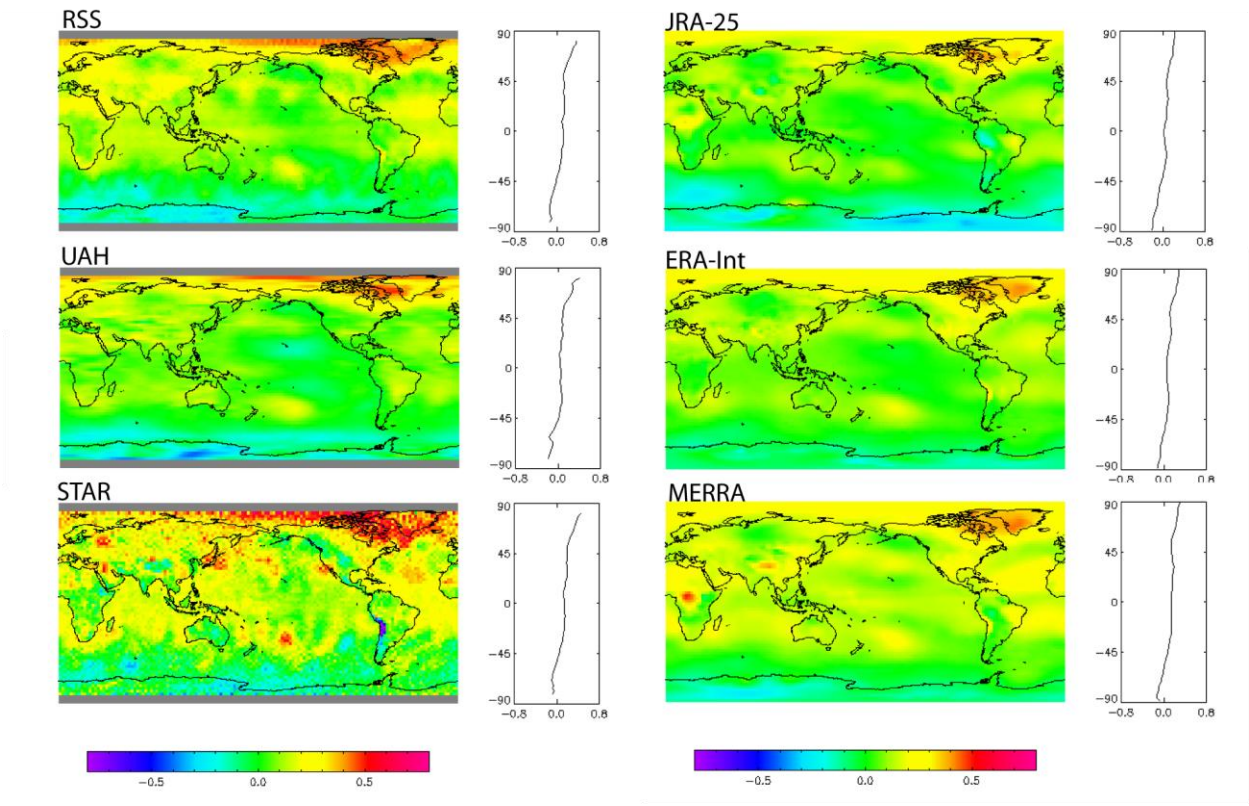


Figure 3. TMT trend maps (1979-2010) for the three satellite datasets, and 3 reanalyses.

In Figure 3, we plot maps for trends in middle tropospheric temperature (TMT) for the six datasets. All datasets exhibit the same basic features, with more warming in the northern latitudes, and less warming or cooling near the south pole. There are a few differences that stand out. First, the STAR dataset shows signficantly more spatial variability in the trends than either of the two other satellite datasets. Second, none of the reanalyses show as much warming in the arctic as is seen in the satellite datasets. Third, both MERRA and STAR show more warming in the tropics than is apparent in any of the other datasets. STAR also appears to show a land/ocean trend contrast, particularly surrounding Africa that is not as large in the other datasets.

Single Satellite Comparisons.

Comparisons of results for individual satellites can help isolate the causes of any differences observed for the combined datasets. To date, we have completed a preliminary study of the differences between RSS and STAR data for the AMSU satellites that both groups use. STAR does not yet include data from the AQUA satellite. Unfortunately, individual satellite data are not available from the UAH group. We begin by comparing data for AMSU channel 5, or TMT.

In Figure 4, we show the difference (STAR-RSS) for global (80S to 80N) monthly averages for NOAA-15, NOAA-18, and METOP-A. The data for NOAA-15 show a troubling ~0.1 K discontinuity at the beginning of 2001, with the STAR data warming relative to RSS at this time. Evaluation of the average brightness temperatures for two years on either side of the discontinuity shows that this shift is due to a change in overall brightness in one of the datasets, i.e. the magnitude of the shift is independent

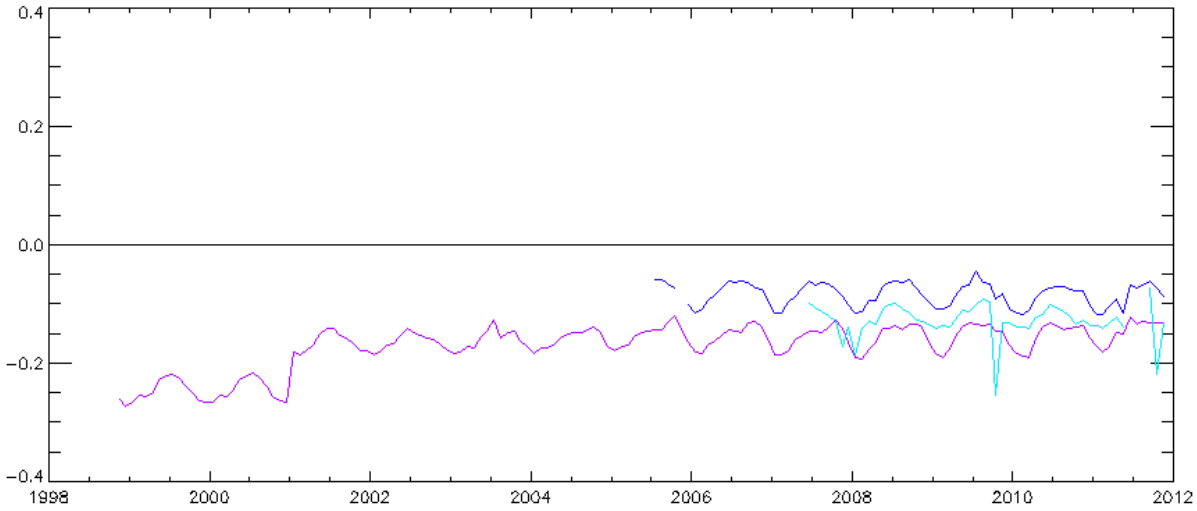


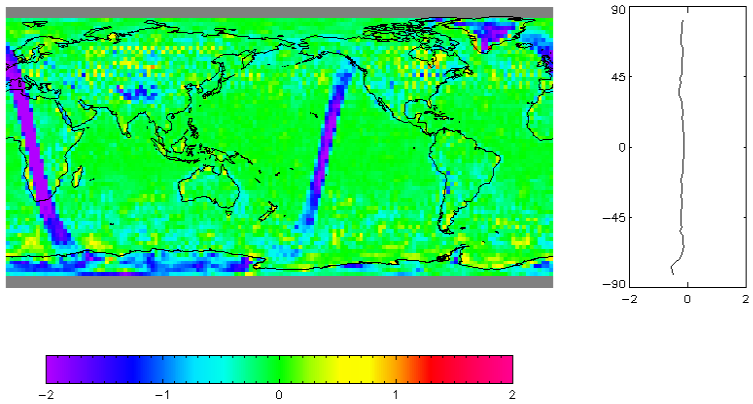
Figure 4. Global mean (80S to 80N) difference (STAR-RSS) time series for AMSU channel 5. All time series show a small offset and seasonal cycle, probably caused by the different calibration procedures used by the two groups. For this channel, the difference time series for NOAA-15 (shown in purple) shows a discontinuous step in January 2001 of about 0.1 K. The difference time series for METOP-A (shown in light blue) shows several months with large differences (October 2009, September and October 2011. These are likely to be due to missing data in one of the datsets (see Figure 5b).

of location or season. This suggests that it is due to a change in calibration, or an error in the application of the calibration algorithm in one of the datasets. More research is needed to resolve the cause of this shift.

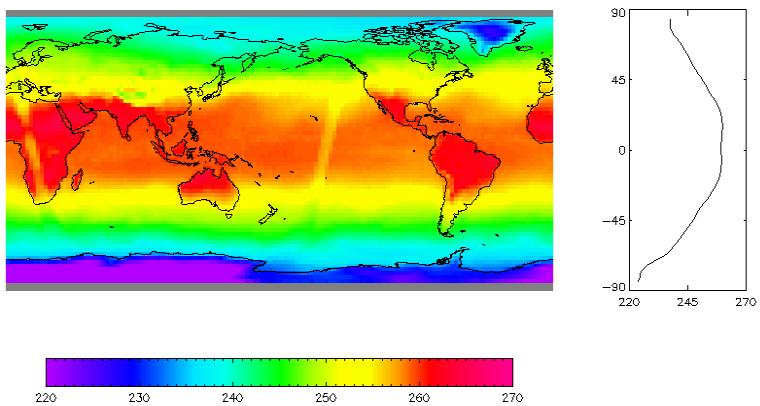
There are also large anomalies in the difference time series for METOP-A in October 2009, and September and October of 2011. Further analysis suggests that this is due to missing data for these months in one of the datasets. As an example, in Figure 5a we show a map of the brightness temperature differences for October 2011, along with maps of brightness temperatures for STAR (Figure 5b) and RSS (Figure 5c). In Figure 5a, we can see regions of large difference that roughly mimic the shape of data from a single orbit. This suggests that a number of orbits may be missing or misanalyzed in one of the datasets. In Figure 5b, we can see a region of anomalously low brightness temperature in the STAR data that is not present in the RSS data (Figure 5c), indicating that the problem for this month resides in the STAR data. A similar set of maps from October 2009 is less conclusive.

Fortunately, the other 2 AMSU channels used for the deep-layer products show fewer problems. In Figure 6, we plot globally averaged difference time series for AMSU channel 7 (TTS). For the satellites plotted (NOAA-15 in purple and NOAA-18 in blue) there are no large discrepancies, though there appears to be a downward trend in the differences starting in about 2008. As a practical matter for the combined time series, this is partly countered by the small upward trend in the NOAA-18 differences.

a)



b)



c)

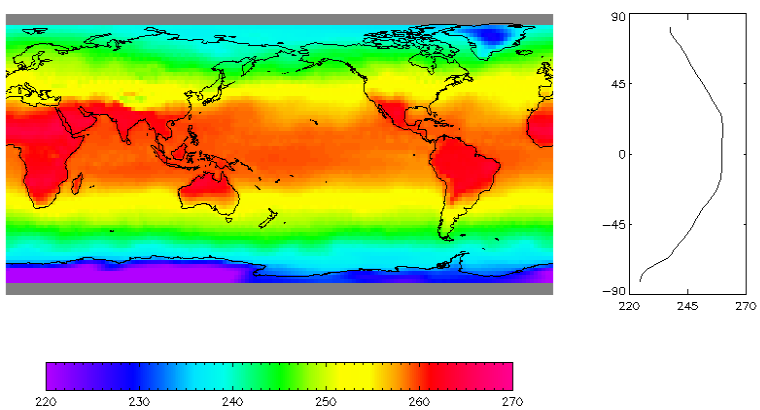


Figure 5. a) Map of AMSU 5 brightness temperature differences (STAR minus RSS) for October 2011, METOP-A. b). Map of STAR AMSU 5 brightness temperatures, October 2011. c) Map of RSS AMSU 5 brightness temperatures, October 2011. Note that the region of large difference is similar to the region for data derived from one orbit.

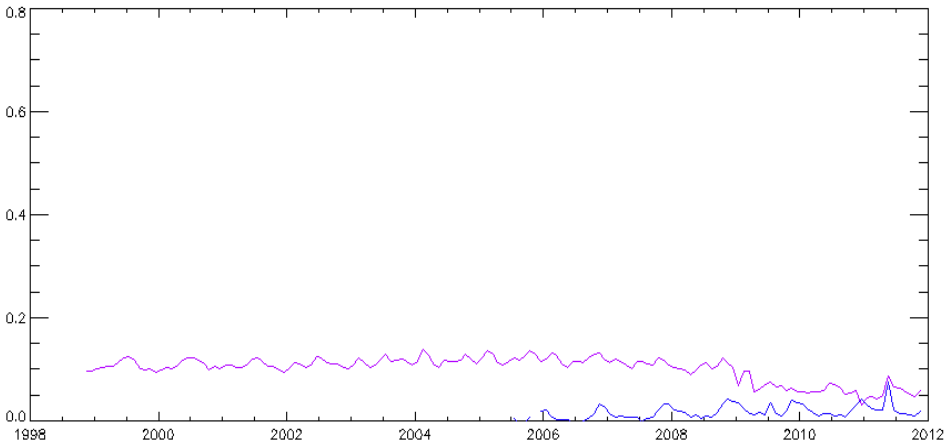


Figure 6. Time series of globally averaged brightness temperature differences (STAR minus RSS) for AMSU channel 7.

In Figure 7, we plot the globally averaged time series differences for AMSU9 (TLS). For TLS, there are no obvious trends in the differences, but the intra-dataset differences are quite different between satellites. This is likely to be because the final intersatellite offsets have not been applied to the single satellite STAR data that are available via the web.

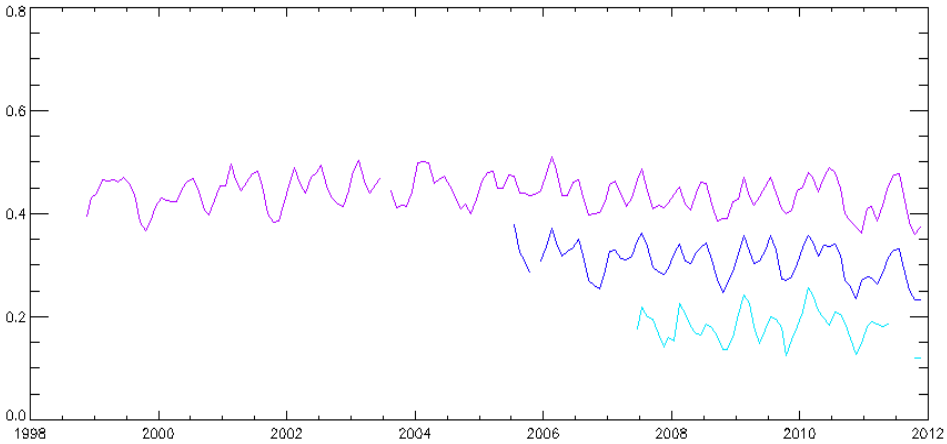


Fig. 7. Time series of globally averaged brightness temperature differences (STAR minus RSS) for AMSU channel 9.

Planned activites for the next 6 months

- Finish manuscript preparations describing short-term trend results
- Complete comparisons with the CSFR reanalysis.
- Perform further work on comparisons with radio occultation data (GPS-RO) and TLS datasets**.**
- Update comparisons with changes in total water vapor column using SSM/I and AMSR-E data.

References

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q.J.R. Meteorol. Soc.*, **137**, 553-597.

Mears, C. A., Benjamin D. Santer and Dotraix, C.Frank J. Wentz, 2011a: Calculating synthetic microwave sounder brightness temperatures from discrete-level data. *Under review*.

Mears, C. A., and F. J. Wentz, 2005: The effect of drifting measurement time on satellite-derived lower tropospheric temperature. *Science*, **309**, 1548-1551.

Mears, C. A., F. J. Wentz, P. W. Thorne, and D. J. Bernie, 2011b: Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique. *Journal of Geophysical Research*, **116**, D08112.

Onogi, K., J. Tsutsui, H. Koide, M. Sakamoto, S. Kobayashi, H. Hatsushika, T. Matsumoto, N. Yamazaki, H. Kamahori, K. Takahashi, S. Kadokura, K. Wada, K. Kato, R. Oyama, T. Ose, N. Mannoji and R. Taira (2007) : The JRA-25 Reanalysis. J. Meteor. Soc. Japan, 85, 369-432.

Rienecker, R. H., M. J. Suares, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. G. Bosilovich, S. D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, and A. Conaty, A. da Silva, 2011: MERRA - NASA's Modern-Era Retrospective Analysis for Research and Applications. *Journal of Climate*.

NOAA Climate and Global Change Program
Semi-annual Progress Report for
December 1, 2011 through May 31, 2012

Generating consistent radiance SDRs and deep-layer atmospheric temperature TCDRs from the MSU/AMSU/SSU temperature-sounding channels

NOAA Award No. NA09NES4400015

Science Team Lead: *Cheng-Zhi Zou*
Center for Satellite Applications and Research (StAR)
NOAA/NESDIS
Camp Springs, MD 20746

Science Team Members: *Sid Boukabara, NESDIS/StAR*
Lidia Cucurull, NESDIS/StAR
Dick Dee, ECMWF
Qiang Fu, University of Washington
Thomas Kleespies, NESDIS/StAR
Carl Mears, Remote Sensing Systems
John Woollen, NWS/NCEP

Collaborators: *Ralph Ferraro/NESDIS/StAR*
Ben Ho/UCAR

Introduction

This project is a complex, multi-institutional effort aimed at generating climate quality temperature datasets from satellite-borne atmospheric sounders. This report only covers the part of this work assigned to **Remote Sensing Systems**, which is to perform comparisons of various satellite products with each other, and with other sources of atmospheric temperature measurements. The purpose of this part of the work is to use such comparisons to help determine the reasons for, and significance of, any discrepancies between satellite datasets. This report covers 6 months of work, from December 1, 2011 to May 31, 2012. During this time period, we have focused on the submission of a paper describing an analysis of short-term trends in RSS, UAH, and STAR satellite data, and data from a variety of homogenized radiosonde datasets. This manuscript was submitted for publication in the Journal of Geophysical Research, Atmospheres and the first round of reviews have been received. The reviewers requested a substantial amount of additional work, so of which is described in this report. A copy of the paper as submitted is appended to the end of this report. In addition, we have compared TLT results from the RSS and UAH groups with a new version of the total column water vapor results from microwave imaging instruments.

1. Short term trend analysis using the Temperature Total Troposphere (T_{TT}) developed by Fu et al.
2. Effects of Sampling and dataset version on short-tem trend results
3. Short term trend analysis in the extratropics

4. Comparison of TLT trends from RSS and UAH with trends in total column water vapor.

Short-Term Trend Analysis using T_{TT}

Our analysis of short term trends in middle tropospheric temperature (T_{MT}), which we reported in a previous progress report, was criticized by reviewers because the results could be influenced by cooling in the lower stratosphere. This influence is easiest to see in plots of short-term (5 –year or 10-year) trend differences between the lower tropospheric temperature, and temperatures higher in the troposphere. This type of analysis was first performed by Randall and Herman (2007) – our work is largely a response and update of this earlier work. When T_{MT} is used for the mid/upper tropospheric temperature, the can be significant influence in the differences due to short or long term trends in the stratosphere. This is because about 10% of the weight in T_{MT} comes from emission in the stratosphere. To address this concern, we repeated the analysis using the "Temperature Total Troposphere" or T_{TT} product proposed by Fu and Johanson, (2005). By forming a linear combination of T_{MT} and lower stratospheric (T_{LS})

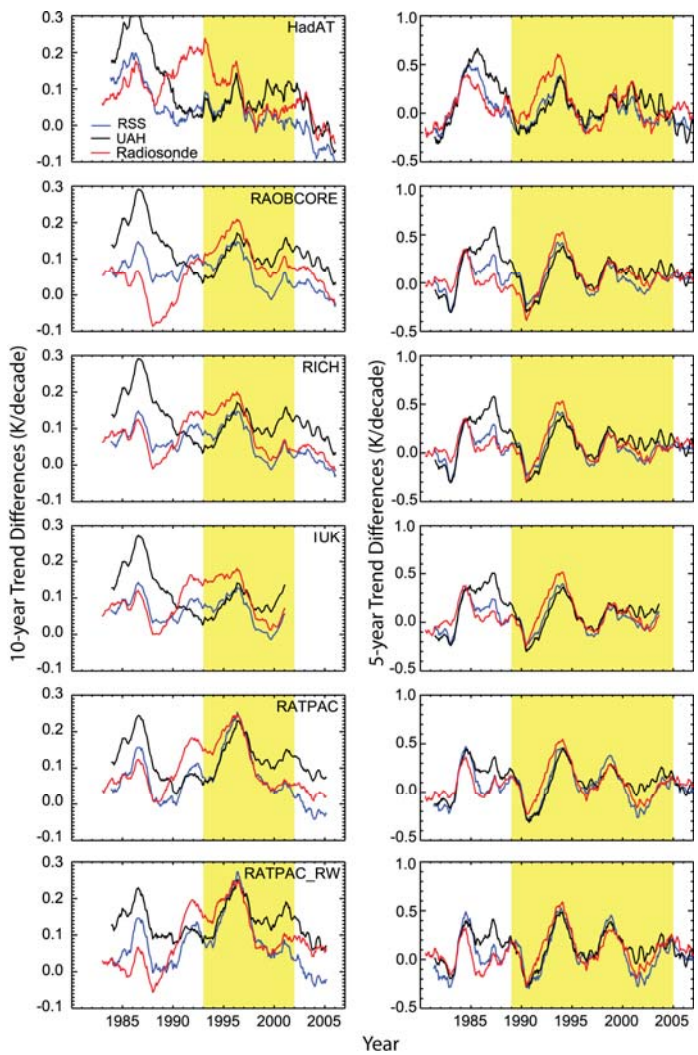


Fig. 1. Differences (TLT minus TMT) in short term trends in atmospheric temperature between adjacent layers for each radiosonde dataset. In each case, the satellite data is sampled at the radiosonde locations, and the differences in spatial sampling lead to the differences between the satellite results for the different panels. The temporal range plotted in RH2008 is shaded in yellow.

