A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes

Peter W. Thorne,[1,2] Philip Brohan,[1] Holly A. Titchner,[1] Mark P. McCarthy,[1] Steve C. Sherwood,[3] Thomas C. Peterson,[4] Leopold Haimberger,[5] David E. Parker,[1] Simon F. B. Tett,[6] Benjamin D. Santer,[7] David R. Fereday,[1] and John J. Kennedy[1]

Received 10 December 2010; revised 18 March 2011; accepted 29 March 2011; published 29 June 2011.

[1] The consistency of tropical tropospheric temperature trends with climate model expectations remains contentious. A key limitation is that the uncertainties in observations from radiosondes are both substantial and poorly constrained. We present a thorough uncertainty analysis of radiosonde-based temperature records. This uses an automated homogenization procedure and a previously developed set of complex error models where the answer is known a priori. We perform a number of homogenization experiments in which error models are used to provide uncertainty estimates of real-world trends. These estimates are relatively insensitive to a variety of processing choices. Over 1979–2003, the satellite-equivalent tropical lower tropospheric temperature trend has likely (5–95% confidence range) been between −0.01 K/decade and 0.19 K/decade (0.05–0.23 K/decade over 1958–2003) with a best estimate of 0.08 K/decade (0.14 K/decade). This range includes both available satellite data sets and estimates from models (based upon scaling their tropical amplification behavior by observed surface trends). On an individual pressure level basis, agreement between models, theory, and observations within the troposphere is uncertain over 1979 to 2003 and nonexistent above 300 hPa. Analysis of 1958–2003, however, shows consistent model-data agreement in tropical lapse rate trends at all levels up to the tropical tropopause, so the disagreement in the more recent period is not necessarily evidence of a general problem in simulating long-term global warming. Other possible reasons for the discrepancy since 1979 are: observational errors beyond those accounted for here, end-point effects, inadequate decadal variability in model lapse rates, or neglected climate forcings.

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

1. Introduction

[2] Over the past twenty years the vexatious issue of whether the troposphere is warming or not and, if it is, then whether it is warming at a rate consistent with climate model expectations, has spawned more than 200 research papers, two dedicated expert panel reviews [*National Research Council Panel on Reconciling Temperature Observations*, 2000; *Karl et al.*, 2006], and has been a focus of reports by Working Group I of the IPCC [*Thorne et al.*, 2011]. Over time, attention has shifted from the global mean to changes in the deep tropics, which are dominated by convective processes and where climate model behavior is strongly constrained [*Santer et al.*, 2005]. Here, any change in temperature at the surface is amplified aloft. The physical reasons for amplification are well understood. On month-to-month and year-to-year time scales, all climate models and observational estimates exhibit remarkable agreement with each other and with simple theoretical expectations. But on multidecadal time scales, many observational estimates of amplification behavior depart from basic theory, while climate models do not. The most recent major assessment [*Karl et al.*, 2006, p. 2] concluded that such discrepancies "may arise from errors that are common to all models, from errors in the observational data sets, or from a combination of these factors. The second explanation is favored, but the issue is still open." In this paper, we explore observational error.

[3] With the notable exception of the Keeling curve of CO_2 concentration changes [*Keeling et al.*, 1976], to date

[1]Met Office Hadley Centre, Exeter, UK.
[2]Cooperative Institute for Climate and Satellites, North Carolina State University and NOAA National Climatic Data Center, Asheville, North Carolina, USA.
[3]Climate Change Research Centre, University of New South Wales, Sydney, New South Wales, Australia.
[4]NOAA National Climatic Data Center, Asheville, North Carolina, USA.
[5]Department for Meteorology and Geophysics, University of Vienna, Vienna, Austria.
[6]School of Geosciences, University of Edinburgh, Edinburgh, UK.
[7]Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, California, USA.

there exists no climate record that is definitively tied to SI standards. Such records require comprehensive metadata, traceability at every step to absolute (SI) standards, and a careful and comprehensive calculation of error budgets [*Immler et al.*, 2010]. They are expensive, time consuming to produce, and difficult to construct and maintain. It is therefore understandable that virtually all of the historical meteorological data available to the community fail, usually substantially, to measure up to such exacting standards. As a result, there will always be uncertainty in establishing how the climate system has evolved, notwithstanding careful attempts to identify and adjust for all apparent nonclimatic artifacts. Despite some claims to the contrary, no single approach is likely to encapsulate all of the myriad uncertainties in the data set construction process. The issue is most critical for multidecadal trends, since residual errors act as red noise, projecting most strongly onto the longest timescales [*Seidel et al.*, 2004; *Thorne et al.*, 2005b].

[4] Upper-air monitoring involves single-use radiosondes (weather balloons) and, more recently, satellites with individual lifetimes of a few years. Radiosonde technologies, and to a lesser extent practices, have changed markedly over the years. As a result, historical radiosonde temperature trends remain uncertain despite substantial efforts by a number of independent groups to address the issue [*Sherwood et al.*, 2008; *Haimberger et al.*, 2008; *Free et al.*, 2005; *Thorne et al.*, 2005a]. Satellite temperature records are similarly uncertain and only represent very broad vertical integrals, which leads to potential issues in interpretation [*Christy et al.*, 2003; *Mears and Wentz*, 2009a, 2009b; *Zou et al.*, 2006; *Vinnikov et al.*, 2006; *Fu et al.*, 2004]. It is essential to have a comprehensive understanding of the structural uncertainty, uncertainty arising from uncertain methodological choices, in the observations [*Thorne et al.*, 2005b]. Although the recent expansion in the number of upper-air temperature data sets has undoubtedly improved our ability to *quantify* structural uncertainty, there still exist only a handful of published data sets, but the number of scientifically plausible data sets is undoubtedly very much larger than this.

[5] Automation of the HadAT radiosonde data set construction procedure [*Thorne et al.*, 2005a] has enabled the creation of an ensemble of HadAT data sets. This has allowed analysts to explore the structural uncertainty (in this particular radiosonde product) arising from a wide variety of subjective choices made in data set construction [*McCarthy et al.*, 2008] (hereinafter M08). In order to evaluate this ensemble, *Titchner et al.* [2009] (hereinafter T09) generated a set of benchmark error models that share many of the complex features of real world temperature monitoring systems, such as spatiotemporal changes in sampling, random errors, and systematic errors. The results from the error models were used to infer the likely relationship of the HadAT ensemble to real world changes (T09). Both M08 and T09 found that the automated HadAT system tended on average to shift the data in the right direction, toward the true solution, but usually not far enough. T09 found that the behavior depended on the assumed error structure. In the case of one error model, they could not capture the true lower tropospheric tropical trends. T09 were therefore able to place a lower bound on the observed trend, precluding a cooling tropical troposphere, but were not able to establish a reliable upper bound.

[6] This paper is a final contribution from the current HadAT radiosonde temperature project (although real-time updates of the HadAT radiosonde data will continue). It complements M08 and T09 in two fundamental ways. First, it includes a wider range of methodological choices, such as input data time resolution, entire adjustment procedure, and dynamical neighbor selection. These results should give a more comprehensive guide to the true structural uncertainty. Second, a conditional probability assessment of "true" atmospheric temperature trends is made. This permits, for the first time, a bounded estimate of the uncertainties in the observations. The paper focuses upon the tropics as apparent discrepancies between modeled and observed lapse-rate changes there continue to receive considerable attention [*Santer et al.*, 2005, 2008; *Karl et al.*, 2006; *Thorne et al.*, 2007; *Douglass et al.*, 2008; *Allen and Sherwood*, 2008; *Klotzbach et al.*, 2009; *Bengtsson and Hodges*, 2009]. Trend estimates are also presented and briefly discussed for the globe and the extratropics in each hemisphere.

[7] Section 2 outlines the data set construction algorithm that is employed and the data that are utilized. In section 3 the suite of systematic experimentation that was performed subsequent to T09 is outlined and analyzed. Section 4 outlines the method used to combine the information from these experiments to create a conditional estimate of the true world behavior. Section 5 discusses the observational results. Sections 6 and 7 provide a discussion and conclusions.

2. Data Set Construction

2.1. Automated HadAT System Methodology Overview

[8] The HadAT system is an iterative neighbor-based breakpoint identification and adjustment algorithm. Neighbors are selected from the contiguous region where, according to the NCAR [*Kalnay et al.*, 1996] or ERA-40 [*Uppala et al.*, 2005] reanalyses over the satellite era the correlation between atmospheric temperatures and the target station is >1/e. Neighbor averages are then constructed from these neighbor composites, and the difference series (target minus neighbor average) is used to define breakpoints and calculate adjustments. Use of the difference series removes common climate signals and hence enhances signal-to-noise ratios for both breakpoint identification and adjustment steps [*Santer et al.*, 2000]. If no breaks existed in candidate or neighbor series, it is assumed that the difference series would simply constitute white noise with variance governed by neighbor density and local climate noise. The system adjusts the most obvious breaks in the data in initial iterations, and then relaxes the breakpoint identification criteria in subsequent iterations so that it picks up smaller breaks. It was originally run manually [*Thorne et al.*, 2005a], but this approach was expensive and irreproducible so the system was subsequently fully automated (M08). In standard form, it uses seasonal resolution radiosonde data for each mandatory reporting level for which a climatology can be calculated.

[9] A Kolmogorov-Smirnov (KS) test [*Press et al.*, 1992] applied to the available difference series at each level, together with the available (known incomplete) metadata (*Gaffen* [1993] and updates), are used to estimate local

maxima in the likelihood of a breakpoint. Breaks having p values greater than a specified threshold (a system tunable parameter) are adjusted in the station series.

[10] The adjustment is a constant applied to all data prior to an assigned break. Following adjustment of all stations, anomalies are recomputed, and the neighbor composite temperature series is recalculated and the critical threshold reduced. The number of iterations is defined by the user.

[11] After the iterations have completed the data are gridded and vertically weighted to produce Microwave Sounding Unit (MSU) equivalent layer temperature anomalies. Global and regional series and trends are then calculated for all pressure levels and the MSU layer equivalents. Full system methodological details are given by M08 and T09.

[12] Although automation has the disadvantage that any individual data set realization is not subject to the same expert scrutiny applied in a manual approach, there are three distinct advantages: (1) It is reproducible (2) it takes hours to weeks rather than years to create a data set version and (3) the process can be parallelized across many computers, enabling large ensembles to be produced. To take full advantage of the automated processing, all subjective methodological choices (such as the number of iterations, the initial breakpoint test critical value, etc.) are regarded as tunable parameters. By randomly selecting parameter values within reasonable bounds (M08 and T09), ensembles of "plausible" climate realizations can be produced. Since the true evolution of historical climate change is uncertain, a set of "radiosonde data" error models constructed from an atmosphere-only climate model run was created to benchmark the performance of the automated data adjustment system. This provides guidance regarding the inferences that can be drawn when the automated data adjustment system is applied to real-world radiosonde data (T09).

2.2. Observational Data

[13] The observational data described in detail by T09 are used in these analyses. The observations are a merge of ERA-40 input data radiosonde holdings [*Uppala et al.*, 2005] and the Integrated Global Radiosonde Archive (IGRA) [*Durre et al.*, 2006] undertaken as part of the RAOBCORE effort [*Haimberger et al.*, 2008]. Where duplicate records existed the ERA-40 ingest data were used. From this combined data set, those station records for which a 1981–2000 climatology could be calculated were retained. Day and night data were considered separately, being defined by an algorithm using the station's recorded longitude and observing time; and that also removes polar latitude sondes where day and night assignment would be dubious. Following T09, who found the automated HadAT data adjustment approach was not sufficiently constrained by the sparser nighttime network in regions outside the Northern Hemisphere midlatitudes, the focus herein is on the daytime record, despite its well documented radiative heating issues with changes in the materials and shielding of temperature sensors [*Sherwood et al.*, 2005; *Randel and Wu*, 2006]. To facilitate direct comparison with the earlier work, the analysis is restricted to the period 1958 to 2003 used by M08 and T09.

2.3. Error Models

[14] Error models were derived from a run of the Hadley Centre's atmospheric model HadAM3 [*Pope et al.*, 2000] forced with observed sea surface temperatures (SSTs) and natural and anthropogenic forcings as used by *Tett et al.* [2006]. The 4-D atmospheric temperature field from this integration was subsampled with the observed spatiotemporal pattern from the daytime radiosonde coverage. White noise was then added to approximate the point nature in both space and time of radiosonde measurements compared to the grid box average values of the model output. Then seasonally invariant break structures were added to approximate nonclimatic influences. All error models tested here assume step changes in bias, with no trend-like bias changes.

