Critically Reassessing Tropospheric Temperature Trends from Radiosondes Using Realistic Validation Experiments

Critically Reassessing Tropospheric Temperature Trends from Radiosondes Using Realistic Validation Experiments

HOLLY A. TITCHNER, P. W. THORNE, AND M. P. MCCARTHY

Met Office Hadley Centre, Exeter, United Kingdom

S. F. B. TETT

School of Geosciences, University of Edinburgh, Edinburgh, United Kingdom

L. HAIMBERGER

Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria

D. E. PARKER

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 8 January 2008, in final form 5 June 2008)

ABSTRACT

Biases and uncertainties in large-scale radiosonde temperature trends in the troposphere are critically reassessed. Realistic validation experiments are performed on an automatic radiosonde homogenization system by applying it to climate model data with four distinct sets of simulated breakpoint profiles. Knowledge of the ''truth'' permits a critical assessment of the ability of the system to recover the large-scale trends and a reinterpretation of the results when applied to the real observations.

The homogenization system consistently reduces the bias in the daytime tropical, global, and Northern Hemisphere (NH) extratropical trends but underestimates the full magnitude of the bias. Southern Hemisphere (SH) extratropical and all nighttime trends were less well adjusted owing to the sparsity of stations. The ability to recover the trends is dependent on the underlying error structure, and the true trend does not necessarily lie within the range of estimates. The implications are that tropical tropospheric trends in the unadjusted daytime radiosonde observations, and in many current upper-air datasets, are biased cold, but the degree of this bias cannot be robustly quantified. Therefore, remaining biases in the radiosonde temperature record may account for the apparent tropical lapse rate discrepancy between radiosonde data and climate models. Furthermore, the authors find that the unadjusted global and NH extratropical tropospheric trends are biased cold in the daytime radiosonde observations.

Finally, observing system experiments show that, if the Global Climate Observing System (GCOS) Upper Air Network (GUAN) were to make climate quality observations adhering to the GCOS monitoring principles, then one would be able to constrain the uncertainties in trends at a more comprehensive set of stations. This reaffirms the importance of running GUAN under the GCOS monitoring principles.

1. Introduction

There has been much debate surrounding tropical tropospheric temperatures since the first attempt to create a satellite-based climate dataset (Spencer and Christy 1990). Climate models predict amplification of the observed tropical warming trends at the surface (Santer et al. 2005; Karl et al. 2006), with maximum warming rates expected in the middle and upper troposphere. The observations are currently inadequately characterized to statistically robustly inform on this issue (Santer et al. 2008). The Remote Sensing Systems (RSS) (Mears and Wentz 2005), the University of Maryland (Vinnikov et al. 2006), and the National Environmental Satellite Data and Information Service (Zou et al. 2006)

Corresponding author address: Holly Titchner, Met Office Hadley Centre, FitzRoy Road, Exeter, Devon EX1 3PB, United Kingdom.
E-mail: holly.titchner@metoffice.gov.uk

DOI: 10.1175/2008JCLI2419.1

Microwave Sounding Unit (MSU) datasets all yield trends that are more or less consistent with model predictions. So do some more recent radiosonde temperature datasets (Sherwood et al. 2008a; Haimberger et al. 2008) and temperatures inferred from radiosonde winds (Allen and Sherwood 2008). However, other recently produced radiosonde datasets (Haimberger 2007; Thorne et al. 2005b; Free et al. 2005) and the University of Alabama in Huntsville (UAH) (Christy and Norris 2006) MSU dataset have all reported less warming aloft than expected since 1979 (Karl et al. 2006). Here we aim to robustly reassess the uncertainty in the manually homogenized Met Office Hadley Centre radiosonde temperature dataset (HadAT) (Thorne et al. 2005b), which should better inform where the truth lies.