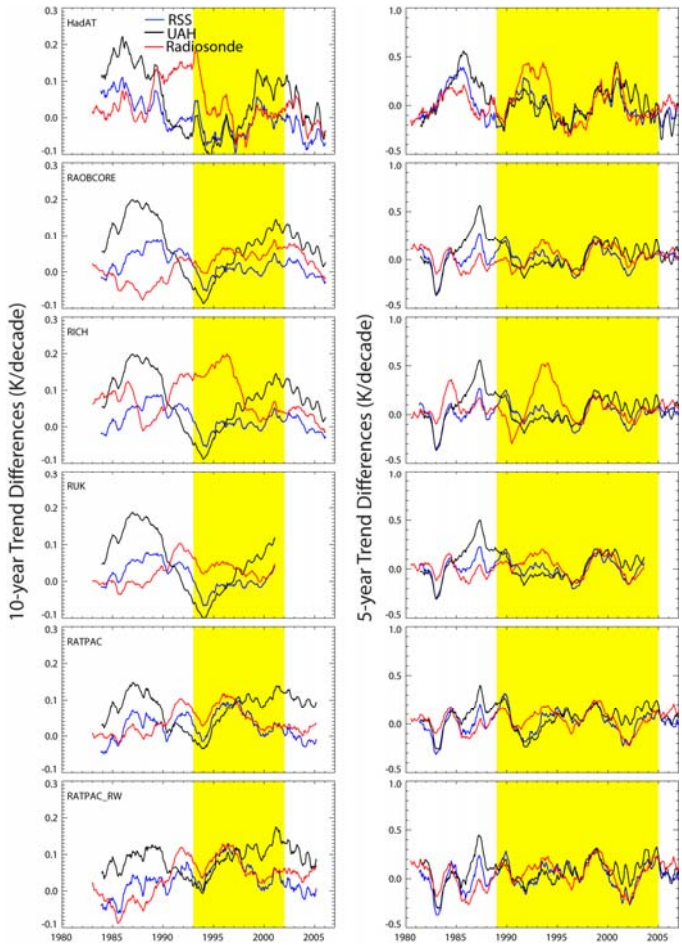


*Fig. 2. Same as Figure 1 in the main text, except calculated using the TT (Total Troposphere) product defined in Fu and Johansan (2005) to be 1.1*TMT -0.1*TLS. This weighted combination of channels reduces the contribution of the stratosphere relative to TMT. The use of TT changes the time series considerably, and in general the variability of the time series is reduced, probably due to the reduction of the influence of volcano-induced stratospheric warming. The overall conclusions drawn from Fig 10 are unaltered, with some of the largest discrepancies occurring outside the range considered by RH2008, and RSS being in better agreement with all radiosonde datasets.*

measurments (see Eq. 1), they found that most of the stratospheric influence could be removed.

$$T_{TT} = 1.1 \cdot T_{MT} - 0.1 \cdot T_{LS} \tag{1}$$

In Figs. 1 and 2, we show plots of the difference between short-term trends in T_{LT} and T_{MT} (Fig. 1) or T_{TT} (Fig. 2). In all cases, the T_{TT} results were calculated using T_{MT} and T_{LS} products from the same research group.

In Fig. 1, there is a large fluctuation in the 5-year trend differences in the 1990-1995 period. This is likely to be associated with the eruption of Mt. Pinatubo in June 1991. The eruption injected the most aerosols into the stratosphere since the eruption of Krakatoa in 1883. The effect of these aerosols was to heat the stratosphere, via direct absorption of solar radiation, and cool the troposphere due to the related reduction of solar radiation at Earth's surface. These effects decayed over the next few years as the aerosols in the stratosphere were depleted. The

result is the any short term trend that ended after June 1991 was smaller than normal in the troposphere, and larger than normal in the stratosphere. This causes the low values for the TLT minus TMT trends in the early 1990's in Figure 1. A similar argument applies to the larger values of the trend difference in the late 1990's. When T_{TT} is used instead (Figure 2), the effects of stratospheric warming are reduced, leading to a reduction in the feature in all datasets. In fact, in almost all cases, the use of T_{TT} reduces the variability of the short term trend difference, indicating that T_{TT} and T_{LT} are more closely coupled than T_{MT} and $T_{LT,}$ and that there are many instances where changes in the troposphere and stratosphere are anti-correlated.

The Effects of Dataset version and Sampling Method on Short Term Trends

We also have performed further work on our short-term trend analysis to determine whether updated versions of the UAH and RSS dataset and method for generating comparison datasets have a significant impact on the results. In Fig. 3, we compare the mean absolute difference between 5-year trends in the satellite data, and 5-year trends in homogenized radiosonde data. In the to row (panels a and b), the satellite data is sampled at the locations of valid radiosonde data. This leads to fairly good agreement between radiosonde and satellite data. In panel a, we use the current versions of the UAH and RSS data, while in panel b, we use two previous versions of the RSS and UAH data. The differences between the results presented in the two panels is not large, though the later version of the RSS data tends to agree better with radiosondes than the previous version. This difference is most apparent for the HadAT dataset. In panel c, we compare the

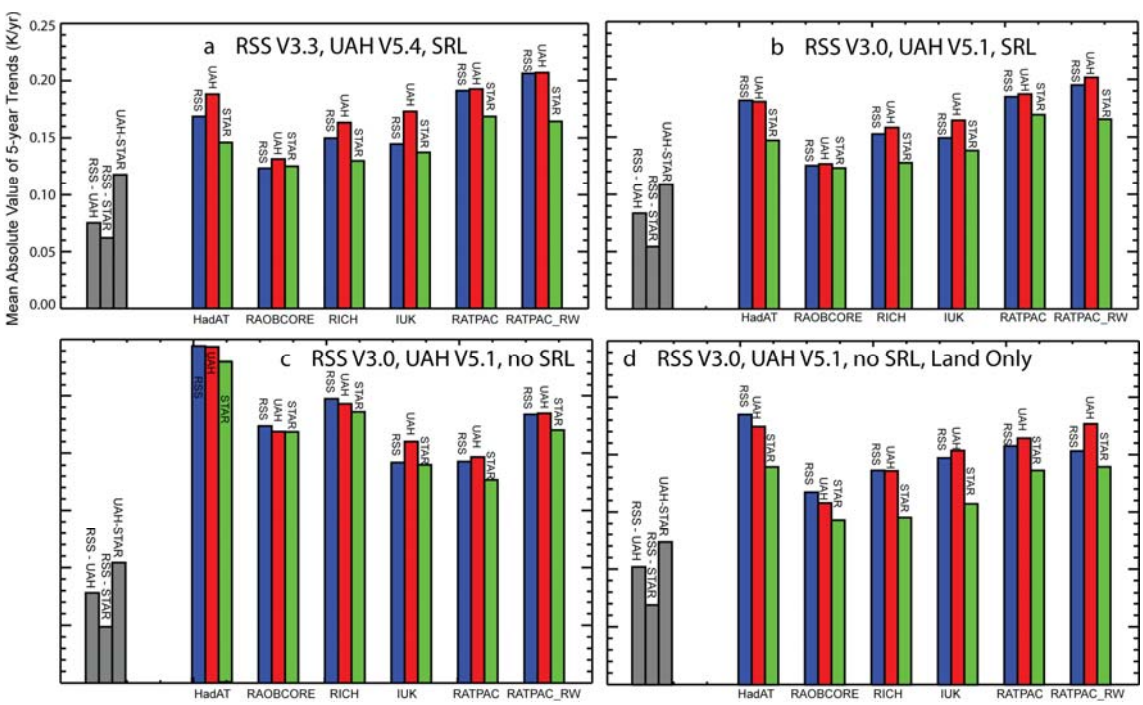


Fig. 3. Bar graphs of mean absolute difference between 5-year trends of satellite and homogenized radiosonde data. a) current satellite data was sampled at radiosonde locations. b) Same as panel a, except calculated using previous versions of the RSS (V3.0) and UAH (V5.1) datasets. The STAR dataset is the same. For this plot, the period of analysis is 1979-2007, so any changes in the STAR results are due to the change in analysis period. c) Same as panel b, except that global, area-weighted averaging was used to construct the global averages. d). Same as panel d, except than land-only area weighted averages were used for the MSU dataset.

radiosonde data with global, area weighted versions of the satellite data. Here the mean absolute differences are much larger, because of the spatial mismatch between the radiosonde and satellite sampling. The situation can be improved slightly by using a land-only subset of the satellite observations as was done by Randall and Herman. This improvement is probably because the land-only sampling more closely matches the hemispheric distribution of the radiosonde data for the HadAT, RAOBCORE, RICH, and IUK datasets. For the full RATPAC dataset, which is relatively uniformly spatially distributed by design, the agreement degrades when the land-only satellite data is used. We conclude that for these kinds of comparisons, the method used to account for spatial sampling is more important that the small changes made between subsequent versions of the satellite datasets.

Short Term Trend Analysis in the Extratropics.

Comments from reviewers of our short-term trend paper caused us to extend our analysis of short term trends, originally performed for global and tropical averages, to extratropical averages in the Southern (80S to 25S) and Northern hemispheres (25N to 80N). We found that the

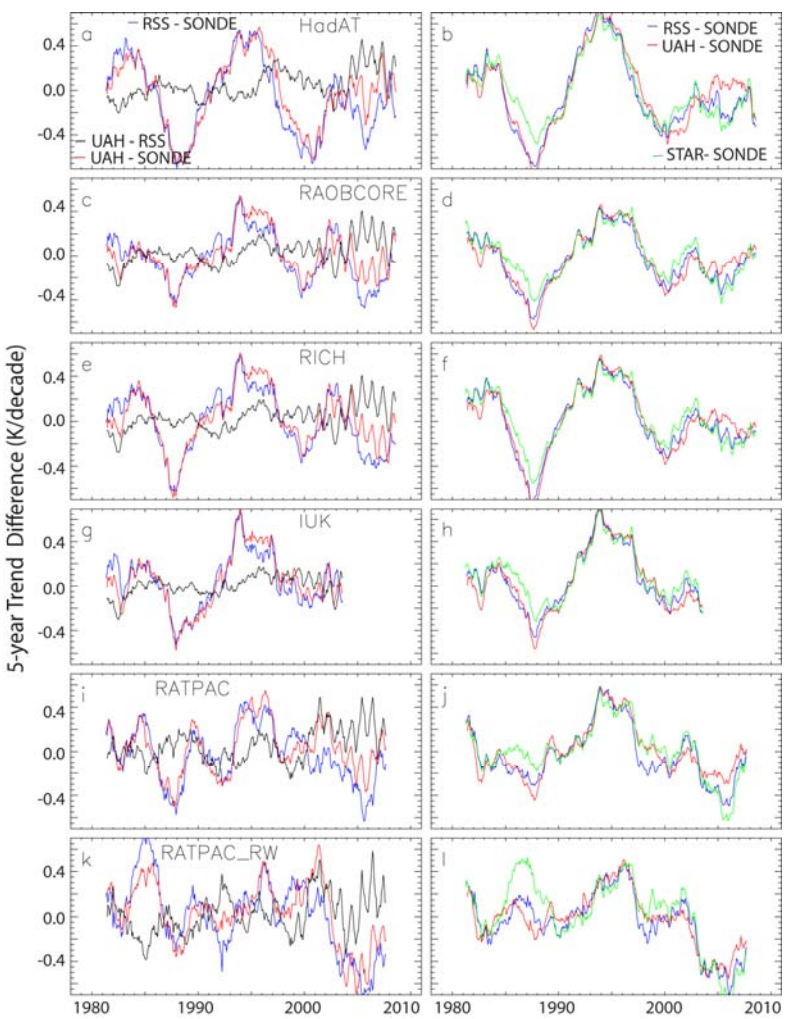


Fig. 4. Short term trend differences for the Southern Extratropics (82.5 S to 20S for TMT, and 70S to 20S for TLT). This spatial subset shows features that are quite different from the other regions, especially for the HadAT, RAOBCORE and RICH datasets. These differences may arise because of the small number of radiosonde stations in the southern hemisphere.

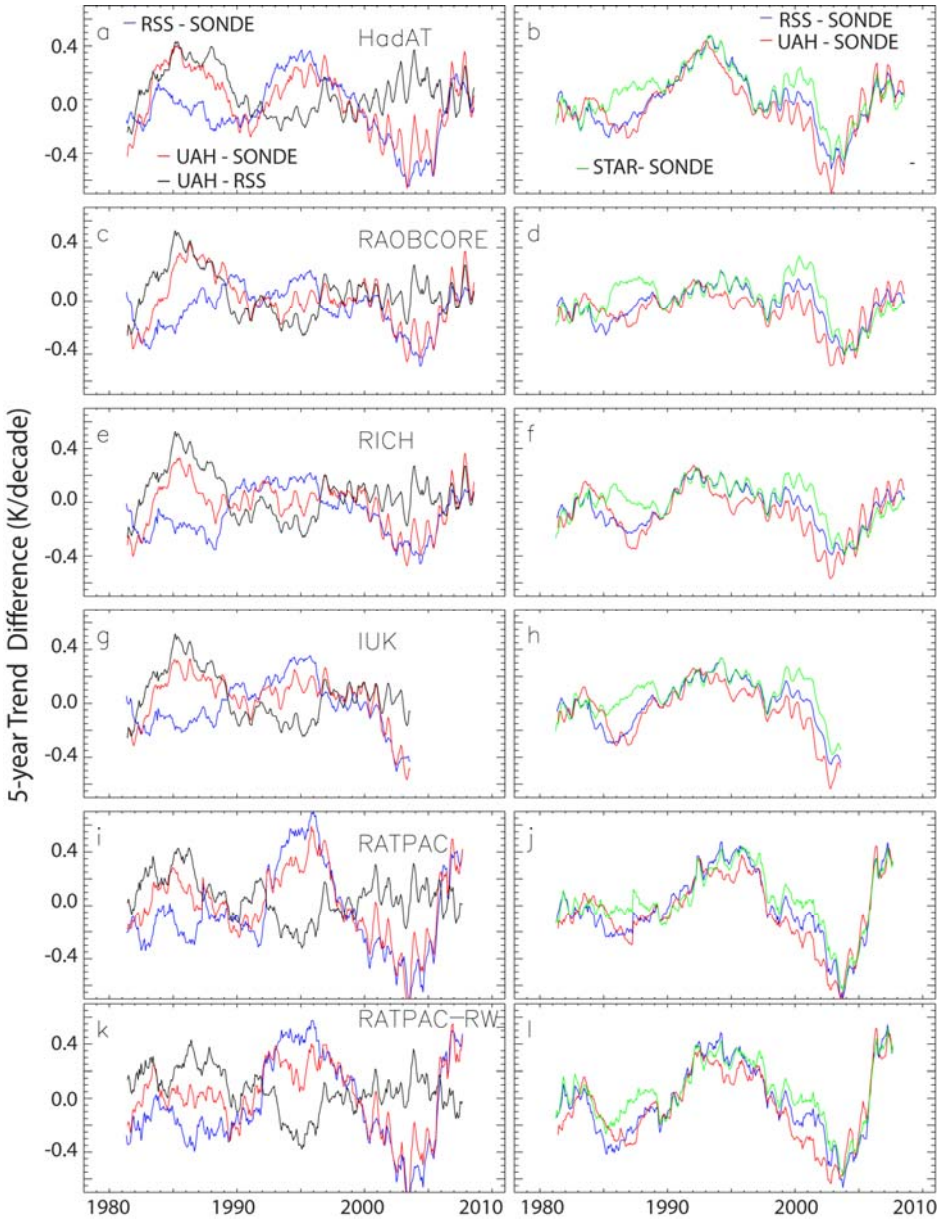


Fig. 5. Same as Fig. 4 except for the Northern Extratropics (20N – 82.5N). The general features, such as the locations of peaks, are similar.

agreement between radiosondes and satellites is substantially worse in the southern hemisphere. This is not surprising, since there are many fewer radiosonde stations in the southern hemisphere in all datasets except for RATPAC, which, perhaps not coincidentally, shows the best agreement with satellites of all the radiosonde datasets. In the Northern hemisphere, the agreement between radiosondes and satellites is also not very good. This is more surprising, since the conventional wisdom is that many of the NH stations provide high quality data. Both regional subset confirm our earlier conclusions that satellite data is more like satellite data than it is like radiosonde data, and the radiosonde datasets are more like each other than they are like satellites. In our submitted paper, we argue that this limits the utility of radiosonde/satellite comparisons for unambiguous satellite dataset validation.

Comparison of Satellite Derived Temperature Datasets with Version 7 SSM/I Total Column Water Vapor.

In Mears et al (2007), we established that it is useful to compare changes in temperature in the tropics with changes in oceanic total column water vapor. Model results, as well as basic physics suggest that temperature and water vapor content should be tightly coupled in the tropics, where the structure of the atmosphere is governed by convection. We recently released Version 7 of SSM/I, SSMIS, AMSRE, and WindSat microwave imager datasets. These datasets consist of retrievals of total column water vapor, surface wind speed, total cloud water, and rain rate. These parameters are only available over the world's oceans. The signals involved in retrieving the total column water vapor are large, giving us substantial confidence in the accuracy of the water vapor product. Most temperature datasets closely track the total column water vapor in the tropics on short time scales. On longer scales, the trend comparisons may be used for diagnosis of problems in one or more datasets. In Fig. 6 we show a time series of the ratio of trends in water vapor, expressed in percent per decade, to trends in lower tropospheric temperature as measured by the MSU/AMSU series of instruments. The plots are from average values averaged over the tropical oceans (20S to 20N). We plot time series of trends derived from each of the TLT datasets, RSS and UAH. Each point on the time series corresponds to the trend calculated from a data subset that begins in January 1988, and ends at the time denoted by the x-axis. Since the plot begins in 2000, the shortest trend represented is 12 years, long enough for the trend to begin to stabilize to a final value. The trend ratio from RSS if fairly constant around 7.5% vapor change per K, while the UAH derived trend has substantially larger fluctuations, with some trend ratios (e.g for 1988-2002) above 15% percent change in vapor per 1K change in temperature, well in excess of the expectations of Clausius-Clapeyron and model results which suggest the ratio should be around 6 to 7% per K.

In Figure 7, we show a similar plot, except using the TT product for the tropospheric temperature. The results are similar for UAH and RSS, with the UAH trend ratio showing more

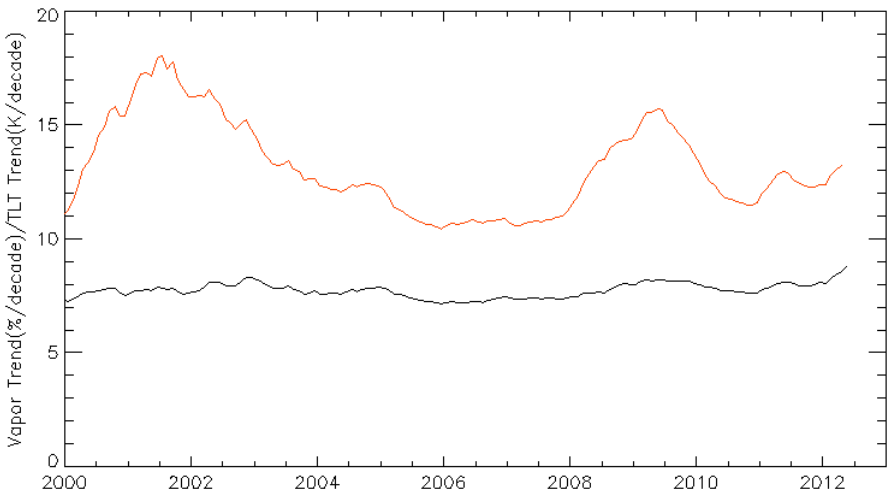


Fig. 6. Ratios of trends in water vapor to trends in TLT, with the trend for each point on the curve starting in Jan. 1988, and ending at the time denoted by the x-axis. The RSS trend ratio (black) shows less variability than the UAH trend ratio (red).

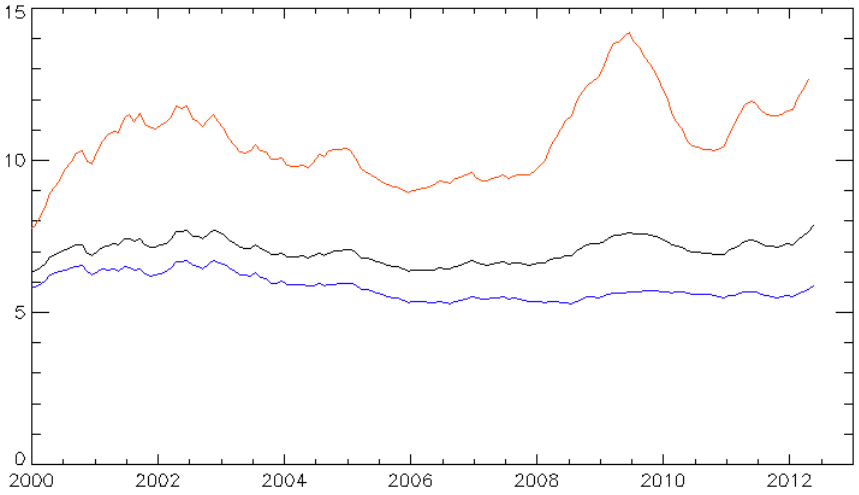


Fig. 7. Ratios of trends in water vapor to trends in TT, with the trend for each point on the curve starting in Jan. 1988, and ending at the time denoted by the x-axis. The trend ratios calculated using RSS data (black) and STAR data (blue) show less variability than the UAH trend ratio (red).

variability. For TTT, it is also possible to include the ratio time series calculated using the STAR satellite temperature dataset, which also shows low variability. These results suggest that there may be significant errors in the UAH datasets on the intermediate time scale.