[15] Four distinct error models were constructed: (1) *current understanding*, using breakpoints based upon existing literature; (2) *many small breaks*, using a large number of small breakpoints and a few large breakpoints; (3) *removal of signal*, where breakpoints were deliberately biased so that above the 150 hPa level they cancel stratospheric cooling while those below cancel tropospheric warming, with a net result of removing the deep layer signal; and (4) *few large breaks*, using fewer breakpoints than other models, where most of these are quite large.

[16] Further details are given in Appendix A of T09. The four error models were designed to be as distinct from each other as possible; each includes at least some of the error characteristics that are likely to affect the real observations including concurrent changes across countries or instrument types. These constitute a small sample of a much larger population of potential error models against which one could benchmark system performance. Principal differences between the error models relate to the prevalence, timing and magnitude of the breaks, how well they are associated with metadata events, and the tendency toward breaks of a given sign. The white noise which was added to approximate measurement sampling effects is similar across the four error models. It is important that the homogenization algorithm is able to cope with error structures other than a priori assumptions regarding the true underlying error structure in case these subsequently prove unfounded. These assumptions could include the prevalence, magnitude, preferential sign bias, geographical coherence or other aspects of the bias structure. Creation of distinct error models should avoid the otherwise insidious potential to overtune homogenization algorithm performance toward a desired outcome.

[17] The only difference from the data used by T09 is in the time resolution. While T09 used seasonal data only, a small subset of the present analysis uses data with monthly or pentadal (5 day means) resolution. Pentads are the finest resolution on which the raw climate model data run used was archived. In the error models, this required randomly assigning the break events to a given date within the stated season to ensure a degree of uniformity in the temporal distribution of breaks. This and the effects of the increased number of data points had only a small impact on the resulting raw large-scale average trend estimates. The pentad resolution model data were available substantially before the equivalent resolution observational data. So it was necessary to assume that if a month reported in the observations then all pentads in that month reported. So there is a very slight mismatch in pentad resolution sampling between the error models and observations where

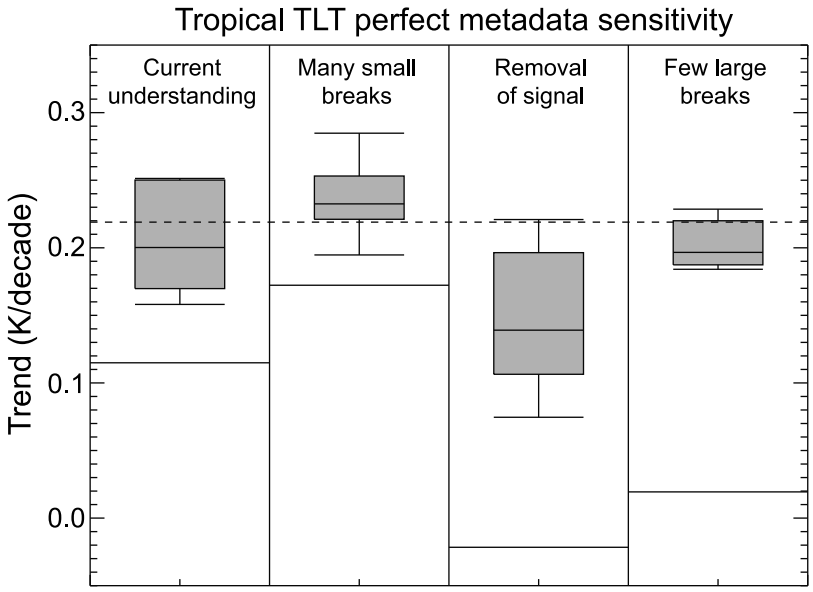

Figure 1. Tukey box plot [*Cleveland*, 1994] for tropical average TLT trends for 1979–2003 from a 10-member "perfect metadata sensitivity" ensemble for each error model. The shaded region denotes the interquartile solution range (with a horizontal line denoting the median value), and the whiskers denote the data range beyond these bounds. Here and elsewhere, tropical averages have been derived by zonally averaging and then cos(lat) weighting over the region 20°S–20°N. The resulting trends have been calculated using a median of pairwise slopes estimation technique that is robust to outliers [*Lanzante*, 1996]. The thick horizontal lines are the trends in the HadAM3 data after imposition of the errors. The dashed horizontal line is the trend in the HadAM3 data before imposition of the errors, and therefore represents the target for homogenized trends.

this assumption is not valid, which will have negligible impact.

3. Systematic Experimentation With Error Models

[18] This section summarizes the results of systematic experimentation since T09. Consideration is deliberately limited to the four error models, where the answer is known a priori. The principal aim is to ascertain whether there were any methodological aspects overlooked in the analyses up to and including T09 that may better constrain the uncertainty in real-world trends. Many of these experiments involve mimicking the choices of other independent teams of analysts who promoted particular strengths of their respective approaches compared to the HadAT methodology. T09 found that *removal of signal* was the only error model for which the HadAT methodology could not capture the true tropical daytime trend with its range of ensemble estimates, so this issue is also addressed.

[19] The section aims to answer the following questions:

[20] 1. Is the homogenization system capable of encapsulating the true trend within its range of estimates if perfect knowledge of breakpoint locations is given? (The answer is yes: see section 3.1.)

[21] 2. Does the phasing of the breakpoint locations affect system performance (yes) and can we reject as a plausible error model the *removal of signal* error model which shows a substantial clustering of the largest breaks? (Our conclusion is no; see section 3.2.)

[22] 3. Does the temporal resolution of the input data affect system performance? (We conclude yes; see section 3.3.)

[23] 4. Does removing apparently breakpoint-impacted neighbors before undertaking adjustments improve system performance? (The answer is no; see section 3.4.)

[24] 5. Does applying a fundamentally different adjustment algorithm improve system performance? (Our conclusion is yes, slightly; see section 3.5.)

3.1. System Performance With Perfect Knowledge of Breakpoint Locations

[25] M08 and T09 underestimated the required shifts in large area average trends, despite capturing the required sign. It is important to ascertain whether this was a result of inadequate adjustment, or of their incomplete break detection (<50%, although better for big breaks). The first additional test was to assess the adequacy of the adjustment step in isolation. If the automated adjustment process cannot capture the true trends, even with perfect knowledge of break locations, it should be rejected. For radiosonde temperatures, homogenization where breaks are known is better than where breaks are unknown [e.g., *Sherwood*, 2007]. T09 undertook a similar experiment using the GUAN network of 161 stations, and found that for this sparser network, the true trend was not consistently captured for the error models. Given that the algorithm is neighbor-based, this may simply reflect the sparseness of the GUAN network. It should be noted that a bias free GUAN would adequately capture global and regional scale changes [*McCarthy*, 2008].

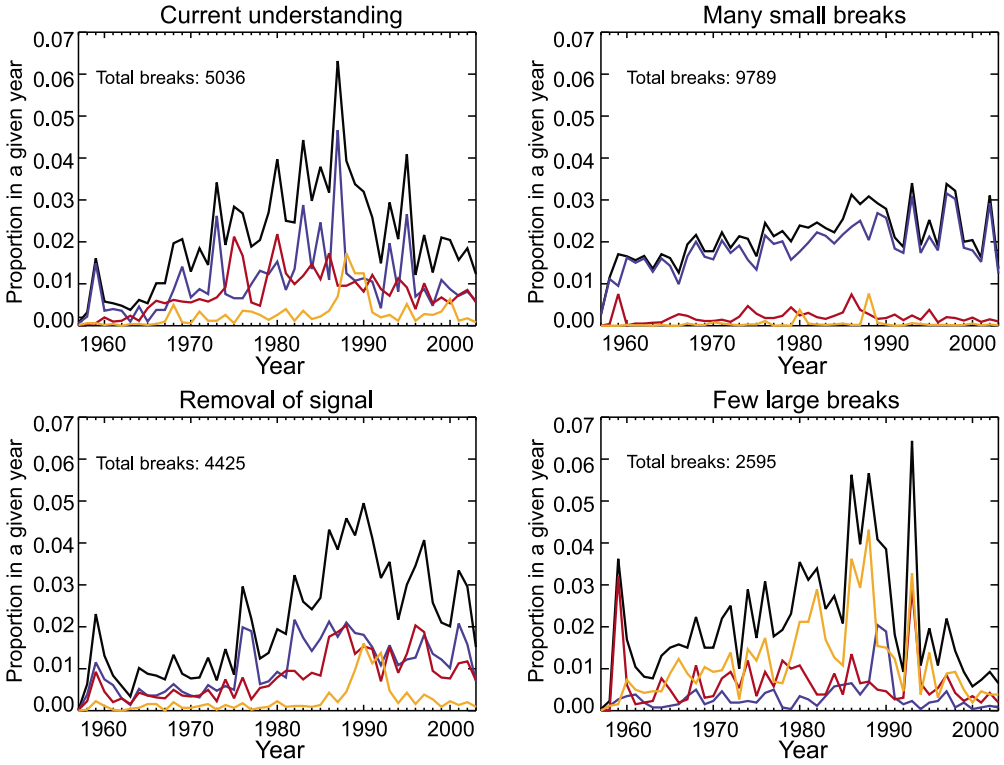

Figure 2. Temporal distribution of worldwide breaks added to the four error models used by T09. Black denotes total breaks, orange denotes large breaks (>1K at three or more levels), red denotes medium breaks (>0.5K at three or more levels), and blue denotes small breaks. Total break number differs by a factor of 4 between the error models, reflecting the fundamental uncertainty in the real-world break structure and frequency.

[26] By tagging all breakpoints through provision of complete error model metadata to the automated system and by tuning system parameters, one can force the automated system to assign a break at all real breakpoints, specifying a single break for multiple real breaks in quick succession and few, if any, additional breaks. A 10-member ensemble for each error model was run with the breakpoint identification parameters set to minimize the chances of finding any breaks not associated with recorded metadata events. The number of iterations and all system parameters associated with the adjustment step were unconstrained.

[27] Results (Figure 1) show that if the system were able to find all the real breaks, and only these breaks, then it would be able to encompass the true trend in all error models, although it is not within the interquartile range of this small ensemble for *removal of signal*. The experiment therefore confirms that the adjustment algorithm taken in isolation is adequate. However, we caution against overinterpretation of this result. The reason for this is that in the real world, there is imperfect a priori knowledge of break locations, and the breakpoint identification and adjustment steps are intertwined in the automated system.

3.2. Sensitivity to Breakpoint Phasing

[28] All four error models had a large number of biases applied at the same time across countries or across commonly recorded metadata event types, such as the VIZ to VIZ-B instrument change (T09's Appendix A). Such biases are not randomly distributed in space and time. Available metadata and previous analysis by multiple groups [*Thorne et al.*, 2005a; *Sherwood et al.*, 2008; *Haimberger et al.*, 2008; *Lanzante et al.*, 2003] strongly implies that some biases with similar characteristics exist in the observational data. However, because the available metadata are grossly incomplete, it is impossible to accurately ascertain the extent of spatiotemporal clustering of nonclimatic effects in the real-world observations. It is therefore of interest to determine to what extent the degree of clustering of such breaks may inhibit system performance.