There have been numerous changes to the global radiosonde observing network throughout the last few decades, many of which have been poorly documented. This has resulted in many sudden changes (inhomogeneities or breakpoints) within the long-term time series. The challenge is to remove these breakpoints and recover the large-scale trends, which are small relative to both the natural variability and the magnitude of many of the identified breakpoints. However, our ability to do this is highly dependent on the decisions made during homogenization (Thorne et al. 2005a). The resulting structural uncertainty is reflected in the different trend estimates produced by the existing datasets (Free and Seidel 2005; Karl et al. 2006).

McCarthy et al. (2008) developed an automated system, adapted from the manual HadAT dataset (Thorne et al. 2005b), that attempts to homogenize a radiosonde dataset using neighbor-based iterative breakpoint identification and adjustment. The homogenization is controlled by a number of system parameters that can be set to different values, akin to making different methodological decisions during the dataset development. The system can be used to output a large ensemble of different dataset realizations. McCarthy et al. (2008) used a very simple validation ensemble and found that, when trends are systematically biased, many experiments did not fully recover the true large-scale trend.

In this study we develop four error (or breakpoint) models based on different, much more complex, assumptions in order to produce breakpoint profiles that could exist in the real world. This extends the idealized experiments performed by McCarthy et al. (2008). The four different error models are applied to homogeneous third Hadley Centre Atmospheric Model (HadAM3) (Pope et al. 2000) climate simulation data from a run with prescribed historical sea surface temperatures and natural and anthropogenic forcings. Temporal and spatial sampling characteristics of the daytime and nighttime observed radiosonde data are imposed on the model data. The resulting heterogeneous data are passed through the automatic homogenization system, and a population of 100 realizations is produced by varying the homogenization system parameters in a series of experiments. The only difference between each of the 100 experiments therefore relates to the homogenization method used. The same 100 experimental setups are used for each of the daytime and nighttime error model input datasets as well as the observations. Knowing the original model ''truth'' we can assess the ability of the system to recover the large-scale trends from heterogeneous data. These validation experiments permit a critical reappraisal of the trends produced when the system is applied to the real observations, enabling us to make inferences about the real world trends. We also run some observing system experiments using the error models in order to assess the impact of future possible changes to the radiosonde network and possible avenues to improve our knowledge of historical data. Although our homogenization experiments are performed on levels between 850 and 30 hPa, we focus on the troposphere, particularly on the tropics. Our main aim is to assess whether the apparent tropical tropospheric lapse rate discrepancy, supported by many, but not all, currently available radiosonde datasets, could be due to uncertainty in the radiosonde records.

2. Input data

a. Radiosonde data

The radiosonde observations used within this study were derived from the raw daily data that were input into the Radiosonde Observation Correction Using Reanalyses (RAOBCORE) dataset (Haimberger 2007). This dataset is a merge of radiosonde ingest to the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) (Uppala et al. 2005) and the Integrated Global Radiosonde Archive (IGRA) (Durre et al. 2006), with preference given to ERA-40 ingest data. Data at 12 standard pressure levels (850, 700, 500, 400, 300, 250, 200, 150, 100, 70, 50, and 30 hPa) were used from 1958 to 2003.

Daytime and nighttime observations were separated as they have different biases (e.g., solar biases in the daytime, Sherwood et al. 2005) as illustrated by the WMO radiosonde intercomparison campaigns (Nash et al. 2005 and references therein). A simple timing criterion was used, counting 90°E–90°W as daytime for 1200 UTC and nighttime for 0000 UTC, and vice versa for all other longitudes. Only stations between 70°N and 70°S

were included, to limit the seasonality of polar day and night. We excluded Indian stations, which have previously been found to be problematic (Thorne et al. 2005b; Lanzante et al. 2003; Parker et al. 1997). Seasonal anomalies were calculated relative to 1981–2000 climatology. This increased the coverage by 28 stations in the tropics (20°S–20°N) during the satellite era, which is the focus of this paper, compared with using 1966–95 climatology, as done by McCarthy et al. (2008) and Thorne et al. (2005b) to maximize the global coverage for the full period. The resulting daytime dataset contained a total of 586 (79 in the tropics) stations and the nighttime dataset contained 513 (29 in the tropics) stations (Fig. 1a).