References

Fu, Q., and C. M. Johanson (2005), Satellite-derived vertical dependence of tropical tropospheric temperature trends, *Geophys. Res. Lett.*, 32, L10703, doi:10.1029/2004GL022266.

Mears, C. A. , B. D. Santer, F. J. Wentz K. E. Taylor, and M. F. Wehner (2007), Relationship between temperature and precipitable water changes over tropical oceans, *Geophys. Res. Letters*, 34, L24709, doi:10.1029/2007GL031936.

Assessing the value of Microwave Sounding Unit-radiosonde comparisons in ascertaining errors in climate data records of tropospheric temperatures

Carl A. Mears and Frank J. Wentz

Remote Sensing Systems, Santa Rosa, CA, USA

Peter W. Thorne

Cooperative Institute for Climate and Satellites, Asheville, North Carolina State University and NOAA's National Climatic Data Center NC, USA

Submitted to Journal of Geophysical Research
Resubmitted, 6/18/2012

Abstract

Multidecadal-scale changes in atmospheric temperature have been measured by both radiosondes and the satellite-borne microwave sounding unit (MSU). Both measurement systems exhibit substantial time varying biases that need to removed to the extent possible from the raw data before they can be used to assess climate trends. A number of methods have been developed for each measurement system, leading to the creation of several homogenized datasets. In this work, we evaluate the agreement between homogenized MSU and radiosonde datasets on multi-year (predominantly 5-year) time scales and find that MSU datasets are often more similar to each other than to radiosonde datasets and vice versa. Furthermore, on these times scales the differences between MSU datasets are often not larger than published internal uncertainty estimates for the RSS product alone and therefore may not be statistically significant when the internal uncertainty in each dataset is taken into account. Given the data limitations it is concluded that using radiosondes to validate multidecadal-scale trends in MSU data, or vice versa, or trying to use such metrics alone to pick a 'winner' is an ill-conditioned approach and has limited utility without one or more of additional independent measurements, methodological, or physical analysis.

1. Introduction

Multi-decadal changes in global atmospheric temperature have primarily been estimated using measurements from two disparate measurement systems, balloon-borne radiosondes (beginning in the late 1950's) and satellite-borne microwave sounding instruments. The microwave measurements are constructed by merging together measurements from the Microwave Sounding Units (MSUs, late 1978-2005) and the Advanced Microwave Sounding Units (AMSUs, mid 1998 to the present). Hereafter we refer to the merged MSU/AMSU datasets as MSU datasets for brevity. Unfortunately, neither MSU nor radiosonde records have been designed with absolute calibration and traceability. Numerous changes in instrumentation, observing practice, time of observation and various other undesirable measurement aspects pervade these records [Thorne et al., 2011a]. A number of techniques to characterize and remove these problems have been developed and refined, resulting in a number of “homogenized” datasets for each type of data. Ideally, the resulting methodologically distinct datasets would report similar changes in atmospheric temperature during the post-1978 period (the satellite era) when both observing systems were in operation. This would yield confidence that data issues had been adequately understood and removed leading to a good estimate of the true climate system evolution. Unfortunately, this has not been the case, which has led to intense debate about the details of recent changes in the Earth's atmospheric temperatures. As the homogenization strategies have evolved over time, trends from radiosonde and MSU data have come into somewhat better agreement for global-scale averages (although still a substantial error as a percentage of the relatively small trend signal), while more substantial discrepancies remain in the tropics [Lanzante et al., 2006]. However, overall trend agreement can hide interesting differences at shorter timescales and may be a result of largely fortuitous cancellation of

substantial differences on these shorter timescales.

In the evaluation of satellite data, particularly in the early part of any mission, it is customary to "validate" the satellite data by comparing retrieved geophysical parameters with *in situ* measurements related to the variable in question. This is a useful exercise, particularly when the principles underlying the remote sensing techniques are still being tested, and when the accuracy of the *in situ* measurements is expected to exceed the accuracy of the satellite-based measurements. It is tempting to extend this approach to the evaluation of long-term trends in geophysical variables with well-established measurement techniques, such as atmospheric sounding. For example, one could use radiosonde/satellite inter-comparison studies to try to make determinations of satellite dataset quality. In fact a number of studies have been used to suggest that one of the MSU datasets is more accurate than the others, based on a closer agreement with various radiosonde measurements. Randall and Herman [2008] compared MSU measurements with the results from a subset of a single radiosonde dataset and concluded that the University of Alabama, Huntsville (UAH) satellite dataset was more accurate than the RSS data. They focused on trends in the dataset differences over 5-year and 10 year periods, but focused on a limited analysis period. Christy *et al.* [2010] reached a similar conclusion using a similar short-term trend analysis, but analyzed only within the deep tropics, and only one time period (1989-1995). Christy et al., [2007] used tropical radiosonde measurements (both raw soundings and a single homogenized dataset) to argue that the RSS dataset contains a spurious warming trend in the tropics during the early 1990s, with the bulk of the analysis of MSU-radiosonde differences focusing on this period. Conversely, Po-Chedley and Fu [2012] argued for a significant discontinuity in the early portion of the UAH record associated with the short life-time NOAA-9 satellite. All these papers used a limited number of radiosonde datasets, and

focused their attention on a limited time period.

Over the last decade, there have been numerous other studies that have included inter-comparisons between MSU and homogenized radiosonde datasets at both global [Haimberger et al., 2008; Lanzante et al., 2006; Seidel et al., 2004] and regional scales [Thorne et al., 2007, 2011b; Titchner et al., 2009]. A recently completed a study of 32-year trends in both MSU and homogenized radiosonde data [Mears et al., 2011] serves in part to update such studies to the current time. When the internal uncertainty in the Remote Sensing Systems (RSS) dataset is taken into account, the trends in this dataset tend to be consistent with those from homogenized radiosonde datasets for the tropospheric channels considered in this paper. The exception to this finding is in the deep tropics, where datasets from RSS and the Center for Satellite Applications and Research (STAR) tend to show trends that are high compared to most adjusted radiosonde trends.

The rest of this paper is structured as follows. Section 2 outlines the overarching methodological ethos and approach. Section 3 provides details of the data used in our study. In Section 4, we compare time series of 5- and 10- year trends derived from each dataset for various regions, and in Section 5 we investigate the impact of estimates of internal uncertainty on our findings. In Section 6 we revisit the analysis of interlayer differences performed by Randall and Herman [2008], and in Section 7 we conclude with a discussion of our findings.

2. Methodological approach and rationale

The purpose of this paper is to comprehensively investigate the question of whether it is appropriate to use MSU/radiosonde intercomparsions as arbiters of MSU dataset quality. Arguably this can only be done if two over-arching conditions are both adequately met. First,

the radiosonde measurements being considered need to be a sufficiently unbiased representation of the true evolution of the climate system. Second, the results of the comparison need to be valid in the statistical sense, given the inevitable uncertainty in both the MSU datasets and homogenized radiosonde datasets.

Several papers, often authored by the developers of the radiosonde datasets themselves, have addressed the first question and have mostly concluded that substantial decadal-scale errors may remain even in the homogenized data [Lanzante et al., 2003; Randel and Wu, 2006; Titchner et al., 2009]. A number of investigators have specifically caveated that it is probable that significant residual errors remain in the tropics where the network is sparse and most observations are day time only when radiation effects are more important [Randel and Wu, 2006; Sherwood et al., 2005; 2008; Titchner et al., 2009].

For the second question, for both MSU and radiosonde datasets, there are two types of important error [Thorne et al., 2005a]. First is the structural uncertainty, which is the uncertainty caused by the choice of a single processing method from the set of all possible reasonable methods. This uncertainty can best be characterized by the spread of results from different datasets constructed by different research groups. Such an estimate is predicated upon the assumptions that all methods are reasonable (peer review being considered a necessary but not necessarily adequate condition) and that the very finite number of published estimates provides an unbiased estimate of the much larger spread of possible estimates. Second, there is internal uncertainty, unique to each product, that arises from uncertainty in parameters used to perform the various adjustments after a processing method is chosen. In the case of microwave sounders, this includes the uncertainty in calibration offsets, and the adjustments for instrument non-linearity and drifts in local measurement time. This type of uncertainty can be determined by an

analysis of the individual uncertainty in each of the adjustment steps used. In Mears et al. [2011], we performed such an analysis for the RSS datasets using a Monte-Carlo approach.

This paper assesses the issue of robustness of MSU/radiosonde intercomparisons to inevitable uncertainty in radiosonde and MSU records. Specifically, it aims to comprehensively address the suitability of such comparisons given the known data limitations. It aims to address the following questions and in so doing update the results of the various intercomparisons [e.g. Randall and Herman, 2008, Christy et al 2007, 2010] that have specifically looked to assess the quality of different MSU products by reference to radiosonde data and datasets:

1. Impacts of use of only 1 (as has been common), a subset, or all available radiosonde datasets
2. The impacts of different approaches to accounting for sampling mismatches between radiosonde and satellite data
3. The implications of undertaking analyses for limited time periods rather than the whole period of record.
4. What impact recently published comprehensive parametric uncertainty estimates for both one MSU and one radiosonde product may have on the results.
5. The physical interpretation of the difference between MT and LT used in some previous studies
6. Whether updated versions of many of the datasets previously considered impact the results and implications of such analyses

3. Datasets used in this study.

We focus our attention on MSU deep-layer diagnostics (and equivalent estimates from

radiosondes) of the temperature of the atmosphere with the bulk of the signal arising from the troposphere. The first MSU product, TMT (Temperature Middle Troposphere) is an average of the atmospheric temperature with a weighting function that extends from the surface to the lower stratosphere, and peaks about 5 km above the surface. The second MSU product, TLT (Temperature Lower Troposphere) involves a mathematical recombination of several of the off-nadir view scenes of the TMT channel [c.f. Mears et al., 2011, Fig. 4]. The TLT product has a weighting function that peaks much lower, about 2 km above the surface [Mears and Wentz, 2009a; Spencer and Christy, 1992].

MSU/AMSU equivalent temperatures have been calculated from the radiosonde data (gridded temperatures on standard pressure levels) using a method that assigns weights for each pressure level that depend on the surface type (land or ocean), and climatological value of the surface pressure for the grid point in each radiosonde dataset, thus taking into account changes due to the large differences in surface emissivity, and changes in the total atmospheric transmissivity due to changes in total mass of the atmosphere above the radiosonde site. Differences between this method and previous approaches, which typically used constant weights, are small for multidecadal changes except in regions where the material surface is at high altitude.

3.1. Satellite-based products

All satellite records considered here are derived from the MSU/AMSU series of microwave sounders, which make measurements near a complex of Oxygen absorption lines centered at 60 GHz [Smith et al., 1979]. For TMT, we consider the most recent versions of the dataset constructed by three different groups, the University of Alabama, Huntsville (UAH V5.4) [Christy et al., 2003], the Satellite Technology and Research Division at NOAA (STAR 2.0) [Zou et al., 2011, 2009; 2006], and Remote Sensing Systems (RSS V3.3) [Mears and Wentz, 2009b;

2009a]. For TLT, only two versions are available, UAH V5.4 and RSS V3.3. TLT is more influenced by radiation emitted by Earth's surface than TMT. Also, the diurnal cycle in surface skin temperatures is much larger than that in the free troposphere or even that of the near-surface air temperatures. Taken together with the uncertainty introduced through the weighted combination of view angles this yields larger uncertainties in TLT than TMT [Mears and Wentz, 2005; 2009a; Mears et al., 2011]. In all cases the independent groups have applied different methods to characterize and remove errors in the raw measurements associated with calibration errors and drifting local measurement times.

3.2. Radiosonde based products

For radiosondes, we take the approach of considering all published homogenized radiosonde datasets: HadAT [Thorne et al., 2005b], RAOBCORE [Haimberger, 2007], RICH [Haimberger et al., 2008], IUK [Sherwood et al., 2008], and RATPAC [Free et al., 2005; Lanzante et al., 2003]. The first four datasets are fully or partially (in the case of HadAT operational version) automated methods to find and estimate the size of “breakpoints” in the time series for a radiosonde station and create adjusted versions of the radiosonde data. The version of the RATPAC data used here, RATPAC-B, uses a manual breakpoint detection and adjustment method for data from 1958 to 1997. After 1997, no further adjustments were made. We use RATPAC-B because it is a gridded dataset, which is required to perform the spatial sampling steps in our analysis. The companion dataset RATPAC-A, contains automated adjustments up to the present time, but is not available in gridded form. We also show results for a subset of the RATPAC dataset (RATPAC_RW) which was developed by Randel and Wu [2006] to remove radiosonde stations with the largest errors in the stratosphere. It is not known the extent to which these errors are also present in

these stations' tropospheric measurements that are studied here. The RATPAC_RW subset is included here solely because it was the focus of the earlier satellite/radiosonde comparison study performed by Randall and Herman [2008].

3.3 Dataset internal uncertainty estimates

For the RSS (satellite) and HadAT (radiosonde) products there exist ensembles that attempt to quantify the dataset uncertainty [Mears et al., 2011, Titchner et al., 2009, Thorne et al., 2011b]. Such estimates can help to inform on whether differences between pairs of datasets arise through chance choices given the chosen methodological frameworks employed or reflect very real and substantial impacts of differences in methodological choices The MSU error ensembles were calculated by combining estimated errors in the adjustments for measurement time drift and realizations of the spatial/temporal sampling noise introduced by incomplete sampling [Mears et al., 2011]. The resulting error datasets are available at the same spatial and temporal resolution as the base dataset. This makes it possible to construct 400 realizations of this dataset that are consistent with the estimated uncertainty by adding them to the baseline, satellite derived temperature dataset. The HadAT ensemble used herein are derived from perturbed versions of the Hadley Centre's automated neighbor based homogenization procedure. Here we utilize the 20-member seasonal ensemble described in [Thorne et al., 2011b] that most closely recreated the real climate system behavior across a range of analogs to the real world created by Titchner et al. [2009]. Further details on the derivation of these uncertainty products are given in Supplementary Information and in the referenced papers.

We would stress that the radiosonde and MSU data ensembles being considered are fundamentally distinct. The Remote Sensing Systems MSU error model [Mears et al., 2011] is a perturbed ensemble around an assumption of essentially zero mean bias in the operational

product version. The radiosonde ensemble makes no assumption about correctness of the operational solution but rather undertakes fundamental end-to-end recalculations of the solution with a focus solely on breakpoint identification and adjustment issues to create the ensemble. Further, the sources of error are distinct for the MSU and radiosonde issues as would be expected given that they are fundamentally different observing technologies, each with unique issues. Therefore although the error estimates may be presented in a similar manner and may ostensibly appear very similar to the reader we would strongly caution against over-interpretation as they are not strictly intercomparable.

Despite these caveats, it is useful to combine the two error analyses to produce an estimate of the expected error in the RSS-HadAT difference time series that we evaluate herein, so that the statistical significance in the HadAT-RSS difference time series can be evaluated. To do this we difference pairs selected from the above ensembles to create an ensemble of difference time series. Because there are far fewer HadAT ensemble members available, each HadAT ensemble member is paired with 20 different RSS ensemble members, to yield an ensemble of 400 possible difference time series.

4. Time Series Comparisons

We begin by comparing the MSU and homogenized radiosonde temperature anomalies time series of large spatial scale averages which are useful because of the significant uncertainties in the measurements from isolated radiosondes and in single MSU grid points. Over larger spatial scales, many components of these uncertainties are reduced by the averaging procedure. We choose to focus on monthly means, averaged over nearly the entire globe (75S to 75N, “global”), or deep tropics (20S to 20N, “tropical”). In both regions, the radiosonde spatial coverage is far from complete. Figure 1 shows a typical radiosonde sampling pattern. We also performed our

analysis for the northern and sounthern extratropics separately. Summary plots from these analyses are presented in the supplementary material.

A simple (and common) way to construct time series from gridded data is to use area-weighted means of all available data. Comparing time series of simple area-weighted (AW) global averages of radiosonde data with the area-weighted means of the more spatially complete MSU data can lead to substantial discrepancies, due to the large areas that are unsampled by radiosondes, and the changes in radiosonde sampling over time. Additional discrepancies occur because the radiosonde sampling patterns in the tropics exclude the eastern tropical Pacific Ocean, where the ENSO signal is often the strongest. A good approach to resolving these issues is to sample the MSU data at the actual radiosonde sampling for each month [Free and Seidel, 2005; Mears and Wentz, 2009b], and then compute an area-weighted average from the sub-sampled data for each month to produce a "global" average. This modifies the MSU means such that they more closely match the area-weighted radiosonde means, and automatically takes into account the presence or absence of a radiosonde measurement for a given location and month and thus changes in spatial sampling over time. We refer to the sampled satellite means as "sampled at radiosonde locations" (SRL).

Because simple AW averages have often been used to perform radiosonde/satellite intercomparisons, we show the results for both methods to assess sensitivity to this choice. In Figs. 2a and 2b, we show AW global time series for the operational HadAT product and each MSU dataset for both TLT and TMT. The large month-to-month variations in all datasets make it difficult to draw conclusions from this plot. In Figs. 2c and 2d, we show the AW difference time series (HadAT – MSU) for each MSU dataset, and in Figs. 2e and 2f, we show the same difference series, except using SRL averaging to calculate the satellite time series. In all cases

shown, the SRL time difference time series exhibit much less variance than the AW time series. We made similar plots for the other 5 radiosonde datasets we consider (see supplemental on-line material). In all cases, the standard deviations of the difference time series were reduced, particularly on short time scales. Even when the difference time series were filtered to remove variability on time scales shorter than one year using a digital filter [Lynch and Huang, 1992], 11 of 12 TLT cases and 13 of 18 TMT cases showed reduced standard deviation, suggesting that the SRL procedure also tends to improve the agreement on interannual and longer time scales. These results reinforce our previous conclusion [Mears and Wentz, 2009b] that MSU/radiosonde comparisons are best performed using SRL averaging.

Even with SRL sampling, the difference time series show significant variability on intra-annual time scales that makes it difficult to draw conclusions. One approach that has been used to help reduce the contribution of short-time scale variability is the analysis of trends in intermediate-length sub-samples of a longer time series. Randall and Herman [2008] used the Randel and Wu [2006] subset of the RATPAC radiosondes to analyze 5 and 10 year trends in tropopspheric UAH and RSS data. In Figs. 2g and 2h, we show plots of trends of rolling 5-year sub-samples of the difference time series, with each slope plotted at the location of the center of the sub-sample. When the 5-year trend is greater than zero, the MSU data warmed relative to the radiosonde data over the 5-year period, and conversely, when the 5-year trend is less than zero, the MSU data cooled relative to the radiosonde data. It is immediately obvious that the 5 or 10-year trends can accentuate both the intermediate and long time-scale differences, as concluded by Randall and Herman. For example, in the TMT data (Fig. 2f) the MSU data (for all 3 MSU datasets) warms relative to HadAT over the 1990-1998 period. This is easier to see in Fig. 2h as large maxima in the short-term trends which reach their greatest magnitude at the center of this

period.

In Figs. 3 (and 4), we plot the 5-year trends in global (tropical) difference time series for both TLT and TMT and for all combinations of radiosonde and MSU datasets (in the on-line supplementary material, we show similar sets of plots for the northern and southern extratropics, see Figs S6 and S7). There are several common features that stand out. For TLT, perhaps the most obvious is a peak in the RSS and UAH minus radiosonde differences centered near 1995 which is consistently somewhat larger for RSS. This feature, due to the warming of MSU data relative to radiosonde data, has been previously discussed in the literature as it occurs during the period where the RSS warms relative to the UAH dataset [Christy and Norris, 2009; Christy et al., 2007; Randall and Herman, 2008]. The better agreement between UAH and the radiosondes (as shown by the lower peak for UAH) during this period, combined with the sign of temperature changes in the tropics during the period surrounding the eruption of Pinatubo, has been used by these authors to argue that the RSS dataset contains a warming bias during this period. The analysis of this period is complicated by the competing effects of volcano-induced cooling and ENSO-induced warming. This is also the period over which general improvements to radiosonde solar radiation shielding yielded an apparently artificial cooling across much of the radiosonde network [Sherwood et al., 2005], an effect which may not entirely have been removed with available radiosonde datasets [e.g. Sherwood et al., 2008]. Hence interpretation of these differences is fraught with physical and instrumental considerations that significantly inhibit a clean inference regarding which MSU product may be closer to the unknown true temperature evolution.

We note that there are two other prominent features in the 5-year trend plots that have not received as much attention. First, there is a second common feature, a low point centered near

2003. This indicates a period when the MSU datasets are cooling relative to the radiosonde datasets. This feature roughly coincides with the end of the data record for NOAA-14, the last MSU satellite, and may be related. We note that for TMT, there is a unexplained trend difference between MSU and AMSU measurements during 1999-2005 [Mears et al., 2011]. Second, UAH tends to show a peak centered near 1986, which is either absent or much smaller for RSS and may be related to calibration problems with the NOAA-9 satellite, a finding confirmed by Po-Chedley and Fu [2012]. Over the entire time series, the effects of the relative warming in the 1990's and the relative cooling in the 2000's tend to cancel, leaving the 32-year trends for MSU and radiosonde data in relatively good agreement [Mears et al, 2011].