[29] Figure 2 shows that the distribution of breakpoint frequency with time differs substantially between the four error models; *removal of signal* has a greater clustering (particularly of the system-detectable breakpoints >0.5K; see M08) than the other error models. This was not a design feature of the error model, and may partially explain its resistance to homogenization found by T09. To understand the relative contributions of geographical and temporal breakpoint clustering, two further versions of *removal of signal* were created. In the first, breaks assigned within 6 seasons of valid (nonmissing) data points at the arbitrarily chosen 300 hPa level at individual stations were averaged at all levels and placed at the location of the largest 300 hPa break. No homogenization procedure can be reasonably expected to identify individual breaks which are very close in time, and the automated system specifically assigns only a single break in such cases. The second new version of

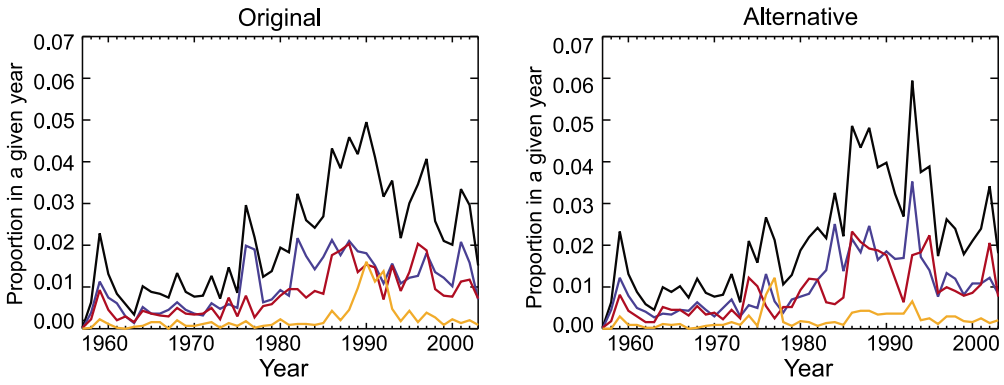

Figure 3. Same as Figure 2 but for the original and alternative versions of *removal of signal*. The total number of breaks applied is identical, but the distribution, particularly for the largest breaks, is substantially different.

removal of signal had similar gross tropical vertical trend profile characteristics, but with a completely new set of breakpoint locations and profiles (Figure 3). Break locations and timings were again derived from a random number generator, under the same assumptions applied in the original error model test. The new breaks were somewhat more evenly distributed in time, with less clustering of medium and large breaks.

[30] The 100-member ensemble of the automated system used by T09 was rerun on both these alternative versions of *removal of signal* (Figure 4). Both ensembles on average perform better than the original, although this may be partly because their overall initial biases are slightly smaller. In particular, the version with less overall breakpoint clustering moves the bulk of the estimates much closer to the true trend and the full spread of trend estimates encompasses the "true" model trend. It is somewhat more consistent with the other T09 error model results, but still more conservative.

[31] T09 could not place an upper bound on the observed ensemble results because *removal of signal* did not capture the true trend and could not be rejected. It is worth considering whether *removal of signal* can be unambiguously rejected as a plausible error model due to the apparently extreme nature of its breakpoint distribution. If so, T09 could be reassessed to provide a bounded estimate. For a randomly selected member of the ensemble used by T09, the system captures the overall shape of the break locations, although it substantially underestimates their frequency (Figure 5, top). This similarity of overall breakpoint identification and real structure is consistent across most ensemble members and all

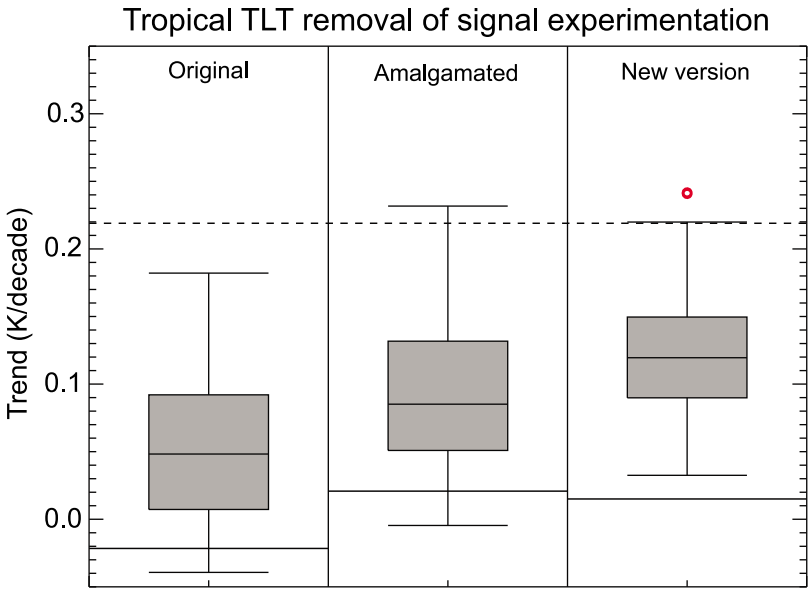

Figure 4. Box-whisker plot for *removal of signal* and the two sensitivity studies of amalgamating rapid-succession breaks and using an entirely new break structure (Figure 3). Results are for tropical TLT for 1979–2003 from T09's 100-member ensemble settings. Legend is as in Figure 1 except that any outliers beyond 1.5 (3) interquartile ranges are shown by open (closed) symbols.

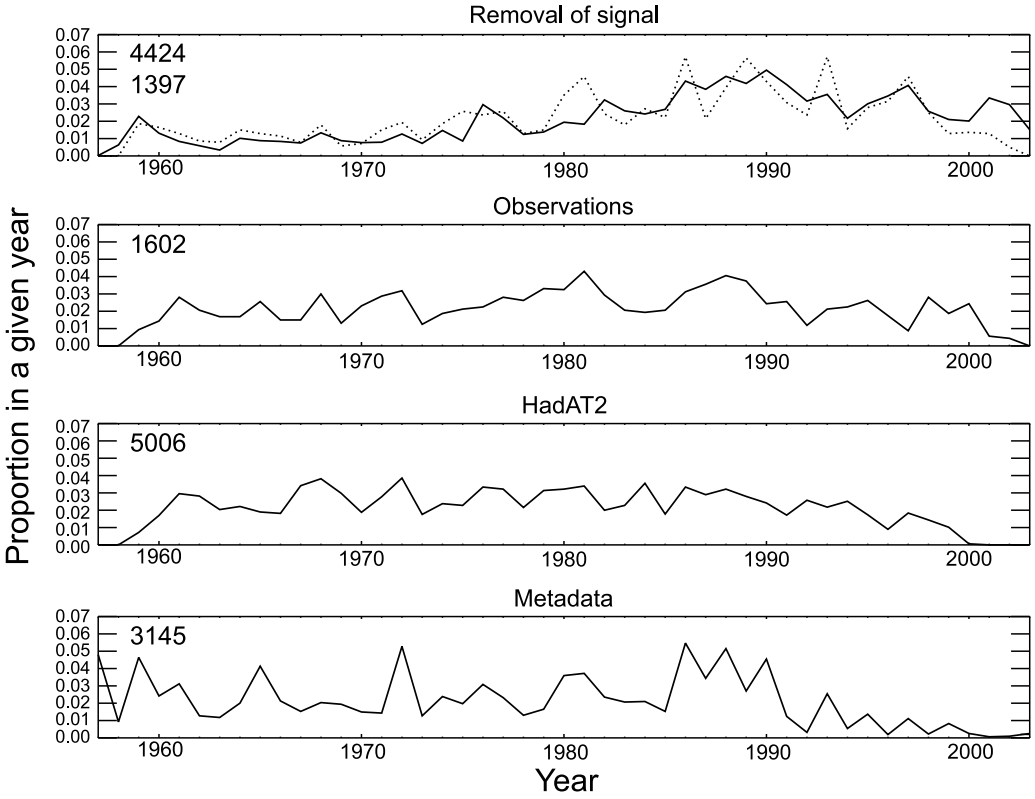

Figure 5. Analysis of structure of real or suspected breaks. First panel shows the real break structure from *removal of signal* (solid line) and the identified breaks (dashed line) from the first of the T09 "top seven" experiments. Note that the pattern matches but the absolute number of breaks (printed at top left) is substantially different with T09's experiment finding only a subset of the real breaks. Second panel is the same experiment but applied to the observations. Third panel is the manually derived HadAT2 break structure. The fourth panel shows available metadata events in the IGRA data holding [*Durre et al.*, 2006] and based upon substantial efforts by Dian Seidel (nee Gaffen) in the early 1990s [*Gaffen*, 1993].

error models (not shown). The same settings applied to the observations yield little or no clustering. HadAT2, which with its manual intervention identifies many more breakpoints (see M08 for further discussion), exhibits even less clustering. The available metadata exhibit some spikes that may be related primarily to when the major metadata collection efforts occurred (the early 1990s) rather than to any real effect. However, *Sherwood et al.* [2008], who did not use metadata in the breakpoint identification step, show a degree of clustering similar to the original *removal of signal* error model (their Figure 2). So, although there is at best limited evidence for real-world breakpoint clustering as severe as that in the *removal of signal* error model, it cannot entirely be ruled out.

3.3. Impacts of Using a Finer Temporal Resolution

[32] Many of the other radiosonde homogenization methodologies that have been developed have used a finer temporal resolution than the seasonal resolution in the manually produced HadAT data set and T09. The RICH (Radiosonde Innovation Composite Homogenisation) methodology considers data at the individual observation level to both identify and adjust for breakpoints [*Haimberger et al.*, 2008]. IUK (Iterative Universal Kriging) considers observation level data in its homogenization step and monthly mean data in its breakpoint identification step [*Sherwood et al.*, 2008]. Both exhibit a greater shift from the raw data than T09 or HadAT. Both groups had claimed that this may constitute a relative strength of their approach.

[33] A new 20-member ensemble using "optimal" system parameter settings from T09 (for those parameters for which these could be ascertained; see Table 1) was therefore applied to the original seasonal resolution data, and to versions of the error models and the observations at monthly and pentad (5 day) resolution. The seasonal resolution ensembles are invariably shifted closer to the true trend than the equivalent larger 100-member seasonal ensembles use by T09 (Figure 6; see also Figure 5 of T09). This reflects the choice of more optimal parameter keyword settings in this ensemble, which removes an artificial tail toward the raw data used by T09: many experiments used by T09 would have been too conservative when identifying breakpoints and applying adjustments.

[34] Temporal resolution has a substantial impact on performance. For each error model, the monthly resolution ensemble generally performs best and contains the true trend, although this is still not within the interquartile range for *removal of signal*. For pentad resolution, the ensembles

Table 1. Apparently Optimal Settings in the Analysis Done by T09, Used in Fixed Configuration in Many of the Further Experiments Here[a]

Parameter	Range Given by T09	Value Used in Present Analyses
Max iteration Number of iterations of the breakpoint identification and adjustment algorithm	3, 6, or 9	9
Adjustment method Whether previously identified adjustments are recalculated for all previously identified breaks with new neighbor composites (adaptive) or not (nonadaptive) at each iteration.	Adaptive or nonadaptive	Adaptive
Adjustment threshold Number of additional tests a calculated break must pass to be adjusted based upon the series behavior. Higher numbers imply harder criteria. The first number is a threshold for the average across all levels and the second that at least one level must attain. See M08.	[1,1], [5,8], [5,11], or [7,11]	[1,1] or [5,8]
Adjustment period The maximum window either side of the break used to derive the adjustment.	5–20 or 40–55 seasons	5–20 seasons (scaled for monthly and pentad analyses)

[a]A complete listing of system parameters is given by Table B1 of T09. If not included above, then the range given by T09 is used here.

incorporate the truth for each error model but, on average, overestimate the required adjustment and exhibit much larger variance.

[35] The iterative HadAT procedure is not compelled to converge to any apparent minimum unbiased solution. So potentially, it can migrate to unrealistic states before it is halted. The error models have substantial value in ascertaining and quantifying the potential for such behavior. Without them, it may have proved tempting to put disproportionate weight on the pentad resolution results as that ensemble gave apparently best agreement with climate model expectations when applied to the observations. But clearly, this would have yielded an unwarranted optimistic assessment of the extent of model expectation-observation agreement as the error models consistently imply that the system will have, on average, shifted the data too far. That is to say there is a propensity to shift beyond the known true solution in the error models.

3.4. Neighbor Selection Effects

[36] The automated system uses a composite weighted average of neighboring stations for both breakpoint identification and adjustment steps (M08). To date, it has weighted the neighbors by the expected correlation coefficient derived

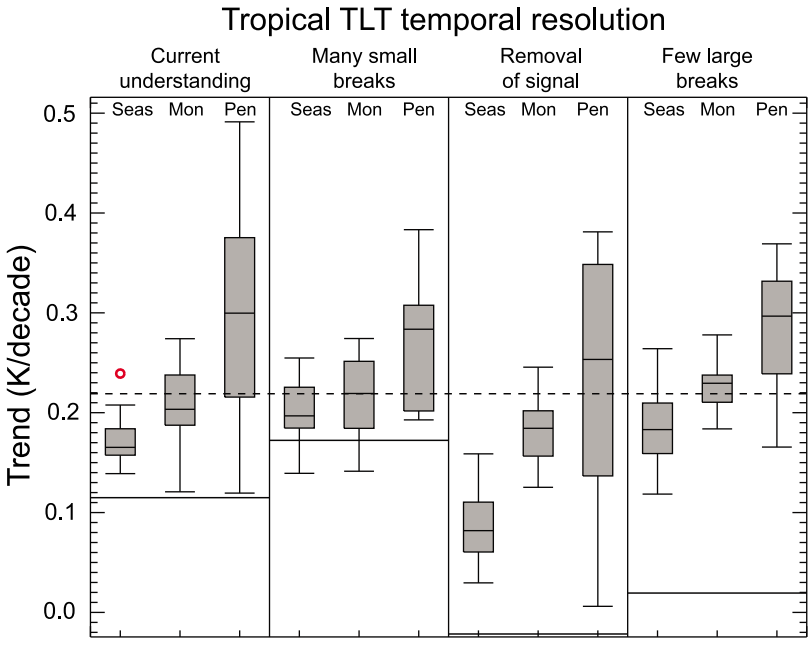

Figure 6. Analysis of tropical TLT trend sensitivity to input time resolution, 1979–2003. Based upon a 20-member ensemble with system tunable settings used by T09 set to their optimal values where these could be ascertained (Table 1). Legend is as in Figure 4.