We also use metadata documenting known changes of instruments and observing practices. These came from the IGRA dataset [Gaffen(1996) and subsequent updates]. While these metadata provide valuable information regarding the timing of potential breaks, they are often incomplete. Around 70% of identified breakpoints in the HadAT (Thorne et al. 2005b) and RAOBCORE (Haimberger 2007) datasets had no known metadata events associated with them. Many of these breaks were large and very likely arose from unrecorded changes at the stations, rather than from false breakpoint identifications.

b. Simulated data

We perform validation experiments using simulated data from HadAM3 (Pope et al. 2000), forced with observed sea surface temperature and sea ice distributions from the Hadley Centre Sea Ice and Sea Surface Temperature (HadISST) dataset (Rayner et al. 2003) and prescribed anthropogenic and natural external forcings (Tett et al. 2007). The advantage of using these data, instead of simply producing a randomly generated series, is that they contain a representation of real variations in the climate (such as ENSO) that may be expected to interact with the ability of any homogenization system to identify and adjust for breaks.

The monthly model data were available at 10 of the 12 radiosonde observation pressure levels (all except 70 and 30 hPa) on a 2.5° latitude by 3.75° longitude grid. Data from the nearest grid box to each station were extracted. The data were subsampled twice, once using the daytime observation coverage and once using the nighttime coverage, to produce two datasets. Seasonal anomalies were created for each station relative to a 1981–2000 climatology. Random noise with a Gaussian distribution and standard deviation of half that of the model gridbox series was added to approximate sampling effects and to ensure that no two station series arising from the same grid box would contain exactly the same data. The two resulting model datasets provide homogeneous records with the same spatial and temporal sampling as the daytime and nighttime observational datasets. We refer to these as control datasets.

3. Methods

a. Homogenization system

The homogenization system developed by McCarthy et al. (2008) uses an iterative neighbor-based breakpoint identification and adjustment technique similar to that employed in the development of the HadAT radiosonde dataset (Thorne et al. 2005b). Reference anomaly series are generated as weighted composites of neighboring stations, with weightings derived from temperature correlation coefficients calculated using either National Centers for Environmental Prediction (NCEP) (Kalnay et al. 1996) or ERA-40 (Uppala et al. 2005) reanalyses since 1979. A station minus neighbor difference series is then calculated, which is intended to remove the majority of the natural climatic variations and large-scale trends and emphasize nonclimatic change points. The success of this depends upon how well errors in the neighbor series cancel when averaged together.

The Kolomogorov–Smirnov (K–S) test (Press et al. 1992), a nonparametric statistical homogeneity test, is used to identify breakpoints in the station minus neighbors difference series. Pressure levels are considered in unison during the breakpoint identification, as the breakpoints are assumed to affect multiple (although not necessarily all) levels. Information regarding the sign of the potential breakpoints is not used at this stage. The statistical breakpoint test result series is combined with information based on the metadata. The metadata therefore provide additional evidence but are not usually crucial for the identification of a breakpoint. A critical threshold is used to assign breakpoints within this combined series. See Fig. 1 in McCarthy et al. (2008) for an example of this breakpoint identification method.

Adjustments are calculated for all assigned breakpoints at all levels, taking the difference between the median values of the neighbor difference series before and after the breakpoint. The process is iterative: the critical threshold for identifying a breakpoint is relaxed between each iteration. Therefore we should identify and adjust only the most severe breakpoints in the early iterations. With each iteration the neighbor series should improve, as well as the station series itself. The system is likely to perform best in areas with a high