For the TLT plots (left hand column) in Fig. 3, we also plot the 5 year trend differences for RSS-UAH. Using the mean absolute value of each of the trend difference curves over the entire time period in Figure 5a in all cases except for RAOBCORE and RICH, the MSU datasets are in closer agreement with each other than they are with the radiosonde datasets. (We also evaluated the differences using the root-mean square difference as a difference metric, which yielded nearly identical results. See supplemental on-line material, Fig S8.) Note that the RAOBCORE dataset, which has been asserted to be corrupted by anomalous warming due to underlying errors in the ERA-40 reanalysis used in its construction, [Christy et al., 2010] is the radiosonde dataset that agrees best with both satellite datasets when evaluated using our method. Given that the reanalysis field used to derive adjustments in RAOBCORE is strongly influenced by the MSU/AMSU data (amongst others) this is perhaps not surprising. The RICH dataset generally agrees second best with the MSU data. This dataset is more independent from the background reanalysis, and thus is less subject to the criticisms put forth in [Christy et al., 2010].

The ordering of the level of agreement between the radiosonde datasets and any of the MSU

datasets (RAOBCORE, RICH, IUK, HadAT, RATPAC) is the same as the ordering of the radiosonde trends over the entire satellite era. If the 5-year trends were strongly influenced by the overall trend, this would be expected on mathematical grounds, as the overall trend is related to the accumulated 5-year trends. However, the difference between 5-year trends is dominated by differences on short time scales. We checked this by performing a second set of calculations with the overall trend removed from each time series before the 5-year trends were calculated. The resulting version of Fig. 5 (see supplemental on-line material, Fig S9) is nearly identical.

For TMT, we also include data from the STAR MSU dataset. Again, there are several common features across all MSU and radiosonde datasets, including the peak in the mid 1990s, and a minimum in the mid 2000s. There is a second sharp minimum centered near 1987 in the UAH curve, a feature that is significantly reduced in the RSS data, and almost nonexistent in the STAR data. Also note the appearance of a strong seasonal cycle in the UAH data after 1998. (This feature is even more prominent in Fig. 2., and may be related to the method used to combine MSU and AMSU measurements. AMSU measurements began in the middle of 1998). Again, the difference between the MSU datasets and the radiosonde datasets is larger than the spread between the MSU datasets themselves, as shown by the mean absolute values of the difference curves that are plotted in Fig. 5b.

Fig. 4 is analogous to Fig. 3, except the data are averaged over the deep tropics (20S to 20N) instead of the entire globe, with summary results presented in Figs. 5c and 5d. The set of tropical plots shares many of the features of the global plots, such as the relative warming in the MSU datasets in the mid 1990's, and the relative cooling in the late 1980's and early in the 21st century. One important difference is that there appear to be more differences between the radiosonde datasets in the 1990's, with the MSU datasets showing strong warming relative to

radiosondes in IUK and RATPAC, with considerably less warming relative to RAOBCORE and (to a lesser extent) RICH. For HadAT, most of the relative MSU warming is shifted to a short time period in the early 1990's. It is not surprising that the differences between datasets are larger in the tropical case because the number of radiosonde stations in the sample has decreased, leading to more variability. Also, in the tropics, the radiosonde coverage is more sparse that in the northern extratropics, making it more difficult to perform the necessary adjustments for those methods that rely to some extent upon comparisons with near neighbors (HadAT, IUK, RICH, and to a lesser degree, RATPAC). Finally, most tropical sites have had daytime-only ascents which are most impacted by solar heating effects and these were significantly mitigated through the 1990s [Sherwood et al., 2005, Randel and Wu, 2006] leading to a false cooling signal in the raw record, which may not have been entirely removed. Differences may relate solely to the efficacy of the various radiosonde products in removing this artifact and it is important to note that the sign and timing implies that possibly none of the radiosonde products have adequately removed this artifact rather than a consistent bias in MSU products. In Fig. 5c (TLT), the best agreement between tropical satellite and MSU results is for the RAOBCORE and RICH datasets, with the UAH dataset in better agreement than RSS except for the IUK dataset, and in Fig. 5d, the best agreement again is with the ROABCORE and RICH datasets, but with the UAH dataset typically showing more disagreement than the RSS and STAR datasets. For TMT, the RSS and STAR datasets are extremely close to each other on the 5-year scale, despite substantial differences in 32-year trends.

The results for RATPAC-RW shown here differ from those presented in Randall and Herman [2008] for two important reasons. First, newer versions of the both the RSS and UAH MSU datasets were used in this analysis. Second, and more importantly, Randall and Herman did not

subsample the MSU data at the radiosonde locations, but instead compared global radiosonde averages to area-weighted, land-only MSU averages. A majority (29 of 47) of the RATPAC-RW stations are located on islands (9 stations) or in coastal regions (20 stations) and thus are not representative of a land-only average (see supplementary material for a map showing these results, and a precise description of our definition of land, coastal, or ocean).

Despite extensive efforts we were unable to exactly replicate the results of Randall and Herman due to insufficient methodological clarity in their paper and thus no direct comparison is possible here. To illustrate the relative importance of variations in sampling method and dataset updates, we present several alternative versions of Fig. 5b in the supplementary material (see Fig S10). We find that the changes in dataset version are less important than changes in sampling method, and that while the use on land-only MSU data reduces the mean absolute difference relative to the use of global land-and-ocean averages, both are substantially worse than sampling at the radiosonde locations.

Another question to investigate is the degree to which agreement on the 5-year time scale is useful for predicting agreement on a longer time scale. In Fig. 6, we plot the absolute value of the difference between multi-decadal trends in globally-average TMT (1979-2010, except for IUK, which ends in 2006) as a function of the mean absolute value of each of the trend difference curves. This measure of 5-year trend agreement is the same as is plotted in Fig. 5. The plot shows very little correlation (correlation coefficient = 0.025) between the level of agreement on 5 year time scales, and the agreement between multi-decadal trends. This suggests that the agreement between 5 year trend time series is essentially useless for predicting agreement on longer time scales.

5. Impact of Uncertainty Estimates.

The preceding section serves as an estimate of the structural (or between method) uncertainty in MSU/radiosonde comparisons. We now investigate the impact of internal uncertainty on both the MSU and radiosonde datasets. Errors in both types of dataset are often correlated in both time and location. Only the RSS and HadAT datasets have associated internal uncertainty estimates that are sufficiently detailed to accurately estimate the error in trends at various time scales (Section 2.3).

For the RSS/HadAT case, in Fig. 7a and 7b, we show the median difference between 5-year RSS - HadAT difference trend error ensemble (we emphasize that the median of the HadAT error ensemble is quite different from the operational version of HadAT considered in Section 3). We also plot the 95% confidence interval (CI) around the median difference. The 95% confidence interval was calculated from an ensemble of 400 realizations of the RSS-HadAT difference. The nth member of this ensemble was constructed by subtracting the (n modulo 20)th member of the HadAT ensemble from the nth member of the RSS ensemble. For both TLT and TMT, the 95% CI range for the differences between RSS and HadAT encompasses the zero line 59% of the time, implying that the differences between the 5-year trends are larger than can be easily accounted for by the combination of internal errors. We note that the HadAT adjustment procedure was not designed to remove errors on short time scales, so that the short-term error represented by the ensemble may underestimate the true error in the radiosonde data.

Figs 8a-c summarize a similar analysis of the differences between the RSS and UAH satellite datasets (8a and 8b), and between the RSS and STAR satellite datasets (8c). Because detailed uncertainty ensembles are not available for the UAH and STAR datasets, this part of the analysis only includes uncertainty estimates for the RSS datasets. For TLT, the RSS-UAH 95%

uncertainty range encompasses the zero line 63% of the time, and for TMT the uncertainty ranges encompass the zero line for 30% (UAH) and 62% (STAR) of the time period. We speculate that if the uncertainty in the UAH and STAR datasets were included, these percentages would increase, but the exact amount is difficult to reliably estimate without comparably comprehensive uncertainty analyses from the UAH and STAR groups.

6. Inter-channel Differences.

Randall and Herman (2008) (hereafter RH2008) also studied the differential trends between the TLT and TMT layers for the UAH, RSS, and RATPAC-RW datasets. They motivated this work as a method for diagnosing the impact of the diurnal cycle on the merged dataset, since the adjustments made for changes in local measurement time are much larger for TLT than TMT. We recommend using such differences with caution. Because of the overlap between the TLT and TMT weighting functions and the subsequent cancellation caused by differencing, the TLT – TMT difference contains a large amount of information from the surface and lower stratosphere. Furthermore, a large portion of the resulting weighting kernel (including most if not all of the atmosphere above 6Km) has a negative weighting which is hard to interpret in a physical manner. Figure 9 shows the temperature weighting functions for TLT, TMT, and the TLT – TMT difference. About 12% of the total absolute value of the weight comes from surface emissions, and about 25% comes from above 12 km, or 200 hPa. We stress that these differences should not be thought of as the difference in temperature between the lower and middle troposphere. This is particularly a concern for radiosonde data because of the increase in the relative weight of radiosonde measurements at pressures ≤ 200 hPa where the adjusted radiosonde datasets may be less reliable [Randel and Wu, 2006].

With these caveats in mind, here we update (to the extent we are able to replicate) and extend the RH2008 analysis. There are 4 important differences between the present analysis and RH2008. First, we use the most recent versions of the available datasets. Second, we consider all available homogenized radiosonde datasets (Section 2.2) rather than a single estimate. Third, we use SRL averaged MSU data instead of land-only area-weighted data (Section 3). And fourth, we consider results from outside the limited temporal ranges plotted in RH2008. Figure 10 shows the 5-year and 10-year trend differences between TLT and TMT for each radiosonde dataset. In each case, the SRL MSU averages are plotted, which accounts for the difference in the RSS and UAH curves between plots. These plots correspond to Fig. 4 in RH2008, which they used to argue that the UAH satellite datasets were more accurate than the RSS versions. The bottom row of plots show results from the RATPAC-RW dataset analyzed by RH2008. In Table 1, we present a summary of the mean absolute differences (MAD) and the number of months that each satellite dataset is closer to the radiosonde data set. For the 5-year trends, the 3 datasets are in reasonably good agreement over the bulk of the time period, in agreement with the findings of RH2008, though in general, we find that the RSS data are in better agreement with the radiosonde data than the UAH data for these metrics (Table 1). For the 10 year time period, we find that the largest differences are outside the region plotted by RH2008, and that within the 1993-2002 period plotted, the RSS dataset is in better agreement with the radiosonde data than UAH, in direct contradiction to the findings of RH2008. The probable reasons for this different result are both the use of MSU data sampled at the radiosonde locations (instead of land-averaged satellite data), and (to a lesser degree) the use of updated versions of the MSU datasets. The largest differences between the MSU datasets and radiosonde datasets tend to occur in the early part of the time series for both the 5-year and 10-year trends, outside the region

plotted in RH2008. Again, we find that the two MSU datasets tend to be closer to each other than to any of the radiosonde datasets.

For the other 5-year period plots, the conclusions are similar to that reached above for the RATPAC-RW data. The MSU datasets are in fairly good agreement with the radiosonde datasets during the period plotted in RH2008. During the pre-1990 period the analyses typically show TLT warming relative to TMT more in the radiosonde data than in the MSU datasets, with the MSU datasets being relatively similar. For the 10 year time period plots, this difference becomes more important, with TLT warming much more than TMT in the radiosonde datasets. During the 1993-2002 period plotted in RH2008, the results differ substantially from radiosonde dataset to radisonde dataset, even for the MSU dataset SRL estimates, which makes it difficult to draw conclusions. Note that the only difference between the different versions of the MSU data is the spatial/temporal sampling used to construct the averages. After about 2000, TLT tends to warm relative to TMT more in the radiosonde datasets, except for HadAT, where TLT cools slightly relative to TMT after about 2004.

An improvement that can be made to the RH2008 method is to consider the difference between TLT and the "total troposphere" (TT) MSU product proposed by Fu and Johanson [2005]. This product has reduced weight in the stratosphere, and thus is less affected by overall stratospheric cooling, and stratospheric warming events caused by volcanic eruptions. In the supplementary material, we show an alternative version of Fig. 10 calculated using TT instead of TMT. This replacement generally reduces the variability of the time series, probably due to the reduction of the influence of volcano-induced stratospheric warming, but does not alter the conclusions.

7. Discussion

We have used methods similar to those presented in RH2008 to analyze 5- and 10-year trends in adjusted radiosonde and Microwave Sounding Unit (MSU) measurements of tropospheric temperature utilizing an inclusive range of MSU and radiosonde products. In all cases we find that there are several time periods during which there is substantial disagreement between 5-year trends in radiosonde datasets and 5-year trends in the MSU datasets. Sometimes these differences cancel over longer time periods, perhaps leading to false or overly confident conclusions about the agreement between satellite and radiosonde datasets on multi-decadal time scales. When data from different MSU – radiosonde pairs are examined, the results indicate that all MSU/sonde differences share many common features, and that in most cases, the differences between radiosondes and MSU is much larger than between different MSU datasets, or between different radiosonde datasets. Given the current state of knowledge, we are unable to determine whether this commonality is due to shared problems in the MSU datasets, or to shared problems with the radiosonde datasets, or a combination of both. It is possible that both types of datasets retain substantial common biases within their respective types. For MSU data the three different versions are derived from identical raw source data. If there is a time-dependent bias in the raw data that none of the merging procedures are able to detect and remove, then the common bias would obviously remain in all three datasets. A similar argument holds for the radiosonde datasets, though in this case, the underlying, unadjusted datasets differ in the number and locations of radiosonde stations used.

In addition, an analysis of the internal error in the MSU datasets suggests that the differences between RSS and UAH 5-year trends are possibly not statistically significant for TLT, while the differences between the RSS, STAR and UAH datasets may be significant for TMT. This type of

analysis is hampered by the lack of a detailed error analysis in the UAH and STAR products.

Although radiosonde MSU comparisons have some information content, on their own they are ill-posed to assess satellite dataset quality issues because both types of data almost certainly retain unknown and poorly quantified biases. Additional entirely independent measurements such as from the Hyperspectral Infrared Sounder, GPS Radio Occultation or reanalyses may help. But these are additionally fraught by a variety of issues relating to sampling (clear sky only for HIRS, a different atmospheric volume and temporal samples of opportunity for GPS-RO), interpretation (both satellite measures are responsive to more than just temperatures), period of record, and independence of record (particularly so for reanalyses). Despite this, bringing in such additional independent estimates may offer a future avenue of investigation. Additional insights may accrue from physical rather than wholly statistical interpretation of the records. Finally, real insights on biases and their causes will only accrue through additional in-depth analyses of the observations and accompanying metadata themselves to better understand the causes of biases and differences in the respective records.

In conclusion, when the similarity of the MSU datasets relative to radiosonde datasets is combined with the lack of statistical significance in many of the difference findings, we conclude that trying to determine which MSU dataset is “better” based on short-time period comparisons with radiosonde datasets alone cannot lead to robust conclusions. This is trivially true for any case where two poorly constrained and understood measurements of the same measurand exist. When they disagree the problem is under-constrained such that it is solely possible to conclude that one or both of the measurements is (are) biased relative to the true state of the measurand. Sadly, this is all too common in climate and is why SI traceable measurement programs such as the GCOS Reference Upper Air Network [Seidel et al., 2009] are vital to our future ability to

monitor the changing climate.

References

Christy, J. R., B. M. Herman, R. Pielke, Sr., P. Klotzbach, R. T. McNider, J. J. Hnilo, R. W. Spencer, T. Chase, and D. H. Douglass (2010), What do Observational Datasets Say about Modeled Tropospheric Temperature Trends sinde 1979?, *Remote Sensing*, *2*, 2148-2196, doi: 10.3390/rs2092148.

Christy, J. R., and W. B. Norris (2006), Satellite and VIZ-Radiosonde Intercomparisons for Diagnosis of Nonclimatic Influences, *Journal of Atmospheric and Oceanic Technology*, *23*(9), 1181-1194, doi:10.1175/JTECH1937.1.

Christy, J. R., and W. B. Norris (2009), Discontinuity Issues with Radiosonde and Satellite Temperatures in the Australian Region 1979–2006, *Journal of Atmospheric and Oceanic Technology*, *26*(3), 508-522.

Christy, J. R., W. B. Norris, R. W. Spencer, and J. J. Hnilo (2007), Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements, *Journal of Geophysical Research*, *112*, doi:doi:10.1029/2005JD006881.

Christy, J. R., R. W. Spencer, W. B. Norris, W. D. Braswell, and D. E. Parker (2003), Error estimates of version 5.0 of MSU-AMSU bulk atmospheric temperatures, *Journal of Atmospheric and Oceanic Technology*, *20*(5), 613-629.

Free, M., and D. J. Seidel (2005), Causes of differing temperature trends in radiosonde upper air data sets, *Journal of Geophysical Research*, *110*, doi:doi:10.1029/2004JD005481.

Free, M., D. J. Seidel, J. K. Angell, J. R. Lanzante, I. Durre, and T. C. Peterson (2005), Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A new dataset of large-area anomaly time series, *Journal of Geophysical Research*, *110*, doi:10.1029/2005JD006169.

Haimberger, L. (2007), Homogenization of radiosonde temperature time series using innovation statistics, *Journal of Climate*, *20*(7), 1377-1403.

Haimberger, L., C. Tavolato, and S. Sperka (2008), Towards the elimination of warm bias in historic radiosonde records -- some new results from a comprehensive intercomparison of upper air data, *Journal of Climate*, *21*, 4587-4606.

Lanzante, J. R., S. Klein, and D. J. Seidel (2003), Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities, and MSU comparison, *Journal of Climate*, *16*, 241-262.

Lanzante, J. R., T. C. Peterson, F. J. Wentz, and K. Y. Vinnikov (2006), What do observations indicate about the change in temperatures in the atmosphere and at the surface since the advent of measuring temperatures vertically, in *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*, A Report by the Climate Change Science Program and the Subcommittee on Global Change Research,

Washington DC.

Lynch, P., and X. Y. Huang (1992), Initialization of the HIRLAM model using a digital filter, *Monthly Weather Review*, *120*, 1019-1034.

Mears, C. A., C. E. Forest, R. W. Spencer, R. S. Vose, and R. W. Reynolds (2006), *What is our understanding of the contribution made by observational or methodological uncertainties to the previously reported vertical differences in temperature trends?*, U.S. Climate Change Science Program, Washington, DC.

Mears, C. A., and F. J. Wentz (2005), The effect of drifting measurement time on satellite-derived lower tropospheric temperature, *Science*, *309*, 1548-1551.

Mears, C. A., and F. J. Wentz (2009a), Construction of the RSS V3.2 lower tropospheric dataset from the MSU and AMSU microwave sounders, *Journal of Atmospheric and Oceanic Technology*, *26*, 1493-1509.

Mears, C. A., and F. J. Wentz (2009b), Construction of the Remote Sensing Systems V3.2 atmospheric temperature records from the MSU and AMSU microwave sounders, *Journal of Atmospheric and Oceanic Technology*, *26*, 1040-1056.

Mears, C. A., F. J. Wentz, P. W. Thorne, and D. J. Bernie (2011), Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique, *Journal of Geophysical Research*, *116*, D08112, doi:10.1029/2010JD014954.

Po-Chedly, S., Q. Fu (2012), A bias in the midtropospheric channel warm target factor on the NOAA-9 microwave sounding unit, *Journal of Atmospheric and Oceanic Technology*, *29*, 646-652.

Randall, R. M., and B. M. Herman (2008), Using limited time period trends as a means to determine attribution of discrepancies in microwave sounding unit-derived tropospheric temperature time series, *Journal of Geophysical Research*, *113*, doi:10.1029/2007JD008864.

Randel, W. J., and F. Wu (2006), Biases in stratospheric and tropospheric temperature trends derived from historical radiosonde data, *Journal of Climate*, *19*(10), 2094-2104.

Seidel, D. J. et al. (2004), Uncertainty in signals of large-scale climate variations in radiosonde and satellite upper-air datasets., *Journal of Climate*, *17*, 2225-2240.

Sherwood, S. C., J. R. Lanzante, and C. L. Meyer (2005), Radiosonde daytime biases and late 20th century warming, *Science*, *309*, 1556-1559.

Sherwood, S. C., C. L. Meyer, R. J. Allen, and H. A. Titcher (2008), Robust tropospheric warming revealed by iteratively homogenized radiosonde data, *Journal of Climate*, *21*(20), 5336-5352.

Smith, W. L., H. M. Woolf, C. M. Hayden, D. Q. Wark, and L. M. McMillan (1979), The TIROS-N operational vertical sounder, *Bulletin of the American Meteorological Society*, *60*, 1177-1187.

Spencer, R. W., and J. R. Christy (1992), Precision and radiosonde validation of satellite gridpoint temperature anomalies. Part II: A tropospheric retrieval and trends during 1979-1990, *Journal of Climate*, *5*, 858-866.

Thorne, P. W., J. R. Lanzante, T. C. Peterson, D. J. Siedel, and K. P. Shine (2011a), Tropospheric temperature trends: history of an ongoing controversy, *Wiley Interdisciplinary Reviews Climate Change*, *2*(1), 66-88, doi:10.1002/wcc.80.

Thorne, P. W., et al. (2011b), A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes, *J. Geophys. Res.*, 116, D12116, doi:10.1029/2010JD015487.

Thorne, P. W., D. E. Parker, B. D. Santer, M. P. McCarthy, D. M. H. Sexton, M. J. Webb, J. M. Murphy, M. Collins, H. A. Titchner, and G. S. Jones (2007), Tropical vertical temperature trends: a real discrepancy?, *Geophysical Research Letters*, *34*, L16702, doi:10.1029/2007GL029875.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears (2005a), Uncertainties in climate trends: Lessons from upper-air temperature records, *Bulletin of the American Meteorological Society*, *86*(10), 1437-1442.

Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan (2005b), Revisiting radiosonde upper-air temperatures from 1958 to 2002, *Journal of Geophysical Research*, *110*, doi:10.1029/2004JD005753.