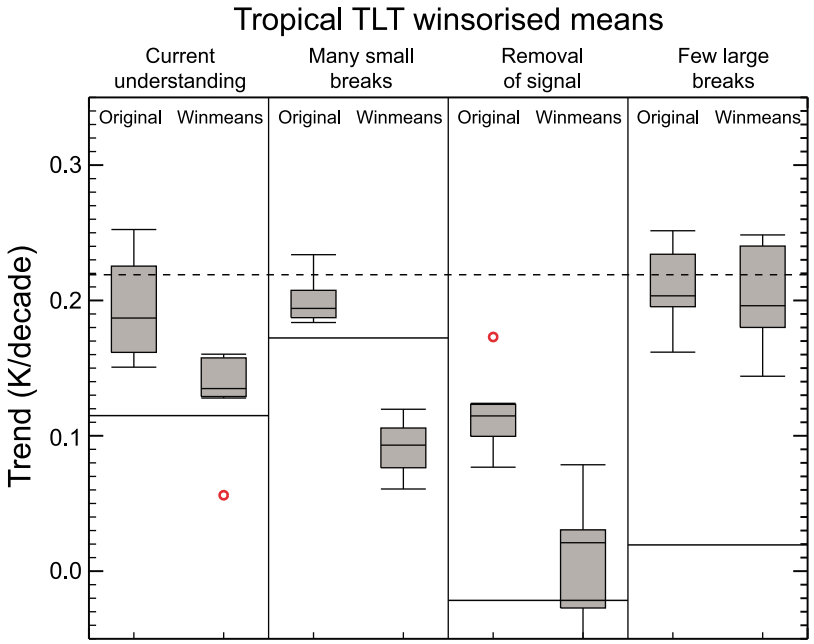

Figure 7. Sensitivity of tropical TLT trend, 1979–2003, to using winsorized means as an alternative neighbor technique. Results are from the seven best experiment configurations identified by T09. Legend is as in Figure 4.

from reanalysis products over the post-1979 era (during which time reanalyses were constrained by assimilated satellite and radiosonde information). However, this explicitly ignores any available information on the actual data quality at the station level.

[37] As a first alternative, instead of using the expected correlation between each neighbor and the candidate station to weight the values, the neighbor average was calculated as the winsorized mean (average within the interquartile range of neighbor values at each time step) of the candidate neighbors. This approach was applied solely to the system settings that yielded the top 7 experiments identified by T09 at the seasonal timescale (Figure 7). In all cases the resulting adjusted trends are biased toward zero trend, very substantially so for some error models. *Gaffen et al.* [2000] found very similar behavior in early assessments of the effects of statistical techniques, with or without metadata, on temperature trends at individual radiosonde stations, though these efforts did not use neighbors. Subsequent analysis in the development of the IUK system also yielded similar conclusions [*Sherwood*, 2007]. Logically, the winsorized mean would tend to be closer to zero at each time step and therefore lead to adjustments (based on neighbor average minus target station), that would on average bias the target station record toward zero trend. However, explicit analysis of the issue was not undertaken.

[38] The RICH data set uses *only* apparently homogeneous neighbor segments to adjust each candidate station at suspected breakpoints [*Haimberger et al.*, 2008]. Although there are two system parameters which can exclude the use of neighbors from the same country and/or with similar metadata records, employing these choices does not guarantee a set of homogeneous neighbors around each identified breakpoint. Ignoring stations afflicted with apparent breaks within the adjustment period in the adjustment step would reduce the chances of simply exporting error structure from these inhomogeneous series through their impact upon the neighbor composite across the network and retaining some systematic mean bias. However, use of adaptive adjustments which are recalculated each iteration based upon the modified neighbor series may mitigate against this. When applying the automated system to humidity data, where data issues are even more substantial, to create HadTH [*McCarthy et al.*, 2009] the use of a first-guess adjustment prior to data input to the system was required to handle this issue.

[39] Two variations on blacklisting were tested: ignoring neighbor data around breaks found in all iterations; and around those breaks found only in the current iteration. In both cases neighbors were recalculated after masking out data within the specified adjustment period used (which is a system tunable parameter) each side of the identified breaks before proceeding to calculate the adjustment factors. The more aggressive blacklisting causes many stations to have too few neighbors when recalculated to form a neighbor average which meant that no neighbor estimate existed around the identified breakpoints and let through an unacceptably high number of breaks unadjusted. This is a substantial issue in data sparse regions (including much of the tropics) and for experiments where many breaks are found. RICH gets around this issue by expanding the search radius until a sufficient number of apparently homogeneous neighbors are identified. However, with increasing distance the degree of correspondence to the candidate station being adjusted will, on average, decrease. Such an approach was not pursued as it would have required a substantial system rewrite.

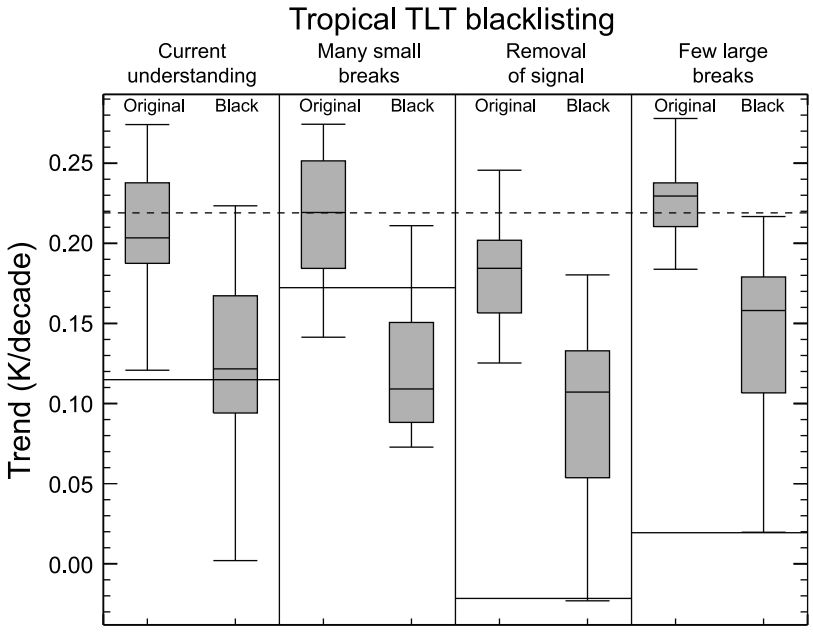

Figure 8. Sensitivity of tropical TLT trend, 1979–2003, to blacklisting apparently inhomogeneous segments identified in the current iteration prior to the adjustment step. Based upon the 20-member monthly ensemble (Figure 6). Legend is as in Figure 1.

[40] For the less aggressive blacklisting these issues were much less obviously pervasive. It was applied to the monthly error model data using the same 20 experimental setups as was used in section 3.3 (Figure 8). The ensembles produced exhibit a larger spread of solutions and are on average biased toward zero compared to the default approach. The impact is largest in data sparse regions, implying that the masking out of neighbors is still the fundamental issue, and that bad neighbors are better than no neighbors: at least the former allow some estimate of an adjustment factor rather than simply letting the data through in raw form without any adjustment at all. Without a complete adjustment step rewrite, therefore, blacklisting does more harm than good.

3.5. Sensitivity to Fundamental Adjustment Approach

[41] In section 3.1 it was shown that in the presence of perfect knowledge regarding breakpoint locations, the automated system could adequately retrieve the true trends in all error models. In the real-world situation, where knowledge of breakpoint location is far from perfect, it is of interest to consider whether an alternative adjustment algorithm may perform better. The Iterative Universal Kriging (IUK) system explicitly separates the breakpoint identification and adjustment steps [*Sherwood et al.*, 2008]. It was therefore possible to feed the IUK adjustment step with the breakpoints identified by the automated system after its final iteration. This enables a clean assessment of whether an alternative, published, adjustment methodology offers any advantages. Certain performance indicators of the IUK system on the error models not covered here are detailed by *Sherwood et al.* [2008, section 4]. Importantly, like the automated system, IUK recovered the true trend when all break locations were specified ("perfect metadata," section 3.1) for all four error models [*Sherwood et al.*, 2008].

[42] The IUK adjustment step involves fitting the entire global data set, with missing values imputed, to a regression model that includes leading modes of variability, the trend, breakpoints and noise terms [*Sherwood et al.*, 2008; *Sherwood*, 2007]. It iterates to a maximum likelihood unbiased estimate given the data availability, characteristics, and the specified break timing and locations. Therefore, unlike in the automated HadAT system, convergence is an integral component of IUK.

[43] IUK was developed to be applied at the individual observation level. Although it is trivial to apply IUK to seasonal averages, it is possible that much of its power may be lost in coarsening up the temporal resolution to the seasonal level. The selection of time resolution needs to be balanced against the substantially reduced computational overhead necessary to produce an ensemble of realizations.

[44] In addition to the documented IUK, a variant was developed in which the trend basis function in the regression was allowed to take a polynomial form (up to fifth order) rather than a linear form if the data supported this. True climate evolution may well be better described by other than simple linear functions [*Seidel and Lanzante*, 2004; *Thorne et al.*, 2005a; *Karl et al.*, 2006]. Allowing the temporal evolution term to take an optimal form based upon the data may help in retrieving the true trend so long as there is sufficient information in the data to specify the form of the variation and distinguish it from the bias pattern.

[45] The break locations output from each of the 100-member ensembles used by T09 for the four error models and the observational data were used as input to the IUK system. In the observations (but not the error models) four of the ensemble members used by T09 were found to contain too few breaks for the IUK system to run and so these ensemble members were not included for either error models

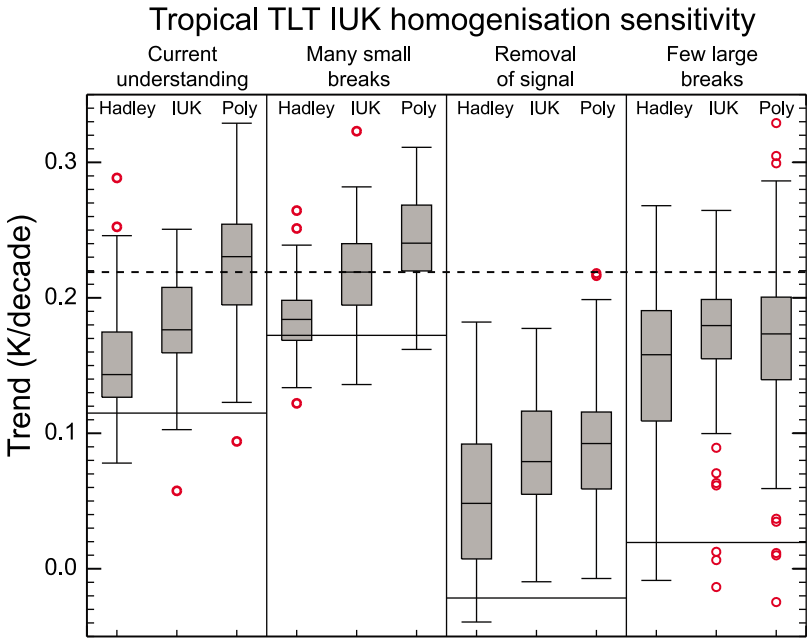

Figure 9. Assessment of sensitivity of tropical TLT trend, 1979–2003, to using the IUK adjustment step rather than our system's adjustment step. Results are from T09's 100-member experiment except four ensemble members that could not be run on the observations. Legend is as in Figure 4.

or observations. The published IUK methodology consistently outperforms T09 across the four error models, moving the estimated trend slightly closer to the truth (Figure 9). However, it does not move the ensembles all the way to the truth and as shown by T09, IUK-based adjustment does not encompass the truth for *Removal of Signal*. Results allowing a polynomial time series basis function are much more mixed than when IUK assumes an underlying linear trend function. In *current understanding* it does better, for *many small breaks* it substantially overshoots and for the remaining error models it shows no discernible difference from the standard version. This implies a critical dependence with this polynomial trend basis function upon the underlying error structure that is in the real world unknown. The spread in solutions also tends to be greater than either the original adjustment approach or the IUK default approach.