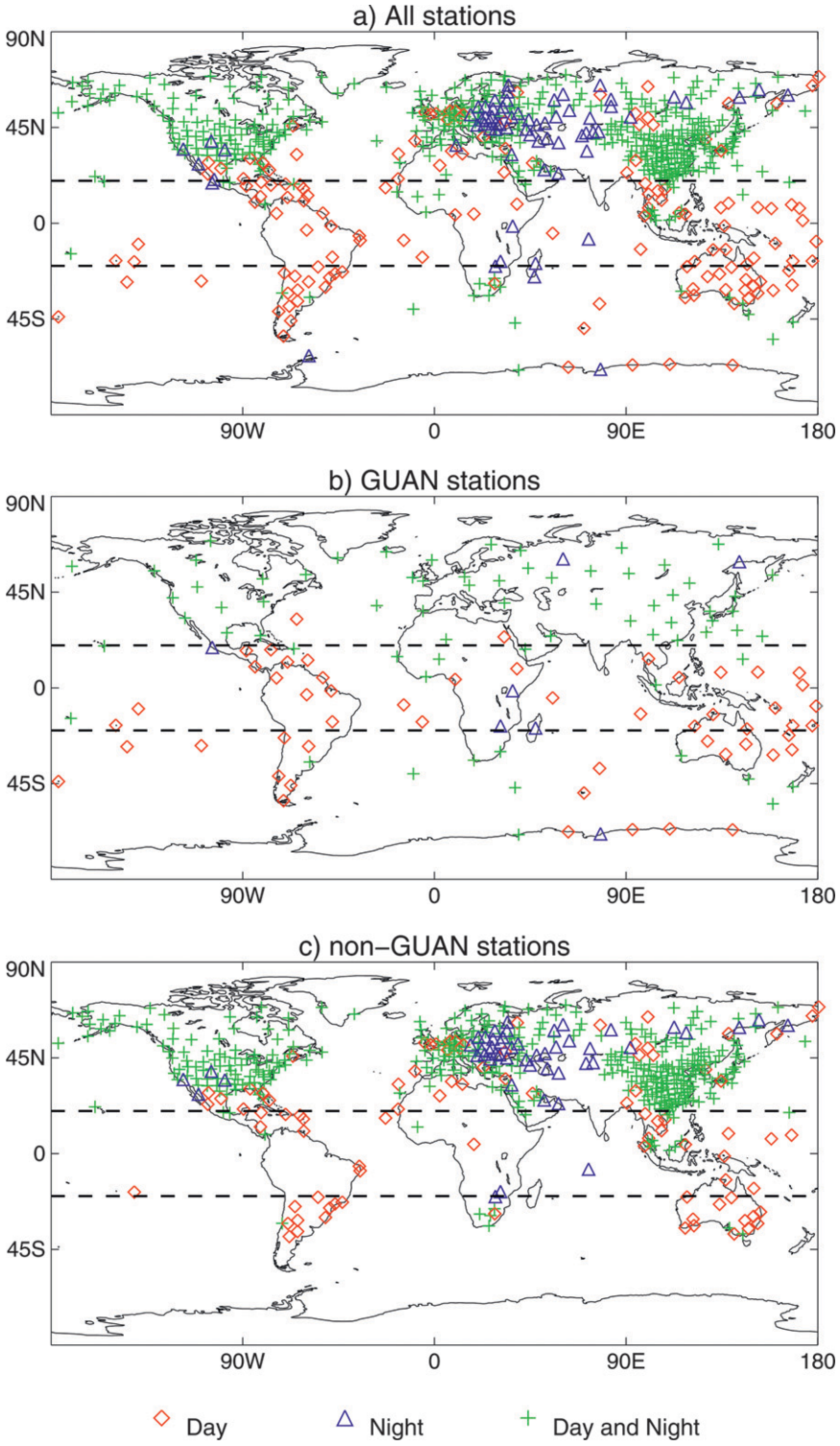

FIG. 1. Station coverage for (a) all stations, (b) GUAN stations, and (c) non-GUAN stations. Stations that contain both daytime and nighttime data are denoted by green crosses, whereas daytime-only stations are denoted by red diamonds, and nighttime-only stations are denoted by blue triangles.

TABLE 1. The total number of breakpoints applied within each error model and a summary of the breakpoint sizes applied at 500 hPa and 50 hPa.