Titchner, H. A., P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, and D. E. Parker (2009), Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments, *Journal of Climate*, *22*(3), 465-485.

Zou, C. Z., M. Gao, and M. D. Goldberg (2009), Error structure and atmospheric temperature trends in observations from the Microwave Sounding Unit, *Journal of Climate*, *22*, 1661-1681.

Zou, C. Z., M. D. Goldberg, Z. Cheng, N. C. Grody, J. T. Sullivan, C. Cao, and J. D. Tarpley (2006), Recalibration of microwave sounding unit for climate studies using simultaneous nadir overpasses, *Journal of Geophysical Research*, *111*, doi:10.1029/2005JD006798.

Zou, C.-Z., and W. Wang (2011), Intersatellite calibration of AMSU-A observations for weather and climate applications, *J. Geophys. Res.*, *116*, D23113, doi:10.1029/2011JD016205.

Figure Captions:

Fig. 1. Sampling pattern for the HadAT dataset for January, 2003. Grey boxes represent locations of valid data (5 degree latitude by 10 degree longitude gridboxes with at least one reporting radiosonde station for the month), and white regions are areas where data is not available.

Fig. 2. Graphical representation of various stages in processing the radiosonde and satellite data. The left column is TLT data, and the right side is TMT. The top row (a and b) shows unmodified globally average time series for the HadAT radiosonde dataset, and each of the satellite datasets. The second row (c and d) shows the difference between HadAT and each of the satellite datasets (satellite - HadAT) using area-weighted averaging. The third row (e and f) shows the difference between HadAT and each of the satellites datasets using SRL averaging for the satellite data. Note the large reduction of differences on most time scales. The bottom row (g and h) shows 5-year trends calculated from the SRL differences on the row above.

Fig. 3. Difference in global SRL-averaged 5-year trends (Satellite – Radiosonde) for each radiosonde and satellite dataset. The left column is for the TLT channel, and the right is for TMT. For TLT, we also plot the SRL difference between the two satellite datasets (UAH – RSS).

Fig.4. Same as Fig.3, except that the data are averaged over radiosonde locations in the deep tropics (20S-20N). Similar sets of plots for the southern and northern extratropics are available in the supplemental material (Figs. S6 and S7.)

Fig. 5. Summary of the differences between 5-year trends using the mean absolute difference as a metric. 5a. TLT, global. 5b. TMT, global. 5c. TLT, deep tropics (20S-20N). 5d. TMT, deep tropics. In most cases, the difference between different satellite datasets is substantially less than

the difference between the satellite datasets and the radiosonde datasets. The exception to this is for the RAOBCORE and RICH datasets for TLT.

Fig. 6. Scatter plot of the absolute value of the 32-year trend difference (1979-2010) in global means of TMT as a function of mean absolute deviation of the 5-year trend difference time series computed from the same difference time series. Each point represents the comparison of one dataset (satellite or radiosonde) with another (satellite or radiosonde).

Fig. 7. Differences between 5 year trends with uncertainty estimates for channels TLT and TMT. All time series and uncertainty estimates are global averages sampled at the location of HadAT radiosonde measurements (which end in 2003 in this ensemble and hence the series truncation). The thick black line is the median difference time series, and the yellow region surrounding it contains 95% of the instances in the uncertainty ensemble. The uncertainty ensemble includes the estimated uncertainty in both the RSS and HadAT datasets.

Fig. 8. Differences between 5 year MSU trends with RSS-only uncertainty estimates for channels TLT and TMT. a) RSS-UAH for TLT. b) RSS-UAH for TMT, c) RSS-STAR for TMT. In all panels, the thick black line is the median difference time series, and the yellow region surrounding it contains 95% of the instances in the RSS uncertainty ensemble.

Fig. 9. Weighting functions as a function of height for TLT, TMT, and the TLT minus TMT difference for ocean scenes. The rectangular areas located below zero height show the portion of the total weight due to surface emission. For land scenes, this contribution would be about two times larger, with a corresponding reduction in atmospheric weight. Because of the large

absolute weight in the stratosphere and the surface, the TLT minus TMT difference should not be interpreted as the difference between the lower and middle tropospheric temperatures.

Fig. 10. Differences (TLT minus TMT) in short term trends in atmospheric temperature between adjacent layers for each radiosonde dataset. In each case, the satellite data is sampled at the radiosonde locations, and the differences in spatial sampling lead to the differences between the satellite results for the different panels. The temporal range plotted in RH2008 is shaded in yellow.

Table 1. Interlayer Difference Statistics for 5 and 10 year trends.

	5-year Trend Differences				10-year Trend Differences			
	Number of Months RSS Closer	Number of Months UAH Closer	MAD RSS-Sonde (k/decade)	MAD UAH-Sonde (K/decade)	Number of Months RSS Closer	Number of Months UAH Closer	MAD RSS-Sonde (k/decade)	MAD UAH-Sonde (k/decade)
HadAT	218	109	0.1006	0.1511	210	57	0.0628	0.0791
RAOBCORE	205	122	0.0745	0.1266	172	95	0.0505	0.0847
RICH	222	105	0.0597	0.1219	202	65	0.0291	0.0793
IUK	178	89	0.0647	0.1289	150	57	0.0373	0.0794
RATPAC	217	99	0.0741	0.0984	205	51	0.0339	0.0670
RATPAC_RW	215	101	0.0930	0.1009	174	82	0.0456	0.0690

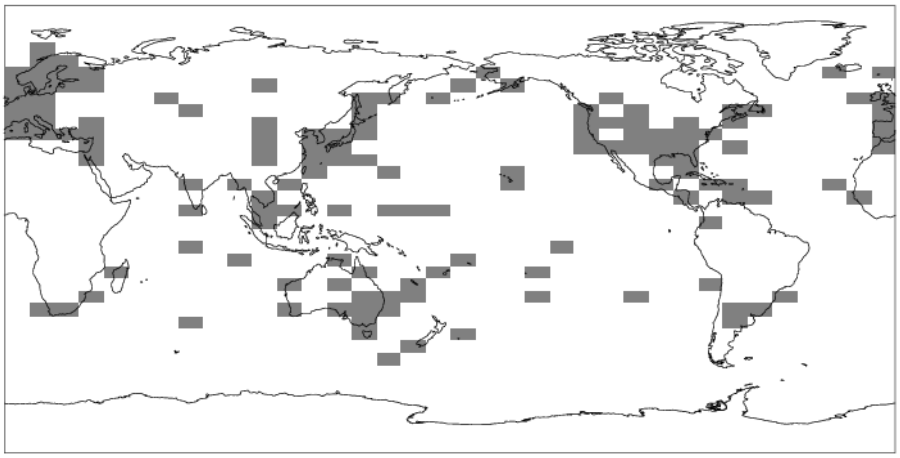


Fig. 1. Sampling pattern for the HadAT dataset for January, 2003. Grey boxes represent locations of valid data (5 degree latitude by 10 degree longitude gridboxes with at least one reporting radiosonde station for the month), and white regions are areas where data is not available.

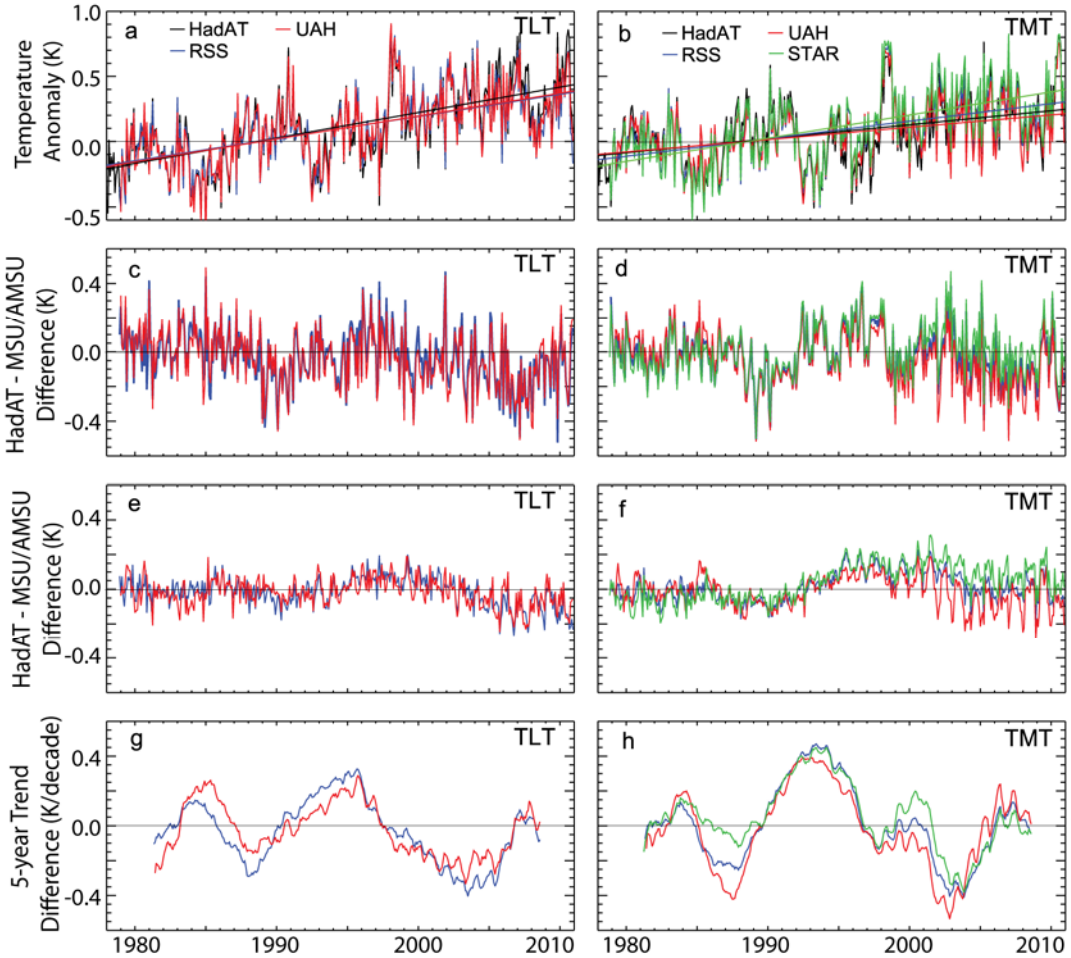


Fig. 2. Graphical representation of various stages in processing the radiosonde and satellite data. The left column is TLT data, and the right side is TMT. The top row (a and b) shows unmodified globally average time series for the HadAT radiosonde dataset, and each of the satellite datasets. The second row (c and d) shows the difference between HadAT and each of the satellite datasets (satellite - HadAT) using area-weighted averaging. The third row (e and f) shows the difference between HadAT and each of the satellites datasets using SRL averaging for the satellite data. Note the large reduction of differences on most time scales. The bottom row (g and h) shows 5-year trends calculated from the SRL differences on the row above.

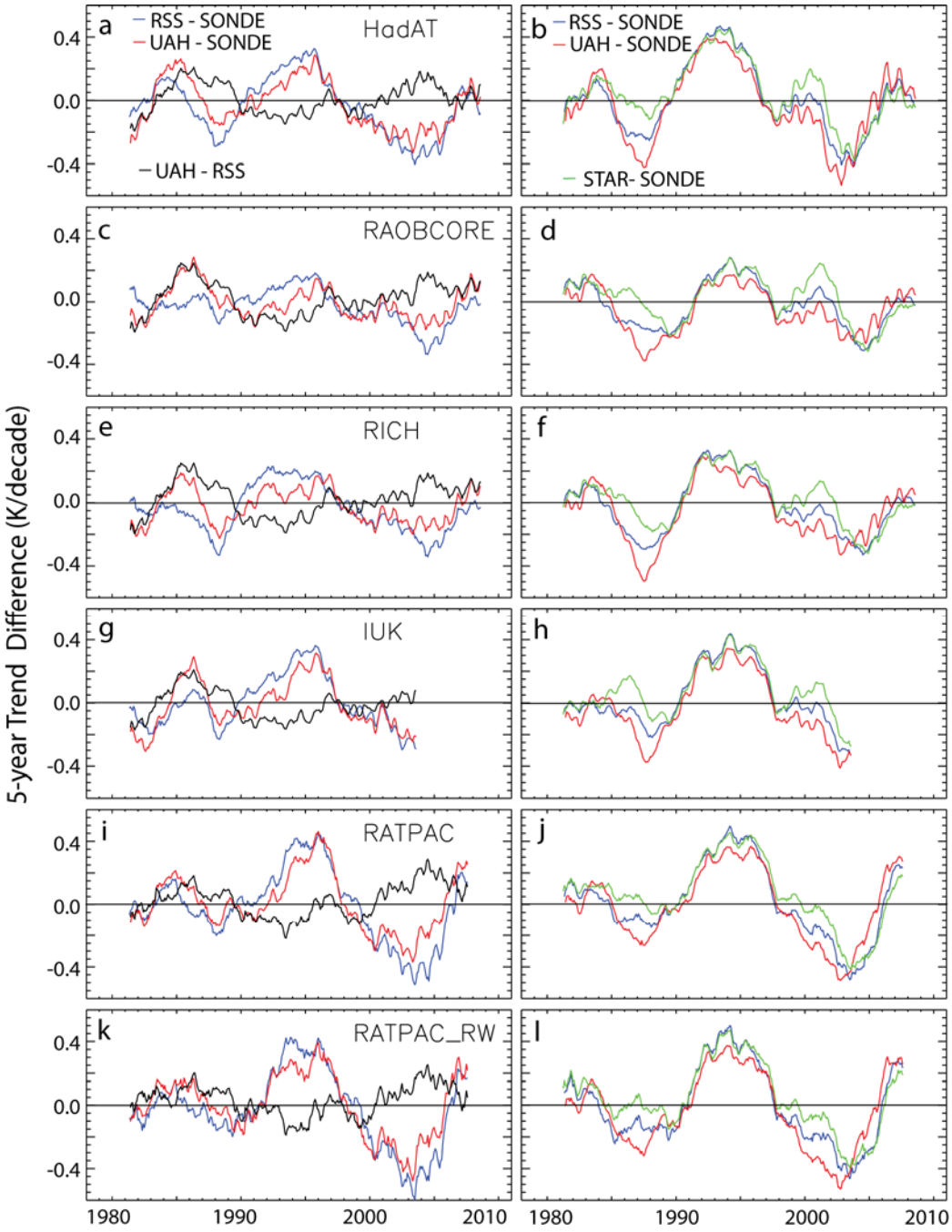


Fig. 3. Difference in global SRL-averaged 5-year trends (Satellite – Radiosonde) for each radiosonde and satellite dataset. The left column is for the TLT channel, and the right is for TMT. For TLT, we also plot the SRL difference between the two satellite datasets (UAH – RSS).

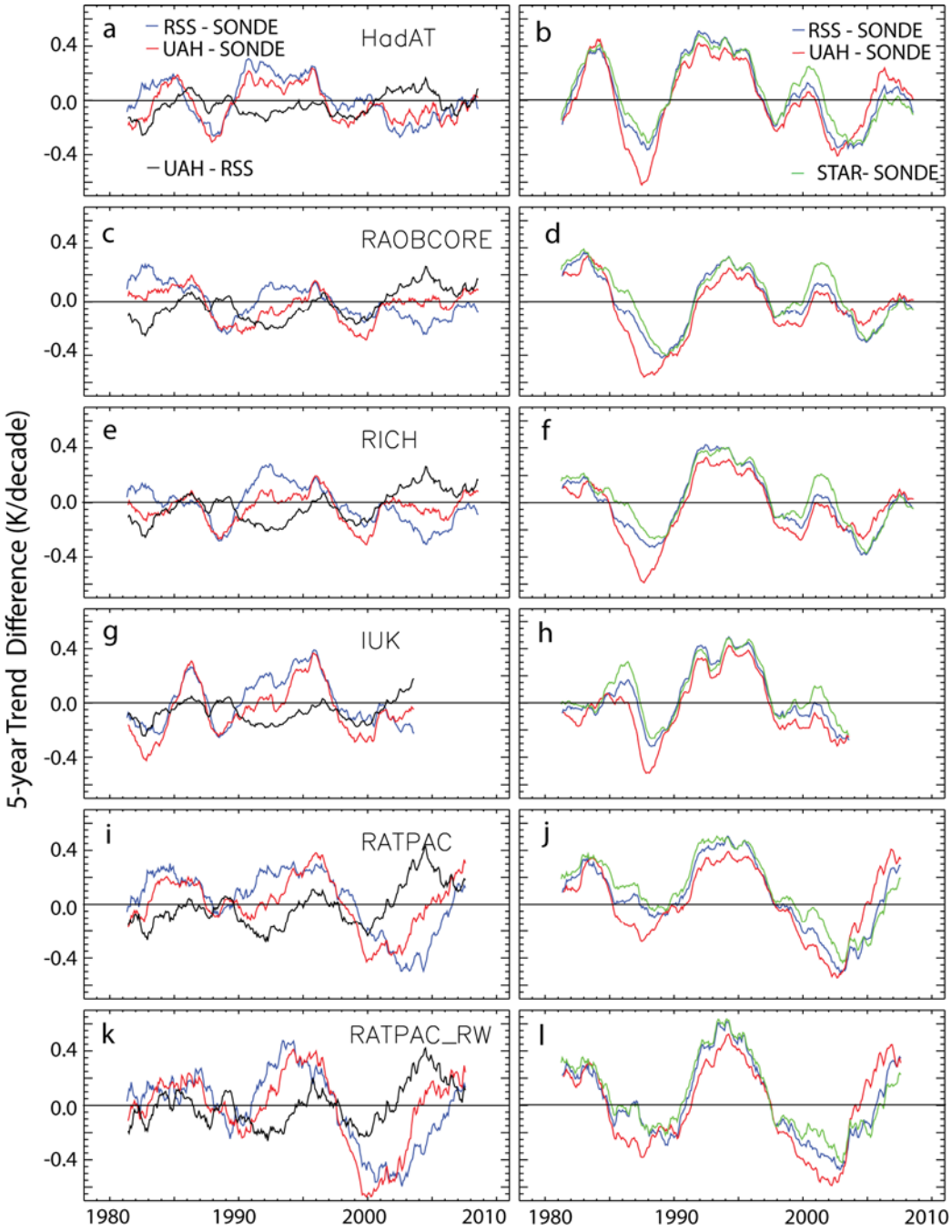


Fig.4. Same as Fig.3, except that the data are averaged over radiosonde locations in the deep tropics (20S-20N). Similar sets of plots for the southern and northern extratropics are available in the supplemental material (Figs. S6 and S7.)

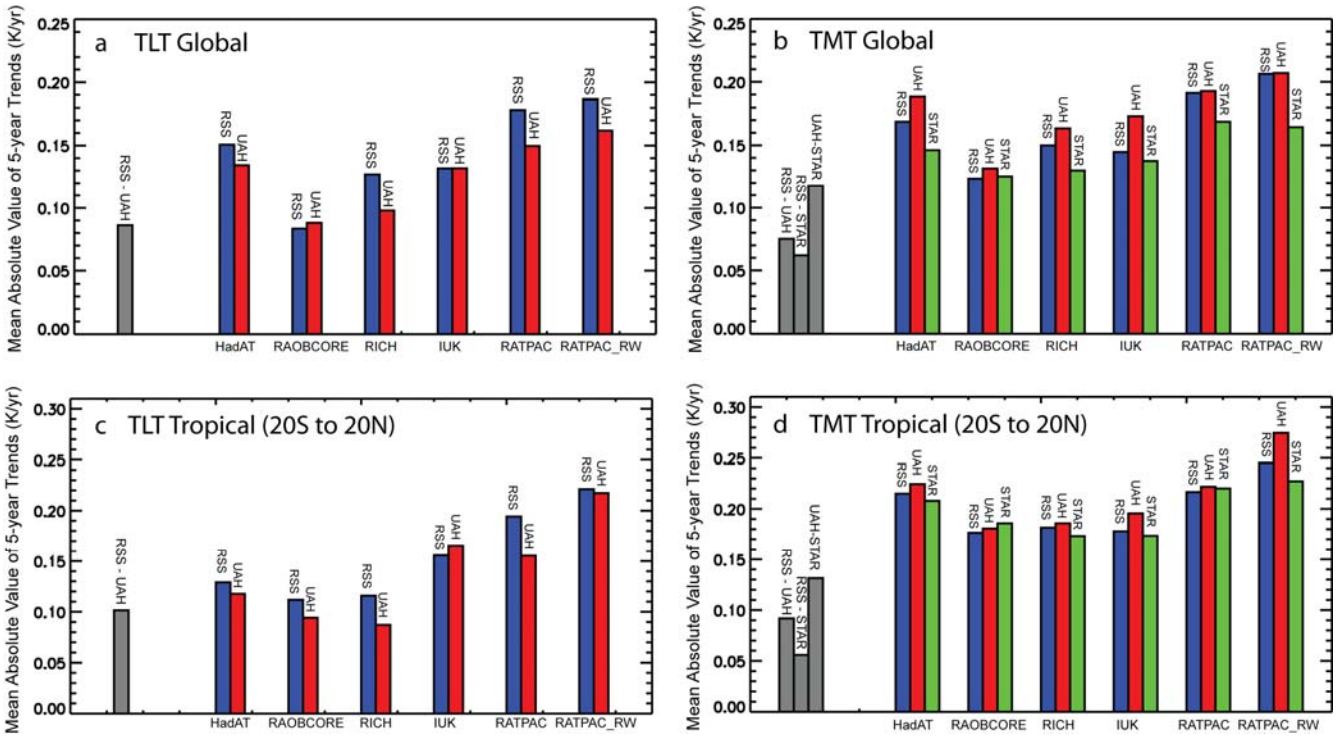


Figure 5. Summary of the differences between 5-year trends using the mean absolute difference as a metric. 5a. TLT, global. 5b. TMT, global. 5c. TLT, deep tropics (20S-20N). 5d. TMT, deep tropics. In most cases, the difference between different satellite datasets is substantially less than the difference between the satellite datasets and the radiosonde datasets. The exception to this is for the RAOBCORE and RICH datasets for TLT.