4. Combining Estimates to Produce a Final Estimate of Real-World Trends

[46] In the study by T09, the use of the error model results was limited to making qualitative logical inferences about real-world trends when the same ensemble of settings was applied to the real-world observations. From the analysis by T09 and the suite of analyses summarized in section 3, there now exists an expanded set of estimates showing how the system will behave when applied to simulated data from the four error models. The next logical step is to attempt to combine this information to produce a quantitative estimate of the real-world trends predicated upon the HadAT methodological framework when the same settings are applied to the real-world data. To make the problem tractable, it was decided to limit further consideration to those ensembles where the behavior is broadly consistent across all error models and the skill in trend retrieval at least comparable to T09's ensemble. This yields: T09 (100 members), the three 20-member time scale experiment ensembles (section 3.3) and the default IUK adjustment (96 members, section 3.5).

[47] For each ensemble, the analysis provides a set of adjusted trends for each error model and for the real world. There are also unadjusted trends for each error model and for the real world, and *crucially* for the error models the true trend is also known. This enables an assessment of absolute performance not afforded in the real world where the solution is unknown. Figure 10 shows all these values for the tropical TLT trend over 1979–2003. Each ensemble/error-model combination has both a systematic and a random error component: the random error is shown by the width of the ensemble of adjusted trends, the systematic error by the difference between the mean of the adjusted ensemble and the true trend. To illustrate these distributions quantitatively it is useful to estimate a PDF from each ensemble, by convolving each ensemble value with a smoothing function (a Gaussian with standard deviation 0.05K/decade was used here). These PDFs are also shown in Figure 10. It is assumed that the small ensembles are an unbiased sample of the population that would be created from running a much larger ensemble. Given that system parameter settings were derived randomly for those parameters which were permitted to be perturbed within each ensemble this seems reasonable.

[48] It is clear from Figure 10 that, for example, *removal of signal* generally results in a systematic error (adjustments are consistently too small), while *many small breaks* has little systematic error. Also adjustment at the pentad time scale produces a much larger random error than adjustment at the monthly timescale. These effects in the error models need to be accounted for in the analysis of the real world observations, so (for TLT trend over 1979–2003), the real world adjusted trends should be systematically adjusted upwards to account for the expected underprediction in the

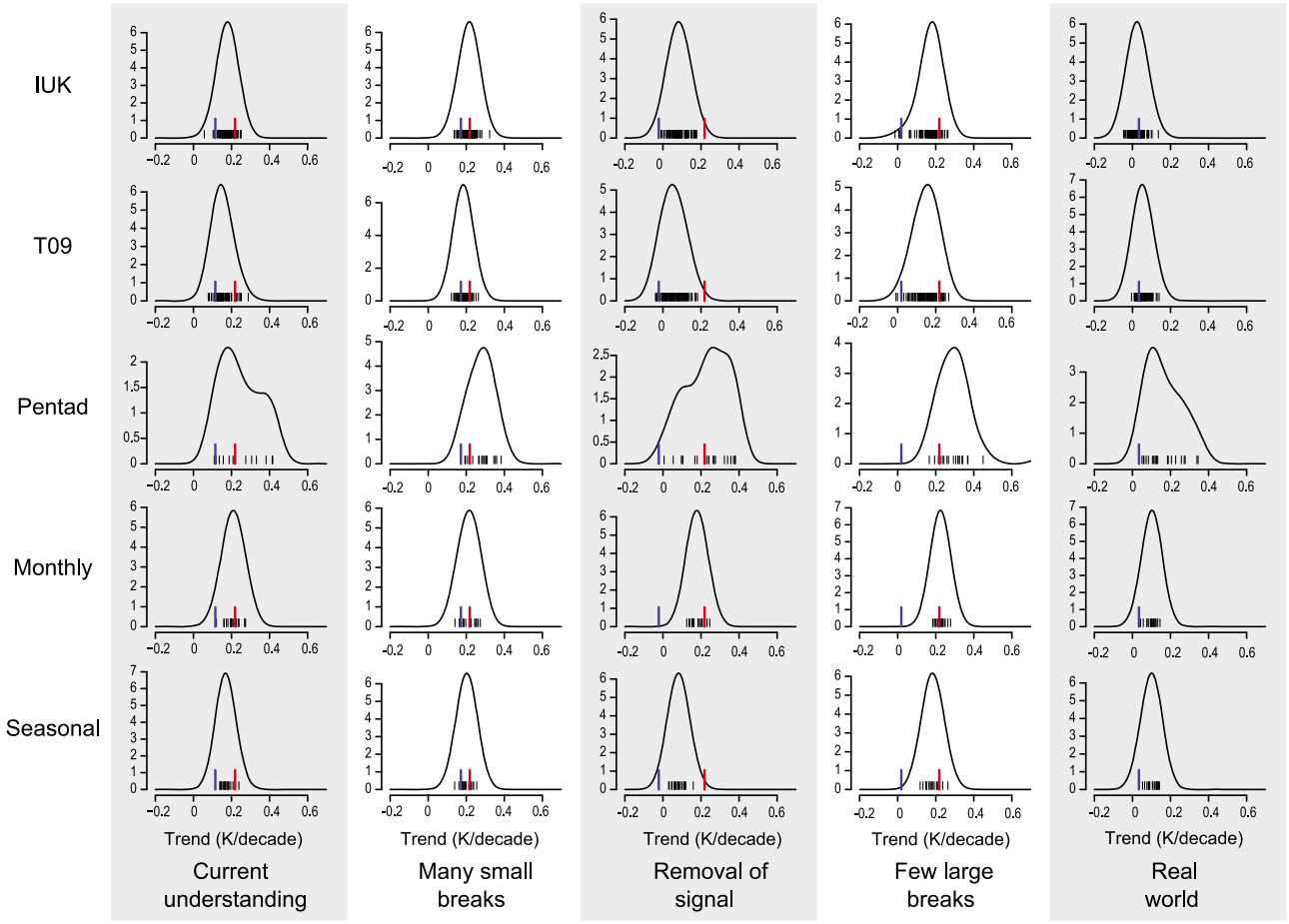

Figure 10. TLT trends 1979–2003. Each plot shows the observed trend including the effects of inhomogeneities (blue), an ensemble of trends after adjustment to remove the inhomogeneities (black rug plot), and a PDF estimate made by smoothing the adjusted ensemble with a Gaussian kernel (black curve). For the four error models the true trend (in the absence of inhomogeneities) is also known (and shown in red).

removal of signal case, and the relatively precise monthly ensemble should have more influence than the more diffuse pentad ensemble. To account for these effects, a conditional probability framework was used to estimate the true solution for the real observations.

[49] Suppose that the *removal of signal* error model were a good representation of the real world. In this case, the real world adjusted trends would need to be scaled up to remove the systematic underprediction. For each ensemble member, the scaling factor needed to convert the adjusted trend to the true trend in the *removal of signal* error model is known. Scaling the real world trend for the same ensemble member by the same factor will remove this systematic error in the real world adjusted trends if the real world error is comparable to that in the error model.

$$O_{s,j} = O_b + \left(O_{a,j} - O_b\right) \times \left(\frac{R - P_b}{P_{a,j} - P_b}\right) \qquad (1)$$

where $O_{s,j}$ is the scaled, adjusted observational trend for ensemble member j, $O_{a,j}$ the adjusted observational trend for ensemble member j before the scaling, O_b the raw observational trend, R the true trend given the error model, $P_{a,j}$ the adjusted pseudo-observations trend for ensemble member j given the error model, and P_b the raw pseudo-observational trend given the error model.

[50] Repeating this scaling for each ensemble member for each of the four error models provides the scaled ensembles shown as the IUK through Seasonal rows of Figure 11. Each of these ensembles is scaled to have no expected systematic error under the condition that the selected error model is a good approximation of reality.

[51] For each error model, each of the five ensembles is an estimate of the same trend, so the likely values of that trend are those that have a high probability in all ensembles. If the ensemble estimates were independent of one another, then the combined probability would be the product of the ensemble PDFs. However, as the ensembles are not independent, this would produce combined PDFs that were too narrow, as the same constraint would be used more than once. Formal methods for combining the ensembles therefore need information on the extent to which the ensembles are independent, and it is not at all obvious how to estimate this. So a simpler approach has been used: a combined PDF has been generated by taking the minimum, at each point, of all of the ensemble PDFs. This uses the unique information

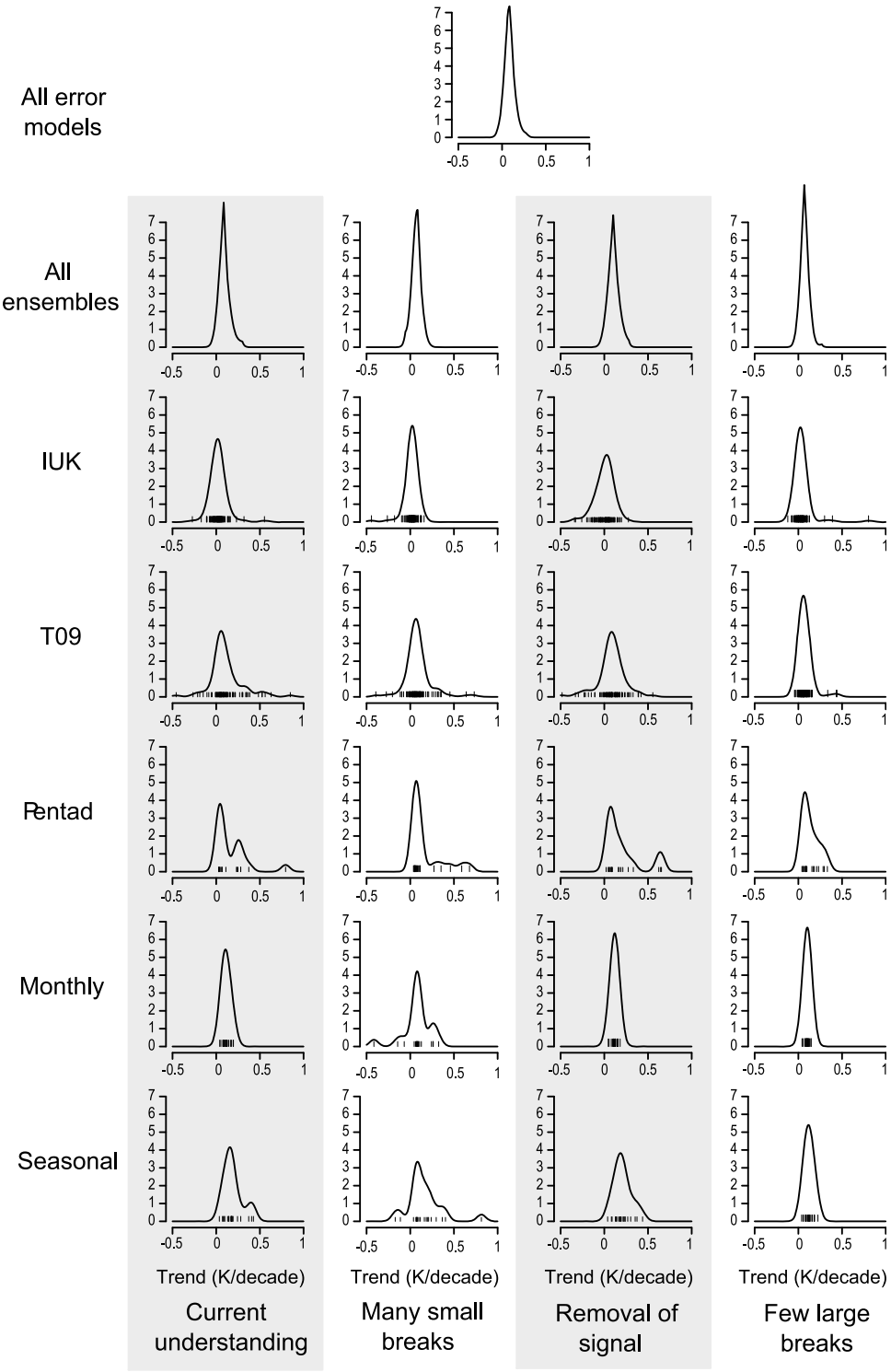

Figure 11. Tropical TLT trends 1979–2003. For each ensemble, the real-world adjusted trends have been scaled by the known errors in each error model, and the resulting ensemble of trends are shown as a rug plot and as a PDF generated by smoothing the ensemble with a Gaussian kernel density estimator. The "All ensembles" row shows the PDFs from the combined ensemble estimates, and the top plot is the mean over the four error models.

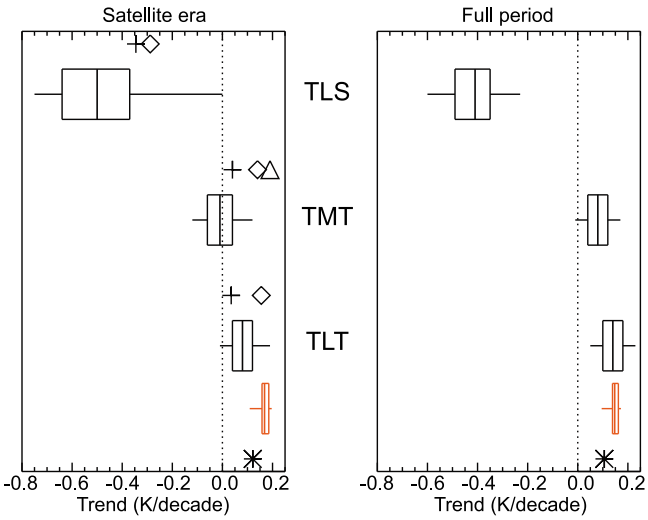

Figure 12. MSU equivalent tropical trend estimates for the satellite era (1979–2003) and the full period (1958–2003). Whiskers show 5–95% range and box interquartile range with median denoted by a vertical bar. Also shown are HadCRUT3 surface trends (asterisks) and model expectations for TLT (red) calculated by multiplying this surface value by the model amplification factor from the model runs used by *Santer et al.* [2005]. Use of either of the two additional commonly cited surface data sets would make relatively small changes [*Santer et al.*, 2005]. It is this amplification factor and not the absolute trend that is strongly constrained in climate models [*Santer et al.*, 2005]. Also shown in Figure 12 (left) are estimates from UAH (pluses), RSS (diamonds) and UMd (triangle) for the same period.

in each ensemble PDF of where the trend probabilities are particularly low, but does not cause narrowing of the combined PDF in regions where multiple ensembles give similar probabilities. The "All ensembles" row in Figure 11 shows this combined probability for each error model.

[52] On the assumption that all of the four error models are equally plausible, it is possible to make a final PDF as a weighted mean of the four error model PDFs ("All error models" graph in Figure 11). The virtue of this approach is that the final PDF makes use of all the ensemble members, and it is possible to see how plausible changes to the system would influence the final PDF. The precision of the PDFs is limited by the small number of members in each ensemble, but in most cases this is not a severe problem. It is clear from Figure 11 that the tails of some ensemble PDFs are undersampled, but it is also clear that adding more ensemble members would be most unlikely to have a large effect on the final PDF. This is true for the tropical lower troposphere trends shown in Figure 11, and for most other regions and heights, but the limited number of ensemble members does become a problem for some stratospheric series (at 50 hPa) where the ensemble PDFs, as estimated from the limited number of ensemble members, overlap insufficiently to allow estimation of a combined PDF. In these cases, no estimate has been made for the combined PDF.

[53] This approach is relatively simple, and illustrates how the uncertainties over choice of error model, process for making the adjustments, and limited ensemble sizes, all contribute to the total uncertainty. But the weighting (prior probability) of the error models is necessarily arbitrary, as are the details of the process for combining ensembles; alternative statistical methods are possible, and would give PDFs that differ in their details. So while these results are likely to be a reasonable indicator of the total uncertainty, the ranges quoted should be treated as approximations rather than precise results.

5. A Reassessment of the Observational Trend Estimates

[54] The PDFs derived in section 4 are now used to provide an analysis of the likely real-world trend behavior. Consistency with preexisting estimates is also assessed (akin to *Mears et al.* [2011]). Over the period 1979–2003, the conditional probability scaled observational estimates of tropical lower tropospheric (LT) temperature trends marginally agree with model amplification behavior [*Santer et al.*, 2005] (Figure 12, left). In addition, the 5–95% range comfortably includes both MSU derived estimates for this layer [*Christy et al.*, 2003; *Mears and Wentz*, 2009b]. The estimated trend of the middle troposphere layer (TMT) is cool biased relative to the available MSU estimates, as is the trend of the lower stratosphere (TLS). Over the full period of record the TLT equivalent measure is in very good agreement with the model expectations (Figure 12, right). Furthermore, the TMT estimate now yields a strong probability of warming of this layer and TLS exhibits less stratospheric cooling.

[55] Tropical TLT trend estimates are relatively robust to choices of inclusion/exclusion of error models and/or ensembles in the conditional probability calculation for both the satellite era (Table 2 and Figures 10 and 11) and full period (Table 3) trends. This provides a degree of confidence in the chosen approach and in the results. Any sensitivity is driven almost entirely by the inclusion/exclusion of ensembles rather

Table 2. Sensitivity of MSU TLT Equivalent Trends (K/decade) in the Tropics, 20°N–20°S, to Choices of Input to the Conditional Probability Assessment Over 1979–2003[a]

	All	Not Current Understanding	Not Many Small Breaks	Not Removal of Signal	Not Few Large Breaks
All	−0.01–0.19	−0.02–0.18	−0.01–0.20	−0.03–0.19	−0.01–0.19
Not Seasonal	−0.02–0.18	−0.02–0.18	−0.02–0.18	−0.02–0.18	−0.02–0.18
Not Monthly	−0.01–0.20	−0.01–0.20	−0.01–0.20	−0.01–0.20	−0.01–0.20
Not Pentad	−0.03–0.19	−0.03–0.19	−0.03–0.19	−0.03–0.19	−0.03–0.19
Not T09	−0.01–0.19	−0.01–0.19	−0.01–0.19	−0.01–0.19	−0.01–0.19
Not IUK	0.00–0.25	0.00–0.25	0.00–0.25	0.00–0.25	0.00–0.25

[a]Columns denote choices of error models and rows choices of ensembles for exclusion.

Table 3. Sensitivity of MSU TLT Equivalent Trends (K/decade) in the Tropics, 20°N–20°S, to Choices of Input to the Conditional Probability Assessment for the Full Period of the Radiosonde Record Considered, 1958–2003

	All	Not Current Understanding	Not Many Small Breaks	Not Removal of Signal	Not Few Large Breaks
All	0.05–0.23	0.05–0.23	0.05–0.24	0.02–0.22	0.05–0.23
Not Seasonal	0.05–0.23	0.05–0.23	0.05–0.23	0.05–0.23	0.05–0.23
Not Monthly	0.05–0.24	0.05–0.24	0.05–0.24	0.05–0.24	0.05–0.24
Not Pentad	0.02–0.22	0.02–0.22	0.02–0.22	0.02–0.22	0.02–0.22
Not T09	0.05–0.23	0.05–0.23	0.05–0.23	0.05–0.23	0.05–0.23
Not IUK	0.05–0.24	0.05–0.24	0.05–0.24	0.05–0.24	0.05–0.24

than error models. In particular the two ensembles that considered higher frequency data appear to differ somewhat from those that considered seasonal resolution data. At least for radiosondes under the HadAT framework, it is more important to adequately assess methodological uncertainty by creating multiple systematic experiments than it is to benchmark against a larger suite of test cases.

[56] Satellite measures are broad vertical integrals that can hide interesting features. A consideration of data on individual levels (Figure 13) shows that the uncertainty estimates from the present analysis do not overlap with model expectations (under a deliberately conservative assessment; see section 6) for the satellite era at any level beyond 300 hPa and only marginally at any level beneath this. Uncertainties are sufficiently large, however, that the estimates are consistent with a local upper tropospheric maximum or an isothermal profile or even a cooling with height through the troposphere. The uncertainty estimates are consistent with all existing data sets up to 400 hPa. Above this several data sets, particularly RAOBCORE, are in disagreement with the conditional probability estimates which are consistently cool biased relative to these preexisting data sets. In contrast to the satellite era, the full period trends are consistent with model expectations up to 150 hPa, the tropical tropopause region. They are also much more consistent with the suite of preexisting radiosonde-based estimates throughout the tropical column. There is stronger support for the existence of an upper-tropospheric maximum as ubiquitously predicted by climate models over this period although the uncertainty estimates are too large to definitively conclude this. Previously published estimates also exhibit good agreement with model expectations throughout the column.

[57] Finally, brief analysis of global, tropical and extratropical hemispheric trend estimates for the full period of record and the presatellite and postsatellite eras was undertaken (Figure 14). In the presatellite era the global troposphere exhibited very little, if any, warming. This was due to a combination of significant Northern Hemispheric tropospheric cooling at all levels up to 250 hPa, and weaker Tropical and Southern Hemispheric warming that was not statistically distinguishable from zero trend at most levels. During the satellite era the global troposphere warmed. In contrast to the earlier period this was primarily due to significant Northern Hemisphere warming while the Southern

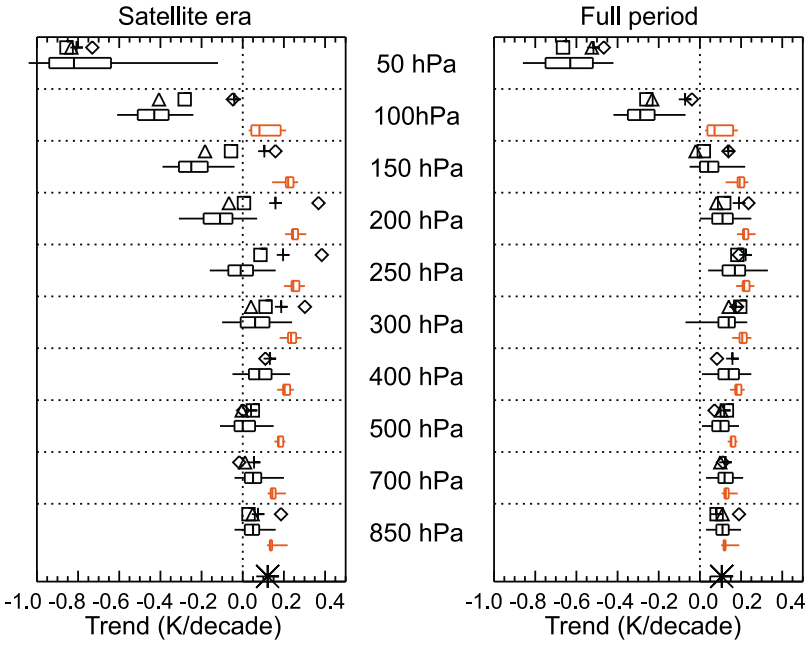

Figure 13. Same as Figure 12 but for individual pressure levels. No model expectation is appended for 50 hPa which is purely stratospheric. A number of other radiosonde estimates are also shown: RICH (pluses), RAOBCORE1.4 (diamonds), HadAT2 (triangles), and IUK (squares).