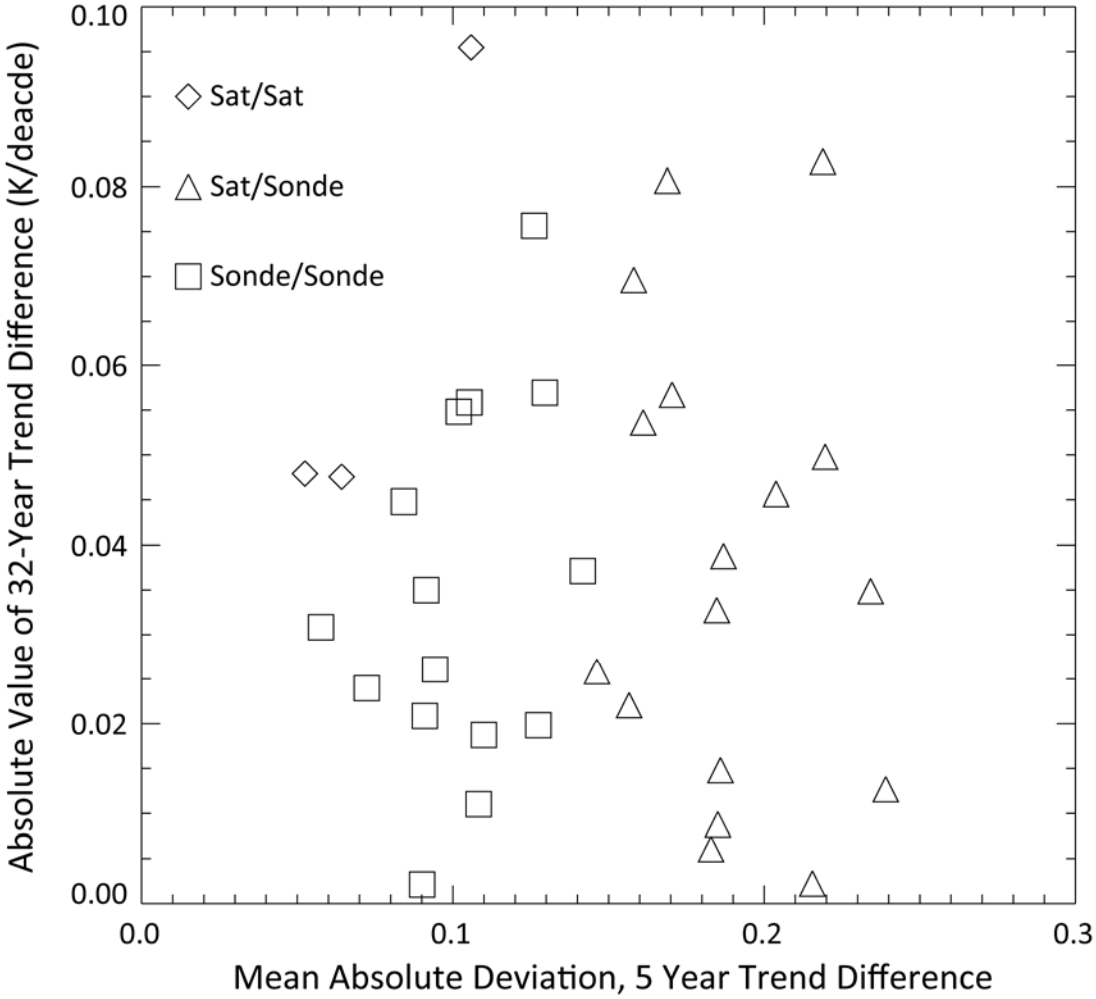


Figure 6. Scatter plot of the absolute value of the 32-year trend difference (1979-2010) in global means of TMT as a function of mean absolute deviation of the 5-year trend difference time series computed from the same difference time series. Each point represents the comparison of one dataset (satellite or radiosonde) with another (satellite or radiosonde).

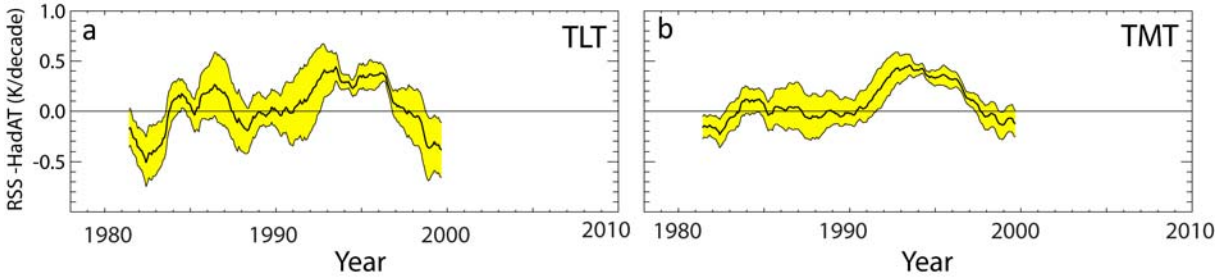


Fig. 7. Differences between 5 year trends with uncertainty estimates for channels TLT and TMT. All time series and uncertainty estimates are global averages sampled at the location of HadAT radiosonde measurements (which end in 2003 in this ensemble and hence the series truncation). The thick black line is the median difference time series, and the yellow region surrounding it contains 95% of the instances in the uncertainty ensemble. The uncertainty ensemble includes the estimated uncertainty in both the RSS and HadAT datasets.

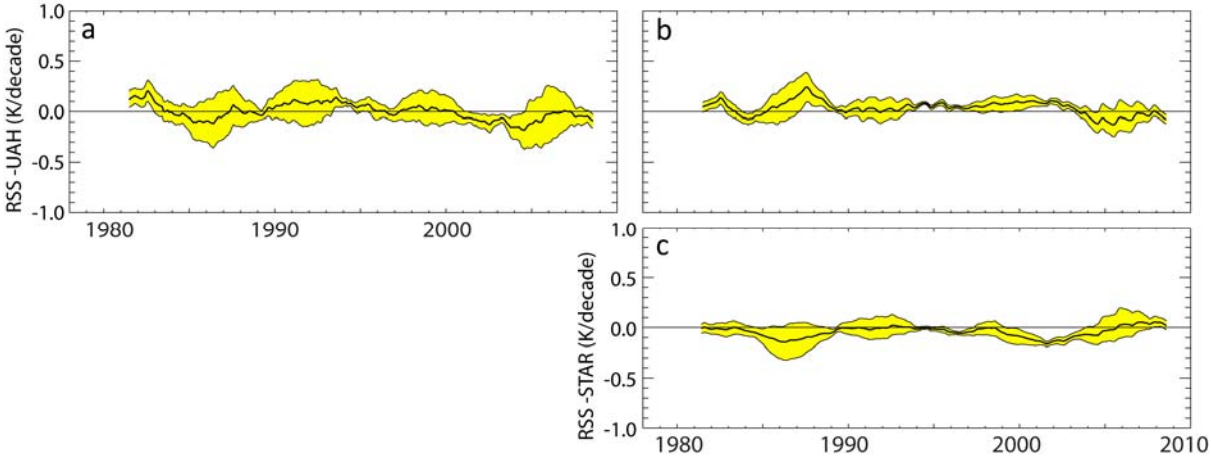


Fig. 8. Differences between 5 year MSU trends with RSS-only uncertainty estimates for channels TLT and TMT. a) RSS-UAH for TLT. b) RSS-UAH for TMT, c) RSS-STAR for TMT. In all panels, the thick black line is the median difference time series, and the yellow region surrounding it contains 95% of the instances in the RSS uncertainty ensemble.

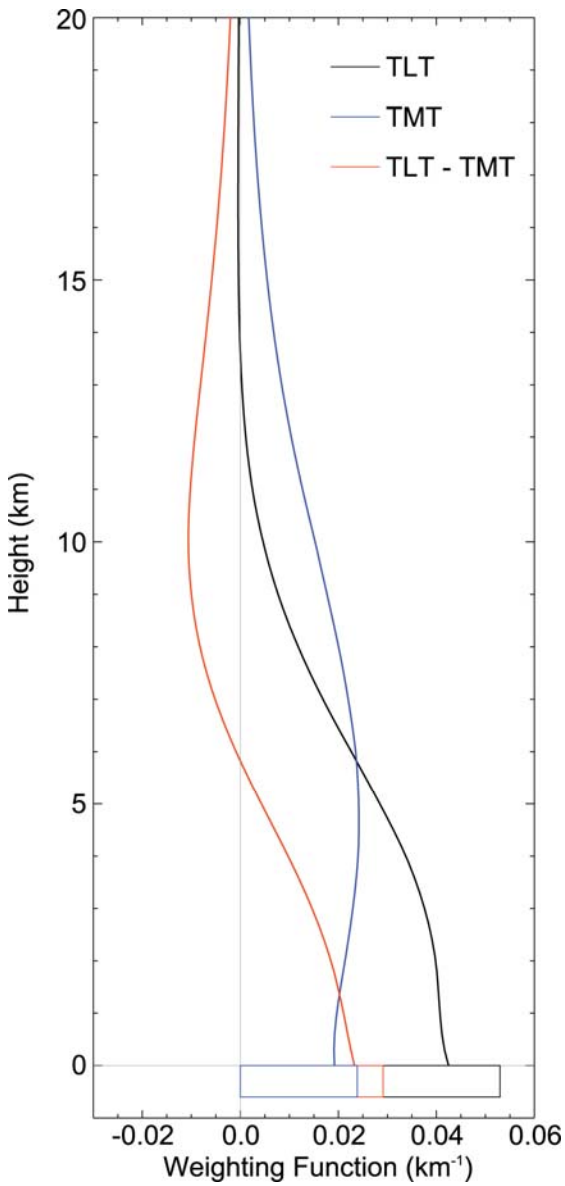


Fig. 9. Weighting functions as a function of height for TLT, TMT, and the TLT minus TMT difference for ocean scenes. The rectangular areas located below zero height show the portion of the total weight due to surface emission. For land scenes, this contribution would be about two times larger, with a corresponding reduction in atmospheric weight. Because of the large absolute weight in the stratosphere and the surface, the TLT minus TMT difference should not be interpreted as the difference between the lower and middle tropospheric temperatures.

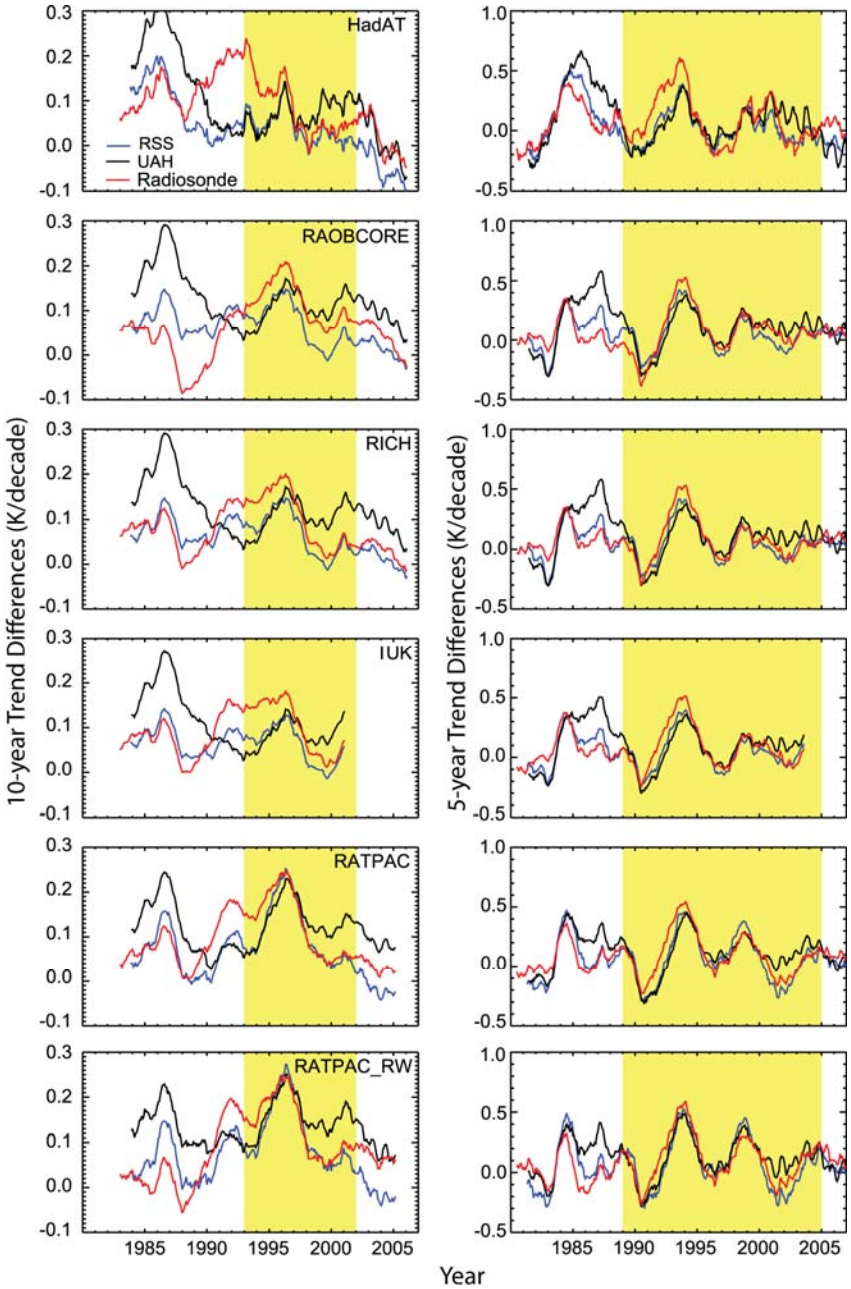


Fig. 10. Differences (TLT minus TMT) in short term trends in atmospheric temperature between adjacent layers for each radiosonde dataset. In each case, the satellite data is sampled at the radiosonde locations, and the differences in spatial sampling lead to the differences between the satellite results for the different panels. The temporal range plotted in RH2008 is shaded in yellow.

Supplementary Material

Additional details on the creation of the RSS and HadAT ensembles

We have recently calculated a set of 400 monte-carlo realizations of possible error for each channel of the RSS dataset [Mears et al., 2011]. The resulting error datasets are available at the same spatial and temporal resolution as the base dataset. This makes it possible to construct 400 realizations of this dataset that are consistent with the estimated uncertainty by adding them to the baseline, satellite derived temperature dataset. The error model is a full end-to-end reprocessing which permits interdependencies in uncertainties between successive steps of the processing chain to be fully expressed. The sources of input error included two sources of error. The first source is the sampling error caused by incomplete temporal sampling of any given gridpoint over a month, caused by gaps between the swaths measured by the satellite on a given day. This error was estimated by sampling the output from a climate model (CCM3) using the actual sampling times from a MSU or AMSU satellite, and comparing these results to the true monthly average. The second source of error is due to a possible error in the model-based adjustments made for changes in local measurement time. These were estimated by evaluating the range of results that arise when different linear combinations of the measurement time adjustments calculated from different, structurally distinct, climate-model based estimates of the diurnal cycle were used. These errors, combined with finite sample sizes, lead to uncertainty in the merging parameters (offsets and calibration adjustments that depend on the temperature of the warm calibration target) used to adjust the various satellites before they are merged together. Sampling errors were found to dominate at small distance and time scales but rapidly diminish for large area averages or longer term changes. On long timescales diurnal cycle correction offsets (particularly for TLT) and inter-satellite offset adjustments were the dominant sources of

uncertainty.

The HadAT ensemble used herein are derived from perturbed versions of the Hadley Centre's automated neighbor based homogenization procedure. The method involves creation of neighbor composites for each station. Breakpoints for each station are indentified using knowledge of metadata events (known to be grossly incomplete) and a statistical test performed on the station-minus-neighbor time series. At these breakpoints, adjustments are applied based on the target-minus-neighbor series. The procedure is iterative with early iterations designed to ascertain the most obvious breaks and the breakpoint detection threshold being relaxed with each iteration. It has no convergence requirement. For full methodology details see Thorne et al [2005] and McCarthy et al. [2008].

The algorithm was originally run on seasonal resolution data with substantial human intervention to confirm breakpoint locations and adjustments to create HadAT [Thorne et al., 2005]. It was then subsequently automated and all obvious subjective methodological choices were made tunable parameters to create an ensemble of potential solutions [McCarthy et al., 2008]. To assess the veracity of the approach, a set of analogs to the real world data with known error characteristics were created. This was done by sub-setting output from an AGCM (forced with observed SSTs and historical 20th Century forcing agents) to the observational datamask and adding four distinct error structures to approximate plausible real-world errors [Titchner et al., 2009]. Then, a set of more substantial methodological perturbations including substantially different input data resolution, neighbor calculation approaches and a whole new adjustment approach were considered [Thorne et al., 2011]. As a final step this analysis recombined a small number of the resulting ensembles under a conditional probability paradigm to yield estimates of trend changes over the longest timescales.

This yields a very substantial set of information which could be used herein. In deciding which data to use we were guided by the need to consider data at the grid box level and for relatively short time periods. The most consistent ensemble in recapturing, in an unbiased manner, the 'truth' in the analogs for several regions and trend periods was the twenty-member monthly resolution ensemble [Thorne et al., 2011]. In particular, during the satellite era tropical trends in the analog worlds it performed very well (see Figure 6 in Thorne et al. [2011]). Although this is far from an absolute guarantee that the ensemble will be unbiased it likely constitutes the best approximation to an unbiased estimator of the true climate evolution from the entire available population of HadAT automated approaches. It also very likely will be a better estimator than the original manually produced HadAT dataset, at least in terms of giving a realistic spread to reflect the real degree of ignorance at smaller time and space scales into the real evolution of the radiosonde record.

Additional Versions of Figure 2 for other radiosonde datasets.

Figures S1 through S5 show versions of Fig. 2 calculated using the RAOBCORE, RICH, IUK, RATPAC, and RATPAC_RW datasets.

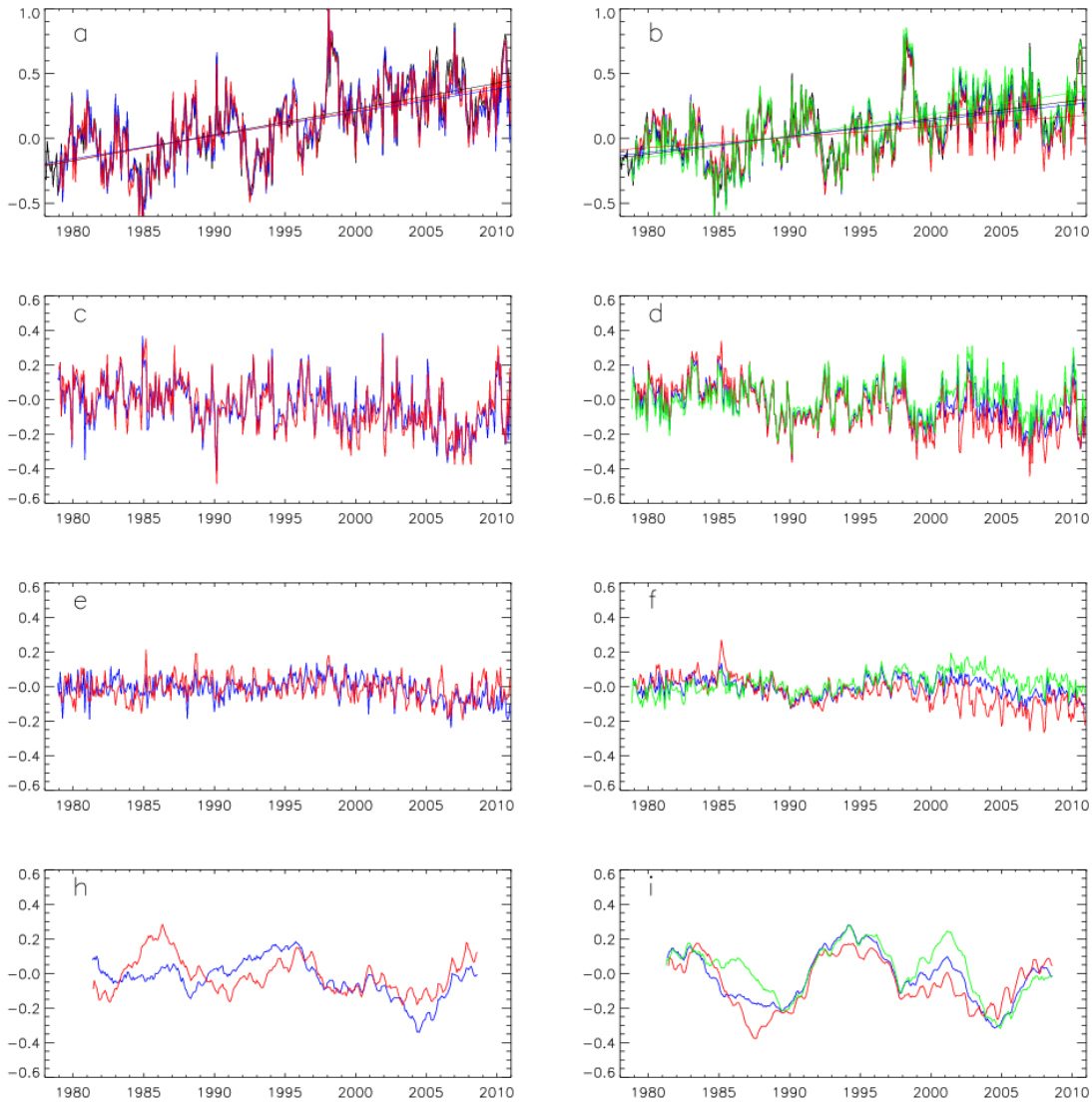


Figure S1. Same as Fig. 2 from the text, calculated using the RAOBCORE dataset instead of the HadAT dataset.