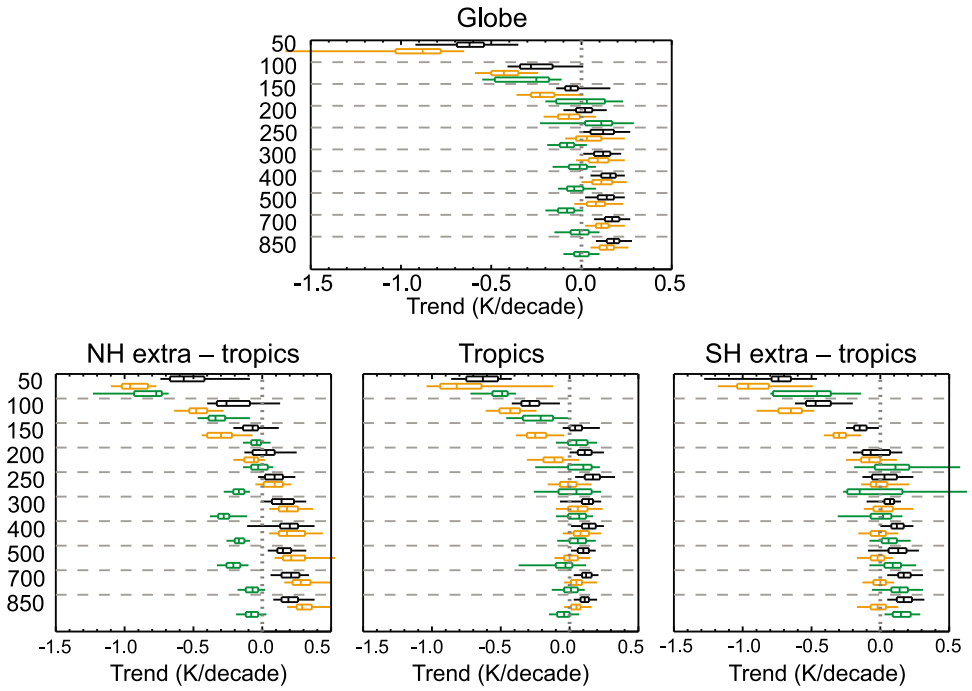

Figure 14. Analysis of trend estimates returned from the conditional probability estimation, by region (Globe, Northern Hemisphere (NH) extratropics (20°N–70°N), tropics, Southern Hemisphere (SH) extratropics (20°S–70°S)) and period (black, 1958–2003; orange, 1979–2003; and green, 1958–1978) for radiosonde levels.

Hemisphere exhibited essentially no trend and the tropics slight warming. Full period trends are more similar to satellite era trends than presatellite era trends in the tropics, Northern Hemispheric extratropics and for the globe. In the Southern Hemisphere the converse is true. In all the periods and regions there is stratospheric cooling which appears to have accelerated over the satellite era.

[58] The estimates for the full period are not, as could be naively assumed, a linear combination of the two subperiods trends at any level. The true climate evolution could equally be explained by a number of other models including step like changes or slope plus steps [*Seidel and Lanzante*, 2004] in both the troposphere and the stratosphere. In part this could be because many anthropogenic and in particular natural forcings are nonlinear, cyclical, or episodic, but also within the troposphere it reflects the role of natural climate variability. Trends from the other data sets considered in Figure 13 support these general findings (not shown for clarity in Figure 14).

6. Discussion

[59] We have outlined a possible approach for better estimating the true behavior of atmospheric temperature over the satellite and radiosonde eras, and for quantifying the inevitable uncertainty in such estimates. Implementation and testing of this approach has yielded a number of useful lessons. First, data set construction must be automated to allow replicability and the fast construction of estimates. Second, it must be modularized so as to allow systematic variation of as many internal methodological choices and external input choices as is feasible. Third, ensembles of data set realizations are required. To make sense of these ensembles, realistic error models are required that mimic likely real-world sampling and atmospheric behavior, and also contain plausible information regarding nonclimatic data issues. Error models should be as distinct as possible, to avoid biasing results to particular preconceived notions on real-world error structure. They provide an important and unique opportunity to understand system performance and limitations through random and systematic experimentation. Finally, estimates of trends in homogenized real-world data need to be interpreted in light of the results of applying the same homogenization criteria to data generated by the error models.

[60] The analysis concentrated on 1979 to 2003 for traceability to earlier analyses, and focused on the tropical troposphere because of continued contention as to whether the observations since the start of MSU satellite observations are in agreement with the strongly constrained climate model amplification behavior [*Douglass et al.*, 2008; *Santer et al.*, 2005, 2008; *Klotzbach et al.*, 2009]. Several of the sets of ensembles that were produced by experimentation with the system yielded comparable behavior across all the four error models against which they were being benchmarked. The results from these ensembles were combined under a conditional probability framework to produce a final two-tailed 90% C.I. estimate of real-world tropical lower tropospheric trends of −0.01K/decade to 0.19K/decade. This result is robust to reasonable inclusion/exclusion of ensembles and/or error models from the conditional probability calculation. For individual pressure levels the uncertainties are larger and the overall agreement with model expectations at some levels is poor, with statistically

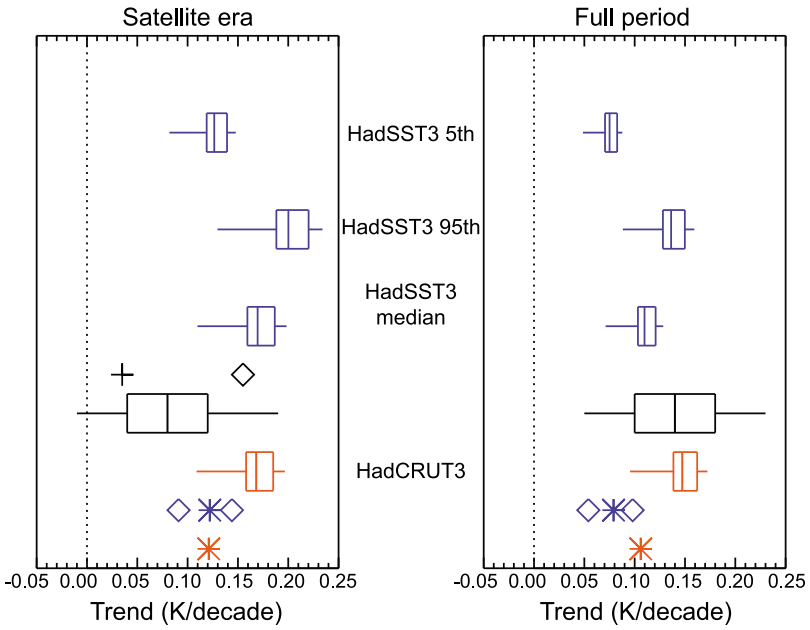

Figure 15. Same as Figure 12 but for TLT only and assessing sensitivity to choice of surface constraint. Figures 12 and 13 used HadCRUT3 (red symbol and whisker), which has only a single estimate. The blue diamonds and symbol denote 5th, 95th, and median estimators from HadSST3 [*Kennedy et al.*, 2011a, 2011b], and the three blue whiskers above denote the scaled model response estimates of the observed TLT response that correspond to each. The black whisker is the conditional probability estimate from the present analysis and the two black symbols existing MSU data set estimates (pluses, UAH; diamonds, RSS).

significant disagreements occurring at several levels, particularly in the upper troposphere, over the satellite era. Trends over the full period of radiosonde records are in statistical agreement with model expectations throughout the tropical troposphere.

[61] The potential residual disparity between the reconstructed tropical tropospheric trends and model expectations over the satellite era could have several explanations. These potential explanations, which are not mutually exclusive, are briefly discussed below.

[62] Residual discrepancies may simply reflect residual biases in the HadAT data, as other data sets have been shown to be closer to model expectations in this region [*Thorne*, 2008, Figure 13]. Indeed, for the current uncertainty analysis to be comprehensive and unbiased would require rejection of several existing satellite and radiosonde data sets as plausible, at least for some layers/levels (Figure 13). Several of these data sets (such as RICH) are much closer to the model based expectations. Although data sets that lie outside the estimates produced here may well be biased, this cannot be definitively asserted from the present analysis. The most likely reasons that the current analysis may not capture the full uncertainty range are (1) that the four error models do not between them adequately capture the error characteristics present in the real world or (2) that the ensemble sizes considered are insufficient. The most obvious potential issue with the error models is that they all assume that nonclimatic influences are all step like changes. Although step-like changes are likely to be the dominant breakpoint type, the presence of more trend-like breakpoints, such as seen at the surface [*Menne et al.*, 2009], cannot be ruled out. Figures 10 and 11 show that in many cases the ensembles are not providing a meaningful constraint. While larger ensemble sizes would be useful, additional (and structurally distinct) error models are probably more important.

[63] Another relatively uncontroversial explanation would be that end-point effects are substantial for the satellite era. *Thorne et al.* [2007] showed that choice of 1979 as a start date was a distinct outlier for all 21 year amplification estimates in both satellite and radiosonde records including all similar length periods in the full radiosonde record. So a 1979 start date may precondition any assessment toward finding a disparity between the observations and model expectations.

[64] The tropical tropospheric trends are likely primarily driven by SSTs in the warmest convecting regions of tropical oceans rather than by combined land surface air temperatures and SSTs [*Santer et al.*, 2008]. Land has been warming faster than oceans in the satellite era globally and in the tropics [*Brohan et al.*, 2006] so the assessment in section 5 is deliberately conservative by choosing use of a combined land and SST data set to constrain model expectations by if this is the case. To assess the potential implications of our surface constraint choice and uncertainty therein, recourse is made to the recently upgraded Hadley Centre SST records [*Kennedy et al.*, 2011a, 2011b]. Rather than being a single estimate this consists of an equiprobable solution set of 100 members and spans similar uncertainty in SSTs, but using a distinct approach from that undertaken here. This allows an assessment of sensitivity to the uncertainty in SST trends in addition to the choice of SST or combined SST and land records

(Figure 15). Use of an SST constraint and uncertainty in the SST trend could easily account for the apparent discrepancy in the satellite era while maintaining consistency in the full radiosonde era if the tropical mean lapse rate is indeed set by the tropical mean SSTs.

[65] The vertical structure of trends over the satellite era, with a relative minimum at around 500 hPa exists in most, if not all, existing radiosonde records and should not be hastily discounted. The 500 hPa level is near the triple point of water and the break between shallow and deep convection in the tropics. There may have been systematic changes related to exchange of heat near the freezing level or the relative frequency of deep vis-à-vis shallow convection that would impart a vertically differentiated temperature trend structure. Diagnosis would require analysis of humidity, cloud and radiation data records that are in poorer shape than the temperature record in this region. There could also be a hitherto neglected or poorly diagnosed climate forcing that can impart this structure. Most logically this would be an aerosol forcing that strongly absorbed radiation substantially in the midtroposphere (and hence warmed it) early in the satellite era and has since rapidly diminished. This would impart a vertically differentiated structure akin to that seen in multiple radiosonde records within the tropics.

[66] Finally, all models may be missing some fundamental climate process such as a nonlinear response to forcing. As discussed by *Santer et al.* [2005, 2008] it is not clear what this could be or why models and observations agree on short timescales but potentially differ on long time scales, given the same fundamental physical processes. There may be natural processes that modulate behavior on decadal timescales that are not captured by any climate models. But with highly uncertain observations it remains most likely that residual observational biases underlie the disagreements with the models. However, if the models lack a basic process, then it urgently needs to be understood and incorporated.

[67] Clearly, these explanations are not mutually exclusive. Equally clearly, the present analysis cannot conclusively inform on these explanations and further research is warranted to elucidate satellite era trends. But it should be stressed that the good agreement between model expectations and our observed analysis over the full 45 year radiosonde record all the way up to the tropical tropopause provides a strong degree of confidence in overall climate model behavior in the tropical troposphere on the longest time scales. This is also seen in the other radiosonde data sets and therefore likely to be real.

7. Conclusions

[68] A comprehensive analysis of the uncertainty in historical radiosonde records has yielded trend uncertainties of the same order of magnitude as the trends themselves. It is highly unlikely that these uncertainties can be unambiguously reduced, at least using the neighbor-based HadAT approach or variants thereof. It remains unclear whether observed tropical tropospheric behavior is consistent with basic theory and the tightly constrained expectations of current climate models. Over the full period of radiosonde record, the estimates produced herein are in statistical agreement with model expectations all the way up to the tropical tropopause. Over the shorter satellite era, a discrepancy remains, particularly in the upper troposphere. Potential explanations range from the relatively uncontroversial involving residual observational errors (either at the surface or aloft), or statistical end-point effects, to more far-reaching reasons involving physical processes or forcings missing from some (known to be the case) or all climate models. The present analysis cannot provide definitive conclusions in this regard. However, the high degree of agreement over the 45 year radiosonde record provides a strong degree of confidence in overall climate model behavior in the tropics on the longest time scales.

[69] **Acknowledgments.** Many Met Office research staff allowed their computers to be used to complete the monthly and pentad ensembles over Christmas 2007 and Christmas 2008. Steve Sherwood, Leo Haimberger, and Thomas Peterson received within-UK travel costs from the Met Office under the Integrated Climate Program while undertaking portions of this work. Discussions with Matt Menne and Claude Williams of NOAA NCDC on a related project helped focus some of the work. Met Office authors were supported by the Joint DECC and Defra Integrated Climate Programme - DECC/Defra (GA01101). NCDC graphics team helped improve figure clarity.

References

Allen, R. J., and S. C. Sherwood (2008), Warming maximum in the tropical upper troposphere deduced from thermal winds, *Nat. Geosci.*, *1*, 399–403, doi:10.1038/ngeo208.

Bengtsson, L., and K. I. Hodges (2009), On the evaluation of temperature trends in the tropical troposphere, *Clim. Dyn.*, *36*, 419–430, doi:10.1007/s00382-009-0680-y.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: A new dataset from 1850, *J. Geophys. Res.*, *111*, D12106, doi:10.1029/2005JD006548.

Christy, J., R. Spencer, W. Norris, W. Braswell, and D. Parker (2003), Error estimates of version 5.0 of MSU-AMSU bulk atmospheric temperatures, *J. Atmos. Oceanic Technol.*, *20*, 613–629, doi:10.1175/1520-0426(2003)20<613:EEOVOM>2.0.CO;2.

Cleveland, W. S. (1994), *The Elements of Graphing Data*, pp. 139–142, AT&T Bell Lab., Murray Hill, N. J.

Douglass, D. H., J. R. Christy, B. D. Pearson, and S. F. Singer (2008), A comparison of tropical temperature trends with model predictions, *Int. J. Climatol.*, *28*, 1693–1701, doi:10.1002/joc.1651.

Durre, I., R. S. Vose, and D. B. Wuertz (2006), Overview of the Integrated Global Radiosonde Archive, *J. Clim.*, *19*, 53–68, doi:10.1175/JCLI3594.1.

Free, M., D. J. Seidel, J. K. Angell, J. Lanzante, I. Durre, and T. C. Peterson (2005), Radiosonde atmospheric temperature products for assessing climate (RATPAC): A new data set of large-area anomaly time series, *J. Geophys. Res.*, *110*, D22101, doi:10.1029/2005JD006169.

Fu, Q., C. M. Johanson, S. G. Warren, and D. J. Seidel (2004), Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends, *Nature*, *429*, 55–58, doi:10.1038/nature02524.

Gaffen, D. J. (1993), Historical changes in radiosonde instruments and practices: Instruments and observing methods, *Rep. 50*, World Meteorol. Organ., Geneva, Switzerland.

Gaffen, D. J., M. A. Sargent, R. E. Habermann, and J. R. Lanzante (2000), Sensitivity of tropospheric and stratospheric temperature trends to radiosonde data quality, *J. Clim.*, *13*, 1776–1796, doi:10.1175/1520-0442(2000)013<1776:SOTAST>2.0.CO;2.

Haimberger, L., C. Tavolato, and S. Sperka (2008), Towards elimination of the warm bias in historic radiosonde records—Some new results from a comprehensive intercomparison of upper air data, *J. Clim.*, *21*, 4587–4606, doi:10.1175/2008JCLI1929.1.

Immler, F., J. Dykema, T. Gardiner, D. N. Whiteman, P. W. Thorne, and H. Vömel (2010), A guide for upper-air reference measurements: Guidance for developing GRUAN data products, *Atmos. Meas. Tech.*, *3*, 1217–1231, doi:10.5194/amt-3-1217-2010.

Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Karl, T. R., S. J. Hassol, C. D. Miller, and W. L. Murray (Eds.) (2006), *Temperature Trends in the Lower Atmosphere: Steps for Understanding*

and Reconciling Differences, 164 pp., U.S. Clim. Change Sci. Program, Washington, D. C.

Keeling, C. D., R. B. Bacastow, A. E. Bainbridge, C. A. Ekdahl, P. R. Guenther, and L. S. Waterman (1976), Atmospheric carbon dioxide variations at Mauna Loa Observatory, Hawaii, *Tellus*, *28*, 538–551, doi:10.1111/j.2153-3490.1976.tb00701.x.

Kennedy, J., N. A. Rayner, R. Smith, D. E. Parker, and M. Saunby (2011a), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties, *J. Geophys. Res.*, doi:10.1029/2010JD015218, in press.

Kennedy, J., N. A. Rayner, R. Smith, D. E. Parker, and M. Saunby (2011b), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization, *J. Geophys. Res.*, doi:10.1029/2010JD015220, in press.

Klotzbach, P. J., R. A. Pielke Sr., R. A. Pielke Jr., J. R. Christy, and R. T. McNider (2009), An alternative explanation for differential temperature trends at the surface and in the lower troposphere, *J. Geophys. Res.*, *114*, D21102, doi:10.1029/2009JD011841.

Lanzante, J. R. (1996), Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data, *Int. J. Climatol.*, *16*, 1197–1226, doi:10.1002/(SICI)1097-0088(199611)16:11<1197::AID-JOC89>3.0.CO;2-L.

Lanzante, J. R., S. A. Klein, and D. J. Seidel (2003), Temporal homogenisation of monthly radiosonde temperature data, Part I: Methodology, *J. Clim.*, *16*, 224–240, doi:10.1175/1520-0442(2003)016<0224:THOMRT>2.0.CO;2.

McCarthy, M. P. (2008), Spatial sampling requirements for monitoring upper-air climate change with radiosondes, *Int. J. Climatol.*, *28*, 985–993, doi:10.1002/joc.1611.

McCarthy, M. P., H. A. Titchner, P. W. Thorne, S. F. B. Tett, L. Haimberger, and D. E. Parker (2008), Assessing bias and uncertainty in the HadAT adjusted radiosonde climate record, *J. Clim.*, *21*, 817–832, doi:10.1175/2007JCLI1733.1.

McCarthy, M. P., P. W. Thorne, and H. A. Titchner (2009), An analysis of tropospheric humidity trends from radiosondes, *J. Clim.*, *22*, 5820–5838, doi:10.1175/2009JCLI2879.1.

Mears, C. A., and F. J. Wentz (2009a), Construction of the remote sensing systems V3.2 atmospheric temperature records from the MSU and AMSU microwave sounders, *J. Atmos. Oceanic Technol.*, *26*, 1040–1056, doi:10.1175/2008JTECHA1176.1.

Mears, C. A., and F. J. Wentz (2009b), Construction of the RSS V3.2 lower tropospheric temperature dataset from the MSU and AMSU microwave sounders, *J. Atmos. Oceanic Technol.*, *26*, 1493–1509, doi:10.1175/2009JTECHA1237.1.

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

Menne, M. J., C. N. Williams, and R. S. Vose (2009), The U.S. Historical Climatology Network monthly temperature data, version 2, *Bull. Am. Meteorol. Soc.*, *90*, 993–1007, doi:10.1175/2008BAMS2613.1.

National Research Council Panel on Reconciling Temperature Observations (2000), *Reconciling Observations of Global Temperature Change*, Natl. Acad. Press, Washington, D. C.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton (2000), The impact of new physical parametrizations in the Hadley Centre climate model—HadAM3, *Clim. Dyn.*, *16*, 123–146, doi:10.1007/s003820050009.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed., pp. 617–622, Cambridge Univ. Press, Cambridge, U. K.

Randel, W. J., and F. Wu (2006), Biases in stratospheric and tropospheric temperature trends derived from historical radiosonde data, *J. Clim.*, *19*, 2094–2104, doi:10.1175/JCLI3717.1.

Santer, B. D., T. M. L. Wigley, J. S. Boyle, D. J. Gaffen, J. J. Hnilo, D. Nychka, D. E. Parker, and K. E. Taylor (2000), Statistical significance of trends and trend differences in layer-average atmospheric temperature time series, *J. Geophys. Res.*, *105*(D6), 7337–7356, doi:10.1029/1999JD901105.

Santer, B. D., et al. (2005), Amplification of surface temperature trends and variability in the tropical atmosphere, *Science*, *309*, 1551–1556, doi:10.1126/science.1114867.

Santer, B. D., et al. (2008), Consistency of modelled and observed temperature trends in the tropical troposphere, *Int. J. Climatol.*, *28*, 1703–1722, doi:10.1002/joc.1756.

Seidel, D. J., and J. R. Lanzante (2004), An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes, *J. Geophys. Res.*, *109*, D14108, doi:10.1029/2003JD004414.

Seidel, D. J., et al. (2004), Uncertainty in signals of large-scale climate variations in radiosonde and satellite upper-air temperature datasets, *J. Clim.*, *17*, 2225–2240, doi:10.1175/1520-0442(2004)017<2225:UISOLC>2.0.CO;2.

Sherwood, S. C. (2007), Simultaneous detection of climate change and observing biases in a network with incomplete sampling, *J. Clim.*, *20*, 4047–4062, doi:10.1175/JCLI4215.1.

Sherwood, S. C., J. Lanzante, and C. Meyer (2005), Radiosonde daytime biases and late 20th Century warming, *Science*, *309*, 1556–1559, doi:10.1126/science.1115640.

Sherwood, S. C., C. L. Meyer, R. J. Allen, and H. A. Titchner (2008), Robust tropospheric warming revealed by iteratively homogenized radiosonde data, *J. Clim.*, *21*, 5336–5352, doi:10.1175/2008JCLI2320.1.

Tett, S. F. B., R. Betts, T. J. Crowley, J. Gregory, T. C. Johns, A. Jones, T. J. Osborn, E. Ostrom, D. L. Roberts, and M. J. Woodage (2006), The impact of natural and anthropogenic forcings on climate and hydrology since 1550, *Clim. Dyn.*, *28*, 3–34, doi:10.1007/s00382-006-0165-1.

Thorne, P. W. (2008), The answer is blowing in the wind, *Nat. Geosci.*, *1*, 347–348, doi:10.1038/ngeo209.

Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan (2005a), Revisiting radiosonde upper air temperatures from 1958 to 2002, *J. Geophys. Res.*, *110*, D18105, doi:10.1029/2004JD005753.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears (2005b), Uncertainties in climate trends: Lessons from upper-air temperature records, *Bull. Am. Meteorol. Soc.*, *86*, 1437–1442, doi:10.1175/BAMS-86-10-1437.

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?, *Geophys. Res. Lett.*, *34*, L16702, doi:10.1029/2007GL029875.

Thorne, P. W., J. R. Lanzante, T. C. Peterson, D. J. Seidel, and K. P. Shine (2011), Tropospheric temperature trends: History of an ongoing controversy, *Wiley Interdisciplinary Rev. Clim. Change*, *2*, 66–88, doi:10.1002/wcc.80.

Titchner, H. A., P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, and D. E. Parker (2009), Critically assessing tropospheric temperature trends from radiosondes using realistic validation experiments, *J. Clim.*, *22*, 465–485, doi:10.1175/2008JCLI2419.1.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Q. J. R. Meteorol. Soc.*, *131*, 2961–3012.

Vinnikov, K. Y., N. C. Grody, A. Robock, R. J. Stouffer, P. D. Jones, and M. D. Goldberg (2006), Temperature trends at the surface and in the troposphere, *J. Geophys. Res.*, *111*, D03106, doi:10.1029/2005JD006392.

Zou, C. Z., M. D. Goldberg, Z. Cheng, N. C. Grody, J. T. Sullivan, G. Cao, and D. Tarpley (2006), Recalibration of microwave sounding unit for climate studies using simultaneous nadir overpasses, *J. Geophys. Res.*, *111*, D19114, doi:10.1029/2005JD006798.

P. Brohan, D. R. Fereday, J. J. Kennedy, M. P. McCarthy, D. E. Parker, and H. A. Titchner, Met Office Hadley Centre, Exeter EX1 3PB, UK.

L. Haimberger, Department for Meteorology and Geophysics, University of Vienna, A-1090 Vienna, Austria.

T. C. Peterson, NOAA National Climatic Data Center, Asheville, NC 28801, USA.

B. J. Santer, Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA.

S. C. Sherwood, Climate Change Research Centre, University of New South Wales, Level 4, Matthews Building, Sydney, NSW 2052, Australia.

S. F. B. Tett, School of Geosciences, University of Edinburgh, Edinburgh EH9 3JW, UK.

P. Thorne, Cooperative Institute for Climate and Satellites, 151 Patton Ave., Asheville, NC 28801, USA. (Peter.Thorne@noaa.gov)