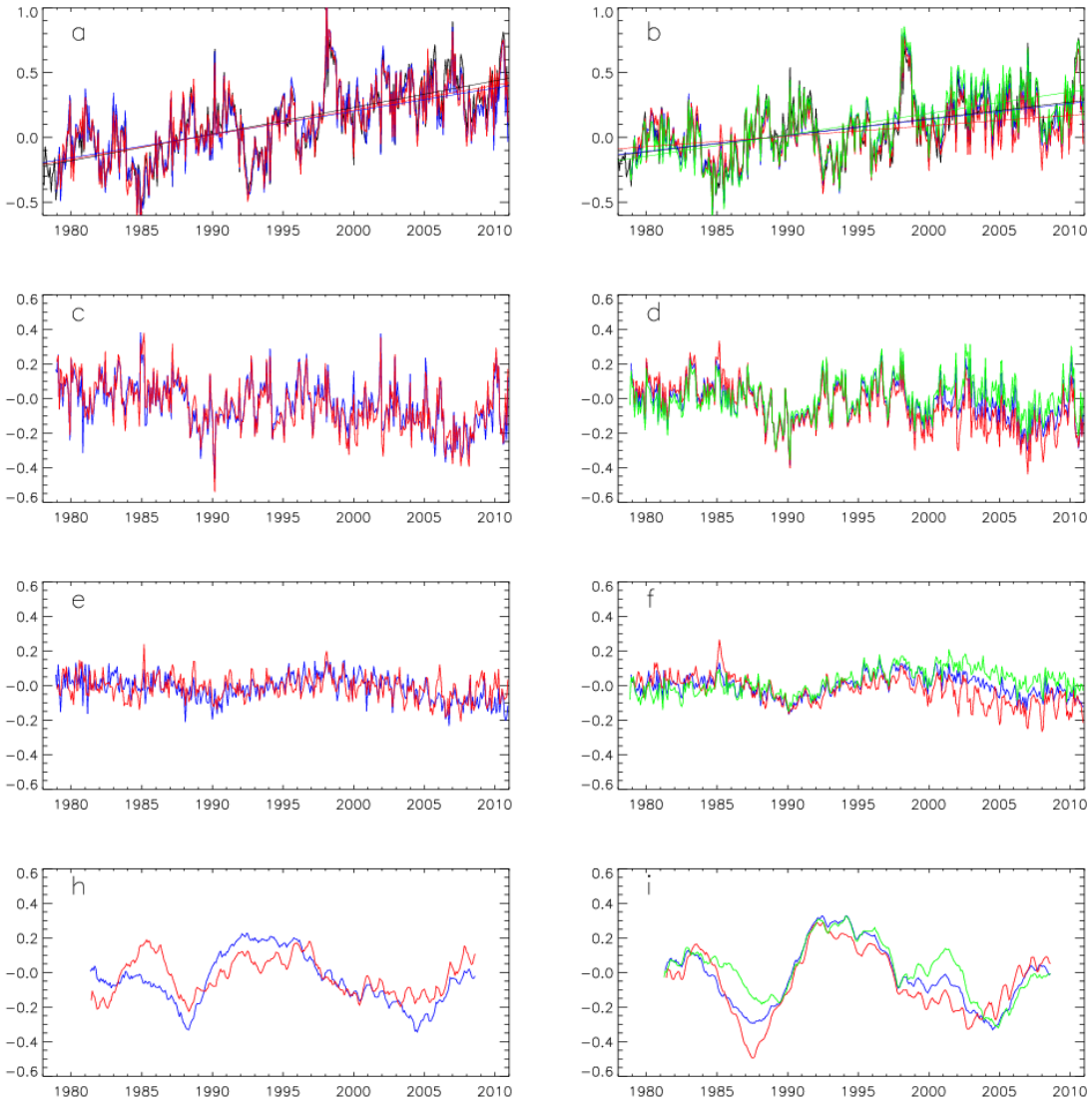


Figure S2. Same as Fig. 2 from the text, calculated using the RICH dataset instead of the HadAT dataset.

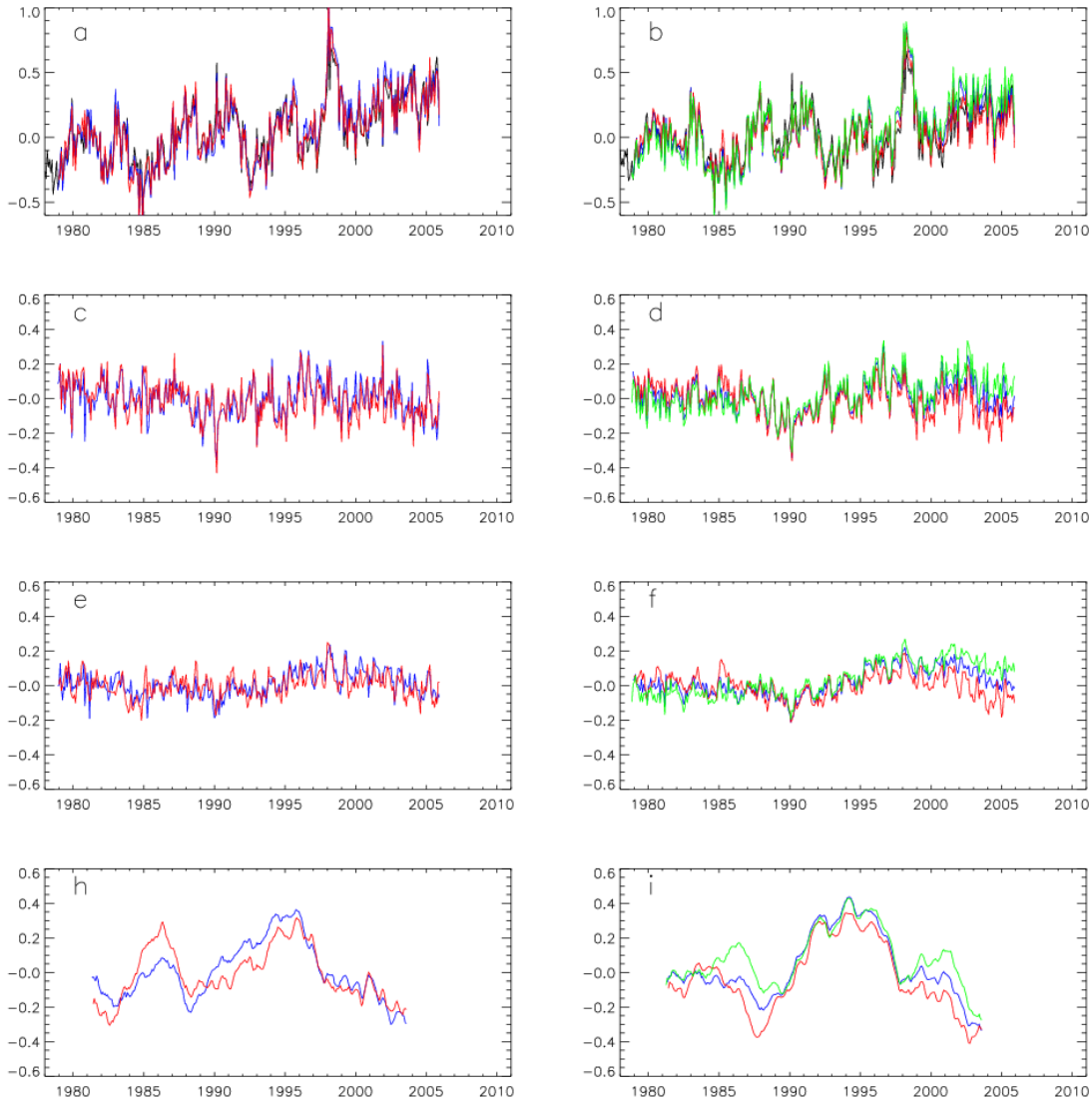


Figure S3. Same as Fig. 2 from the text, calculated using the IUK dataset instead of the HadAT dataset.

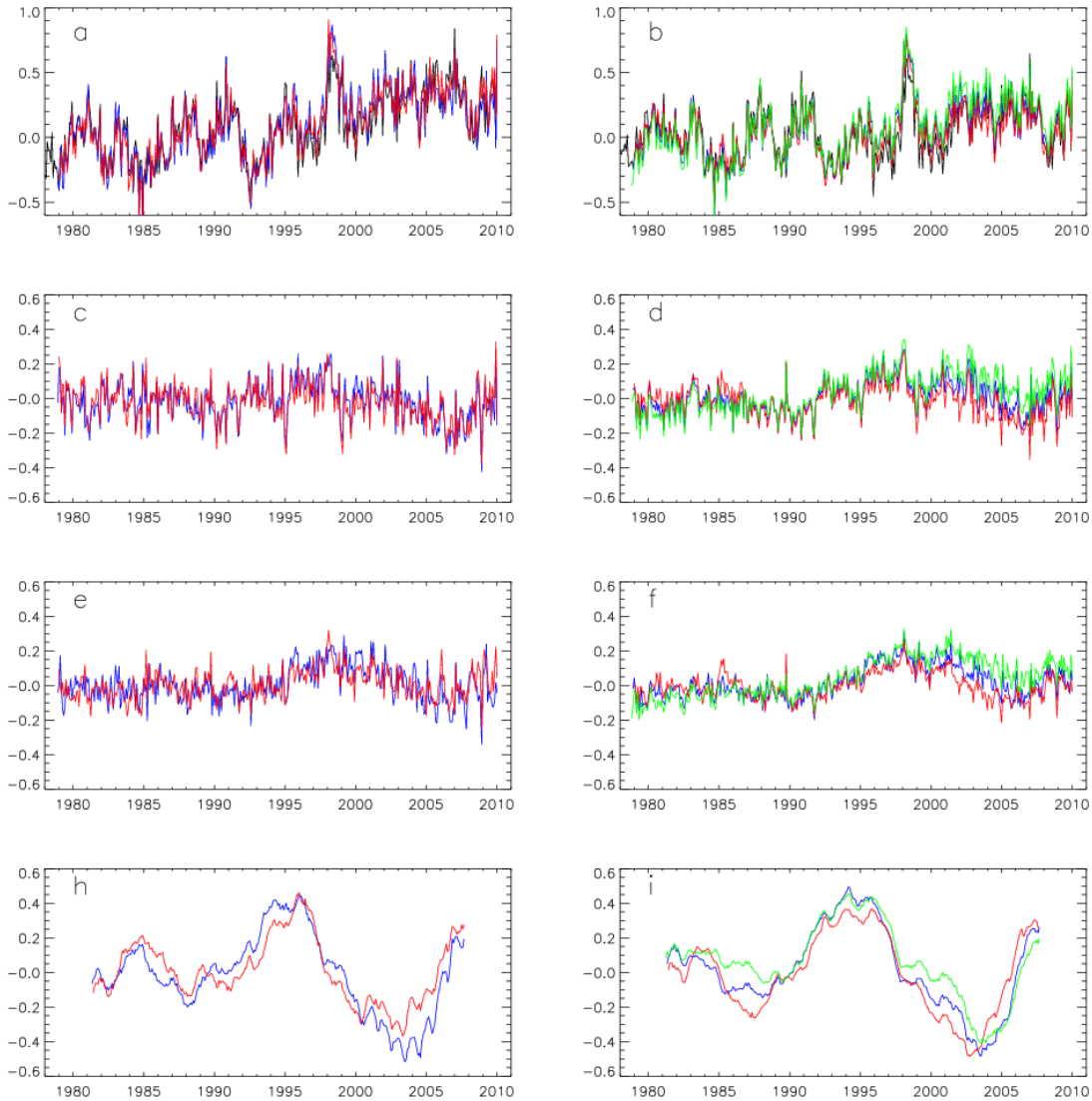


Figure S4. Same as Fig. 2 from the text, calculated using the RATPAC dataset instead of the HadAT dataset.

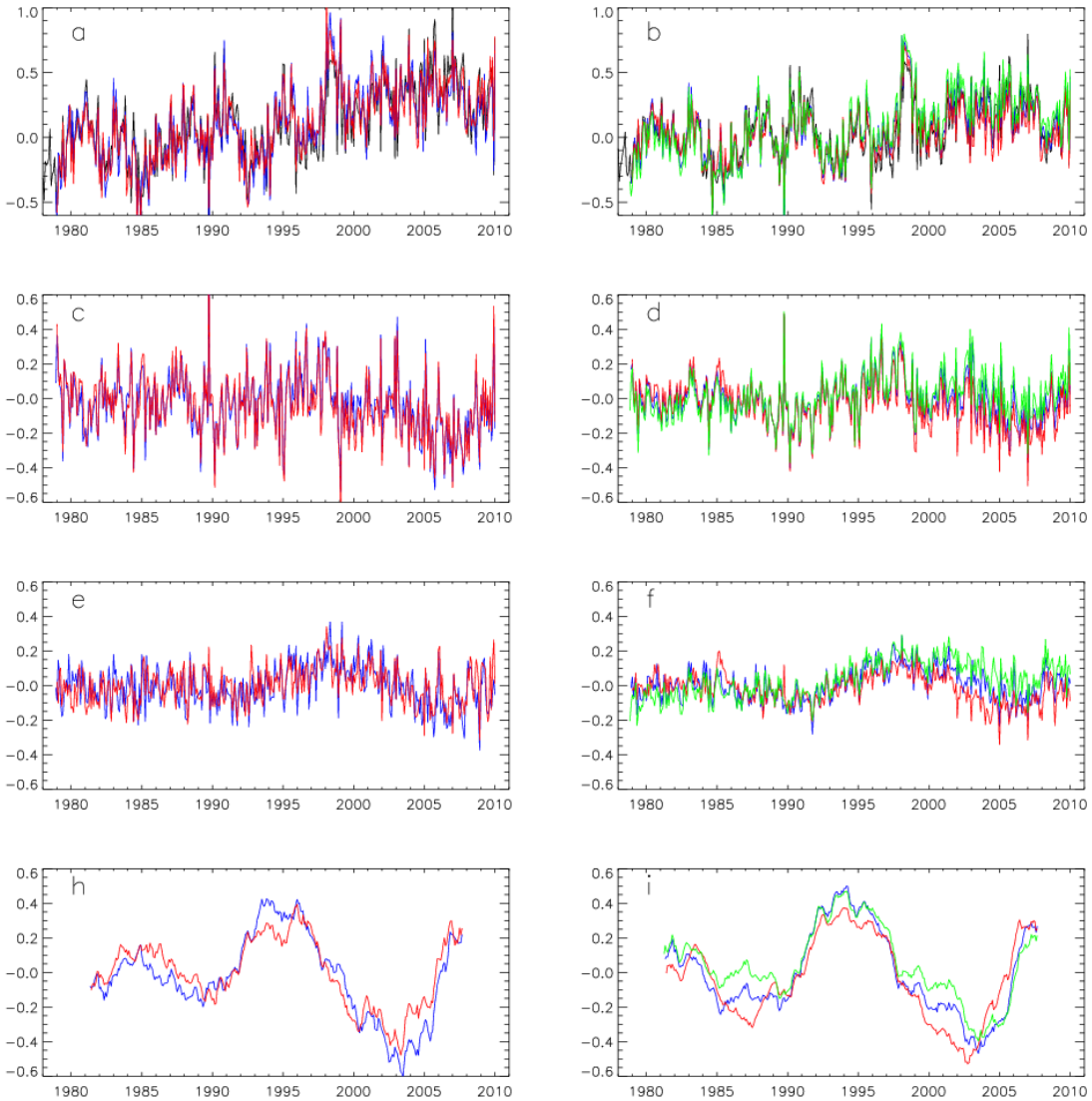


Figure S5. Same as Fig. 2 from the text, calculated using the RATPAC_RW dataset instead of the HadAT dataset.

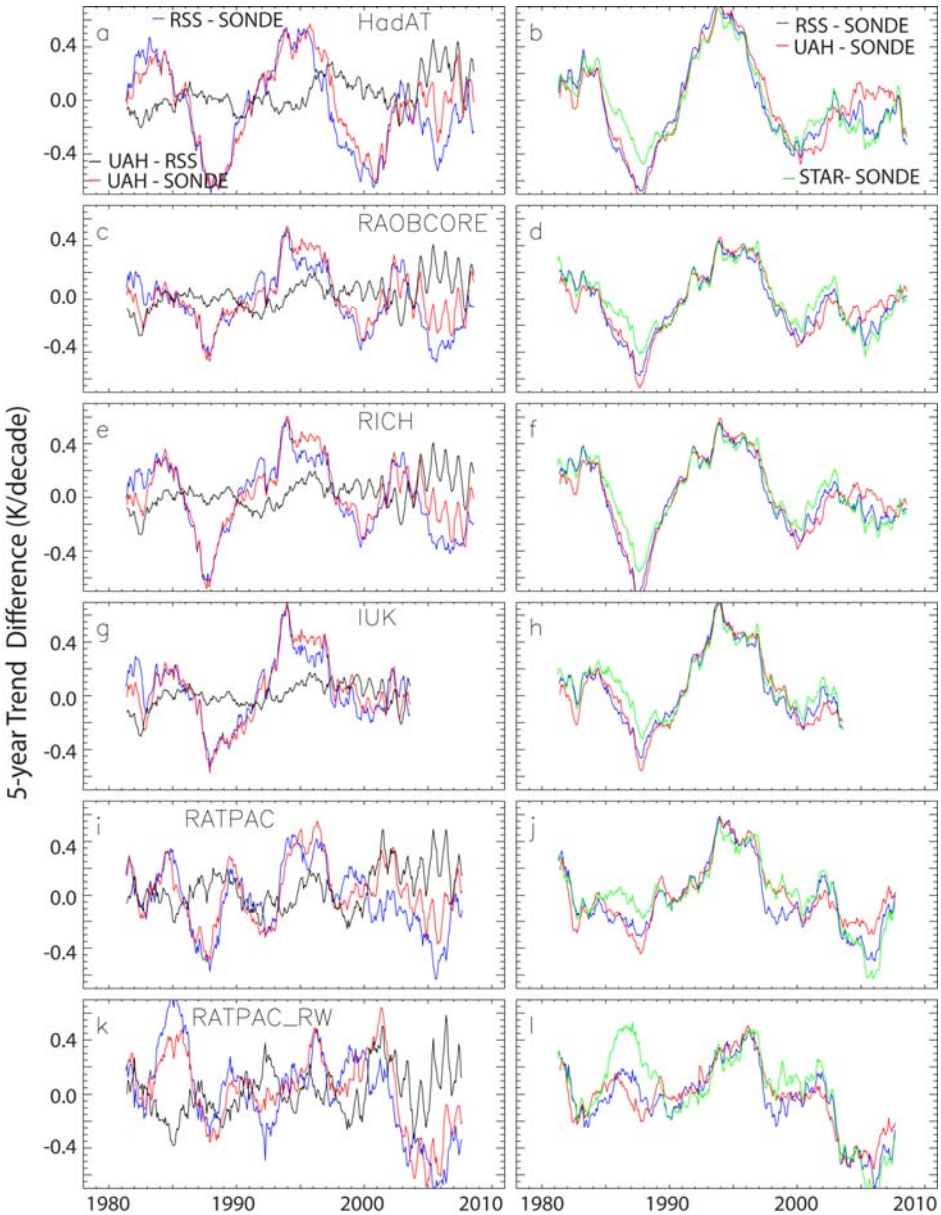


Fig. S6. Same as Fig.4 from the main text, except for the Southern Extratropics (82.5 S to 20S for TMT, and 70S to 20S for TLT). This spatial subset shows features that are quite different from the other regions, especially for the HadAT, RAOBCORE and RICH datasets. These differences may arise because of the small number of radiosonde stations in the southern hemisphere.

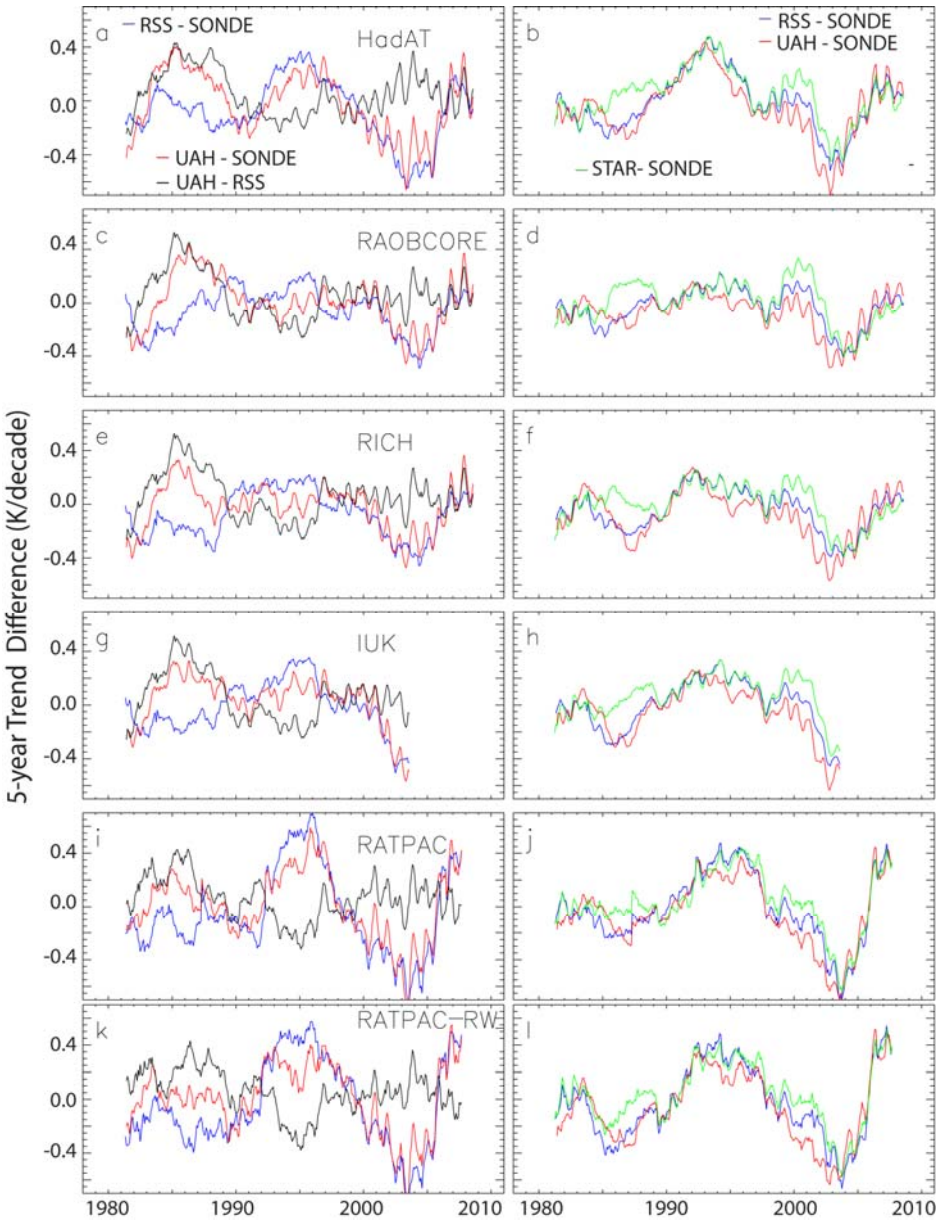


Fig. S7. Same as Fig. 4 from the main text, except for the Northern Extratropics (20N – 82.5N). The general features, such as the locations of peaks, are similar.

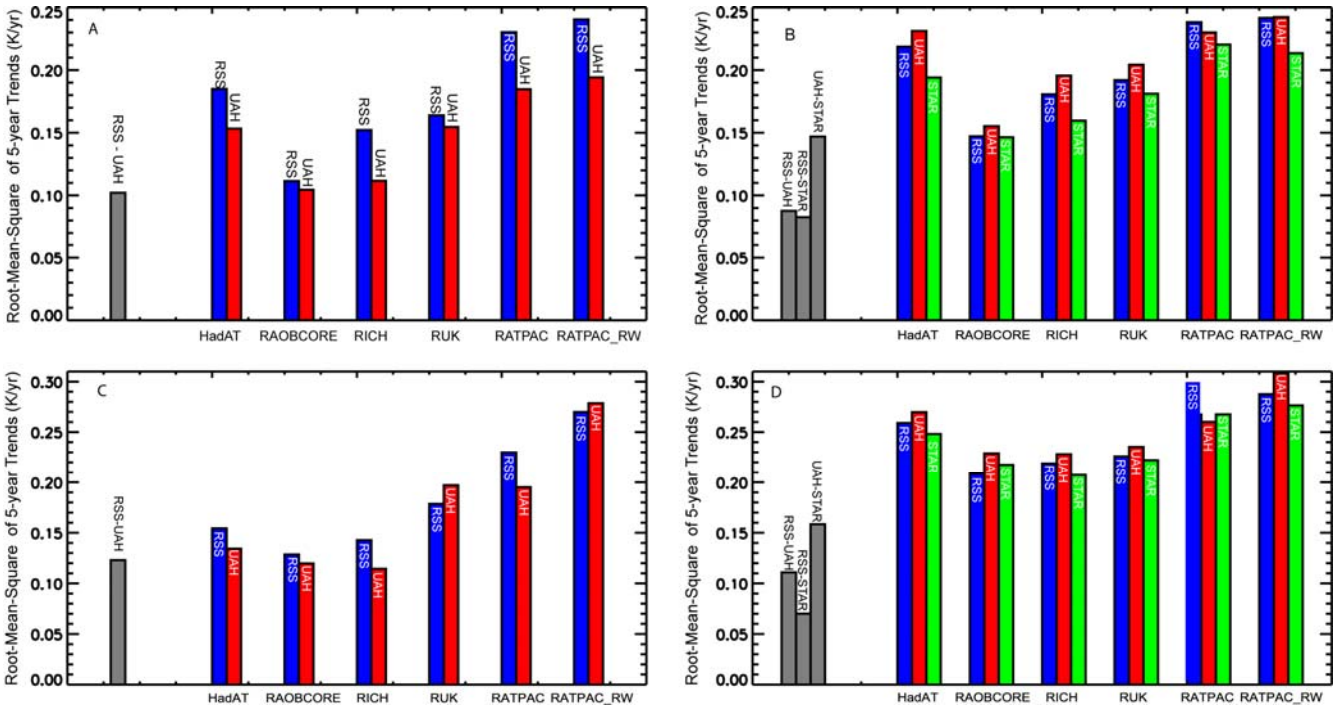


Fig. S8. Same as Fig. 5 in the main text, except calculated using the root-mean-square of the trends in the difference time series instead of the mean absolute value. The relative magnitudes of the trend differences are unchanged, showing the insensitivity of the results to metric chosen.

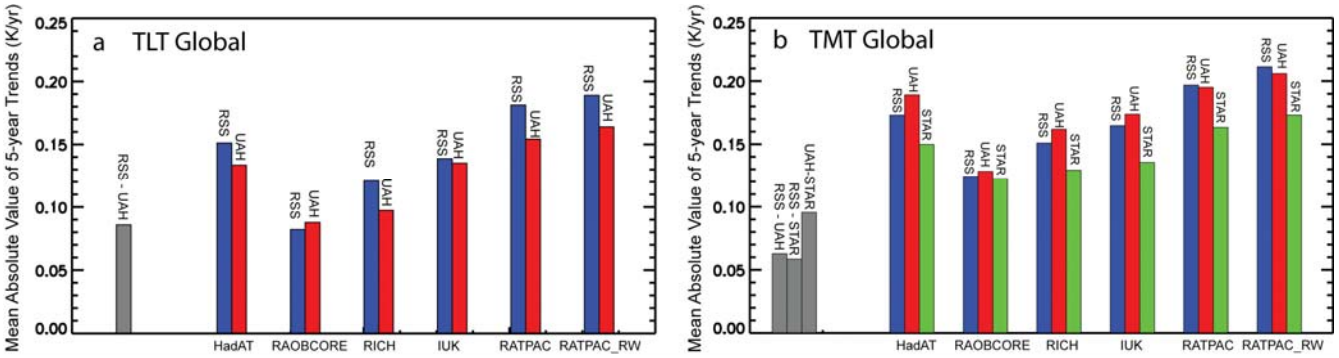


Fig. S9. Same as panels a and b of Fig. 5 in the main text, except that the overall linear trend was removed from each difference time series before the analysis was performed. The results

are essentially unchanged, showing that the 5-year trend differences are dominated by short-term behavior.

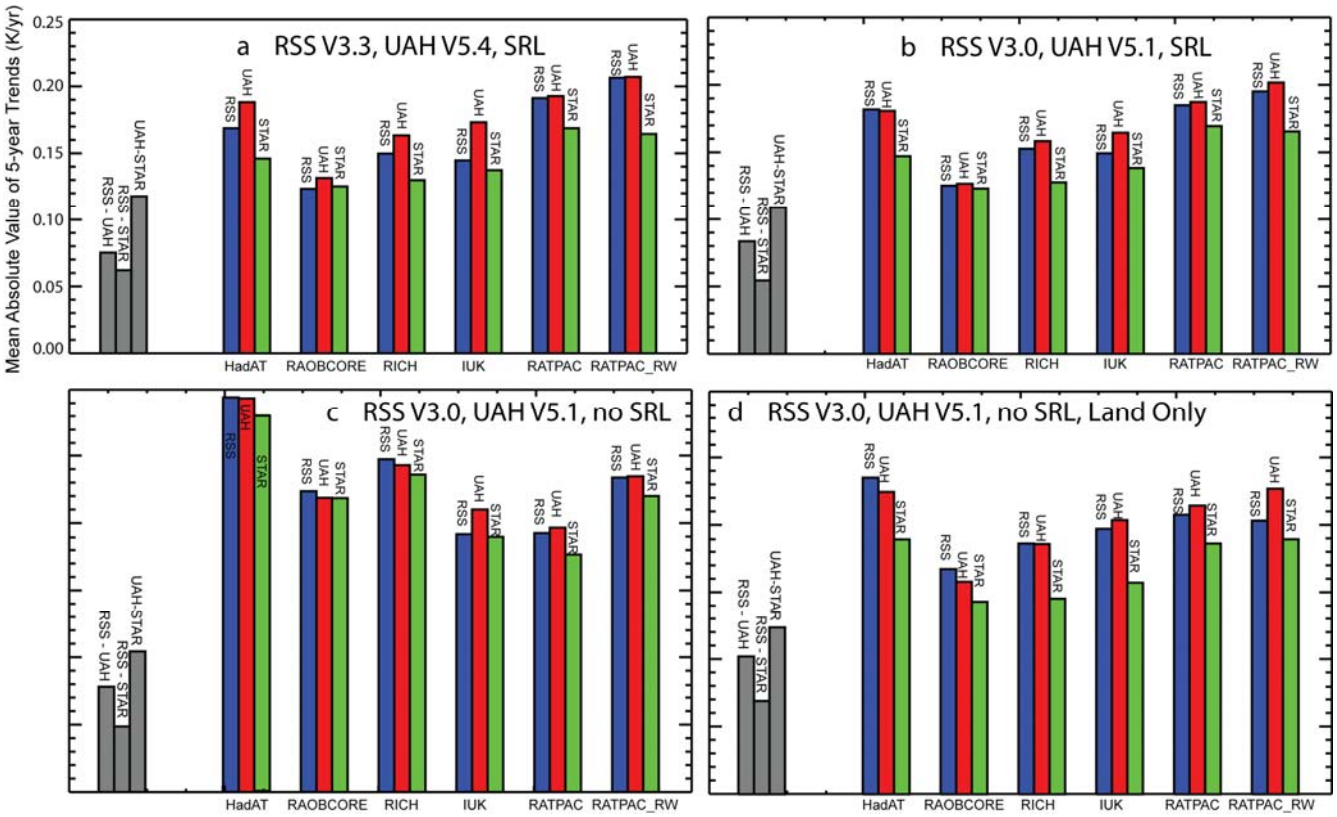


Fig. S10. This figure is multiple versions of panel b of Fig 5 in the main text, which shows the mean absolute difference between 5 year trend time series. a) Reproduced from Fig5b, with SRL sampling and updated satellite datasets. For the updated MSU data, the period of analysis is 1979-2011.

b) Same as panel a, except calculated using previous versions of the RSS (V3.0) and UAH (V5.1) datasets. The STAR dataset is the same. For this plot, the period of analysis is 1979-2007, so any changes in the STAR results are due to the change in analysis period. c) Same as panel b, except that area-weighted averaging was used to construct the global averages. d).

Same as panel d, except than land-only area weighted averages were used for the MSU dataset. For RSS, the update from version 3.0 to 3.3 reduced the mean absolute difference values slightly. Using area-weighted averaging (no sampling at radiosonde locations) reduced the mean absolute difference between MSU datasets, and greatly increases the mean absolute difference between MSU and radiosonde datasets, with the exception of the RATPAC dataset, which was carefully constructed to represent a global average. Moving to a land-only area-weighted average reduces the mean absolute differences, but they are still substantially larger than those in panel a or b.

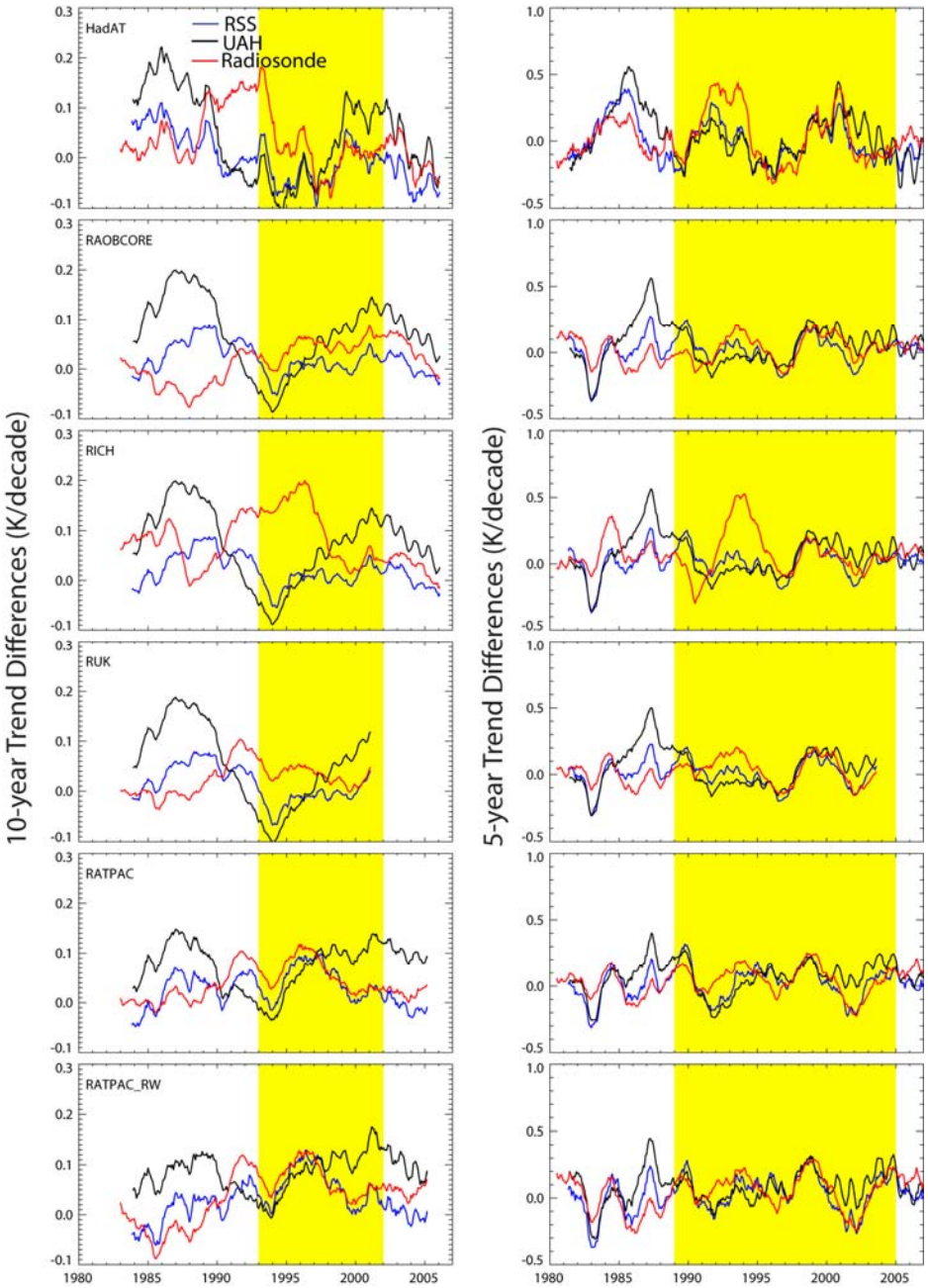


Fig. S11. Same as Figure 10 in the main text, except calculated using the TT (Total Troposphere) product defined in Fu and Johansan (2005) to be 1.1*TMT -0.1*TLS. This weighted combination of channels reduces the contribution of the stratosphere relative to TMT. The use of TT changes the time series considerably, and in general the variability of the time series is

reduced, probably due to the reduction of the influence of volcano-induced stratospheric warming. The overall conclusions drawn from Fig 10 are unaltered, with some of the largest discrepancies occurring outside the range considered by RH2008, and RSS being in better agreement with all radiosonde datasets.

Definition of radiosonde stations as land, ocean, or coastal.

In section 3, we argue that most of the stations considered by Randall and Wu are in either coastal or oceanic locations: To define ocean, coastal, and land we use a gridded land-fraction dataset on the same 2.5 x 2.5 degree spatial scale as the gridded MSU/AMSU data. If the MSU gridpoint that contains the station is less than 30% land, the station is classified as ocean. If not, then we examine a 3x3 subset of the MSU gridpoints that surround the station. If any of these cells have a land fraction that is less than 0.75, then the station is considered to be coastal. If not, then it is considered to be land. One station, ROSTOV-ON-DONU, lat = 47.250, lon = 39.820), which our algorithm evaluated to be coastal due to its location on the shores of the Black Sea, was set to land manually because it is surrounded on three sides by land.

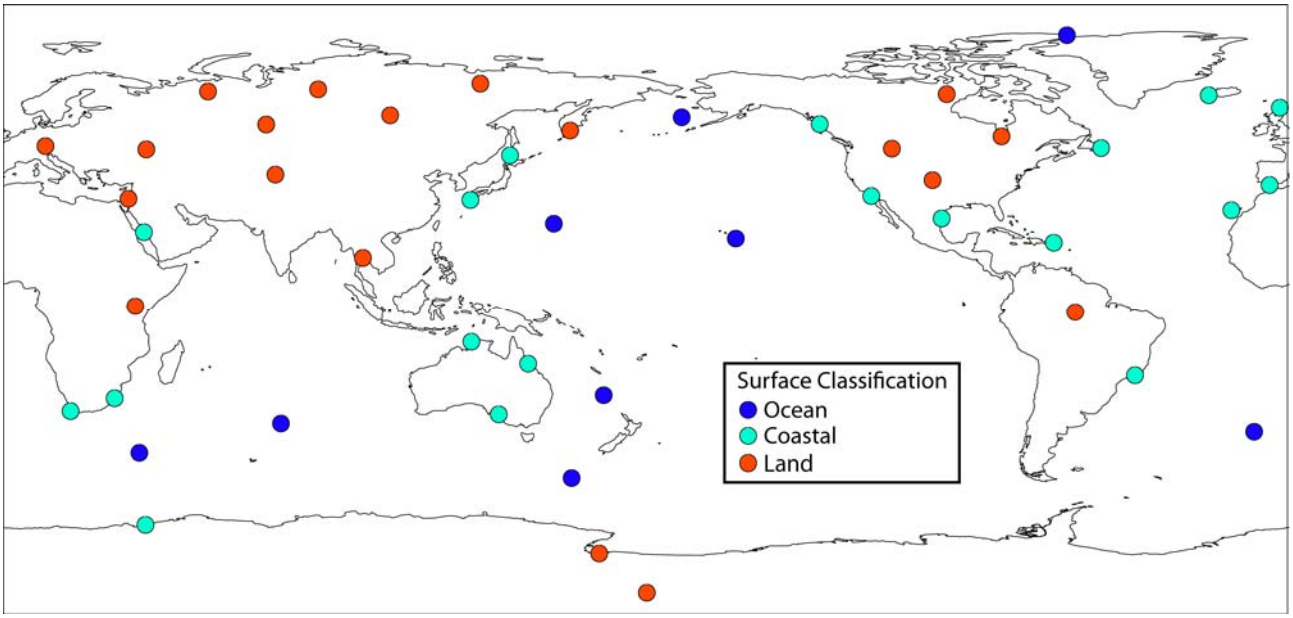


Fig. S11. Map of RATPAC -RW stations, color-coded by surface classification.

Table S1. Number of RATPAC-RW stations assigned to each surface type.

Location Type	Number of Stations
Open Ocean	9
Coastal	20
Dry Land	18

Thus we conclude that only 18 of 47 (less than 40%) the RATPAC-RW stations are truly on dry land locations, making the choice of using land-only MSU/AMSU data for comparison purposes a questionable choice. Coastal locations typically do not exhibit true dry land climatology. For example, these coastal locations show significantly less diurnal variability than dry land locations. For TMT, the amplitude of the diurnal cycle (Mears et al., 2009a, 2009b) used to adjust the MSU/AMSU data is more than 55% larger for an all-land average than for an average over the RW locations, and the part consisting of only even harmonics (the part that dominates

the diurnal adjustment of MSU/AMSU data) is more than 65% larger.

References

Fu, Q., and C. M. Johanson (2005), Satellite-derived vertical dependence of tropical tropospheric temperature trends, *Geophys. Res. Lett.*, 32, L10703, doi:10.1029/2004GL022266.

Mears, C. A., and F. J. Wentz (2009a), Construction of the RSS V3.2 lower tropospheric dataset from the MSU and AMSU microwave sounders, *Journal of Atmospheric and Oceanic Technology*, *26*, 1493-1509.

Mears, C. A., and F. J. Wentz (2009b), Construction of the Remote Sensing Systems V3.2 atmospheric temperature records from the MSU and AMSU microwave sounders, *Journal of Atmospheric and Oceanic Technology*, *26*, 1040-1056.

Mears, C. A., F. J. Wentz, P. W. Thorne, and D. J. Bernie (2011), Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique, *Journal of Geophysical Research*, *116*, D08112, doi:10.1029/2010JD014954.

McCarthy, Mark P., H. A. Titchner, P. W. Thorne, S. F. B. Tett, L. Haimberger, D. E. Parker, 2008: Assessing Bias and Uncertainty in the HadAT-Adjusted Radiosonde Climate Record. *J. Climate*, **21**, 817–832.

Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan (2005b), Revisiting radiosonde upper-air temperatures from 1958 to 2002, *Journal of Geophysical Research*, *110*, doi:10.1029/2004JD005753.

Titchner, H. A., P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, and D. E. Parker (2009), Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments, *Journal of Climate*, *22*(3), 465-485.