Error model	Total number of breaks (average per station)	Time of day	Median bias at 500 hPa (K)	Median absolute bias at 500 hPa (K)	Standard deviation of biases at 500 hPa (K)	Median bias at 50 hPa (K)	Median absolute bias at 50 hPa (K)	Standard deviation of biases at 50 hPa (K)
Current understanding	5232 (8)	Day	−0.06	0.35	0.46	−0.23	0.48	0.64
		Night	−0.05	0.32	0.49	−0.10	0.49	0.64
Many small breaks	9810 (15)	Day	0.00	0.14	0.28	−0.01	0.18	0.38
		Night	0.00	0.12	0.24	−0.02	0.17	0.38
Removal of signal	4578 (7)	Day	−0.06	0.33	0.49	0.17	0.43	0.65
		Night	−0.06	0.30	0.44	0.03	0.42	0.63
Few large breaks	2616 (4)	Day	−0.10	0.62	0.97	−0.20	0.90	1.34
		Night	0.09	0.71	1.14	0.25	0.93	1.28

density of stations that do not have contemporaneous breaks. There are 14 tunable parameters within the system (appendix A of McCarthy et al. 2008), which affect the location and timing of the breakpoints identified and how the adjustments are calculated. For a more detailed description of the system, parameters, and limitations see section 3 of McCarthy et al. (2008).

Once the homogenization is completed, the anomalies are averaged onto a 5° latitude by 10° longitude grid, as in HadAT (Thorne et al. 2005b). They are then vertically weighted to replicate lower-tropospheric T2LT (Karl et al. 2006) temperature anomalies measured by MSU and to allow dataset comparisons (see section 5). Static MSU weighting functions have been provided by the University of Alabama in Huntsville. Latitude bands are averaged and cos(lat) weighted to calculate global, tropical (20°S–20°N), Northern Hemisphere (NH) extratropical (20°–70°N), and Southern Hemisphere (SH) extratropical (20°–70°S) mean time series. Linear trends for the satellite era were estimated using the median of pairwise slopes method (Lanzante 1996) to minimize the effect of outliers. It is important to note that the true time series behavior is not necessarily linear and that alternative time series descriptors may be equally valid (Seidel and Lanzante 2004; Thorne et al. 2005b).

b. Derivation of error models

To rigorously test the homogenization system we apply artificial breakpoint profiles to the daytime and nighttime control datasets from the climate model. The numbers, dates, and profiles of the breakpoints varied between four different error models (Table 1). Each error model was based on different assumptions regarding the size, distribution, etc., of the breakpoints (see appendix A for more details). They were applied at the same dates in the daytime and nighttime datasets, although different breakpoint profiles were used for these two datasets. This is consistent with published results from radiosonde intercomparisons (Nash et al. 2005) that yield different error structures for day and night.

Figure 2 shows the distribution of daytime breakpoint sizes for each error model at 50 hPa. A positive (negative) breakpoint is said to occur when there is an increase (decrease) in the mean of the later part time series compared to the mean in the earlier part of the time series. Although we will later concentrate on the troposphere, we chose to illustrate the breakpoints applied at 50 hPa because they are larger than in the troposphere, as is also strongly believed to be the case in the observations (Karl et al. 2006), and hence the systematic biases can be more easily seen. The main features of each error model are summarized below.

- *Current understanding*. This error model is our current understanding based upon existing literature of the breakpoints that afflict the observed temperature record in the radiosonde network. The breakpoints above 500 hPa have a negative mean (Table 1, Fig. 2a) in order to produce a cooling bias in the long-term trends. Breakpoints in the daytime dataset have an additional negative offset, as McCarthy et al. (2008) and Sherwood et al. (2005) found in observations.
- *Many small breakpoints*. Although this error model does contain some large breakpoints, it contains a large number of smaller ones (Table 1, Fig. 2b). These small breakpoints have only a very small systematic bias.
- *Removal of signal*. Breakpoints at or below 150 hPa have a negative offset and those above have a positive offset (Table 1, Fig. 2c) similar in magnitude to the temperature trend in the model. The net result is

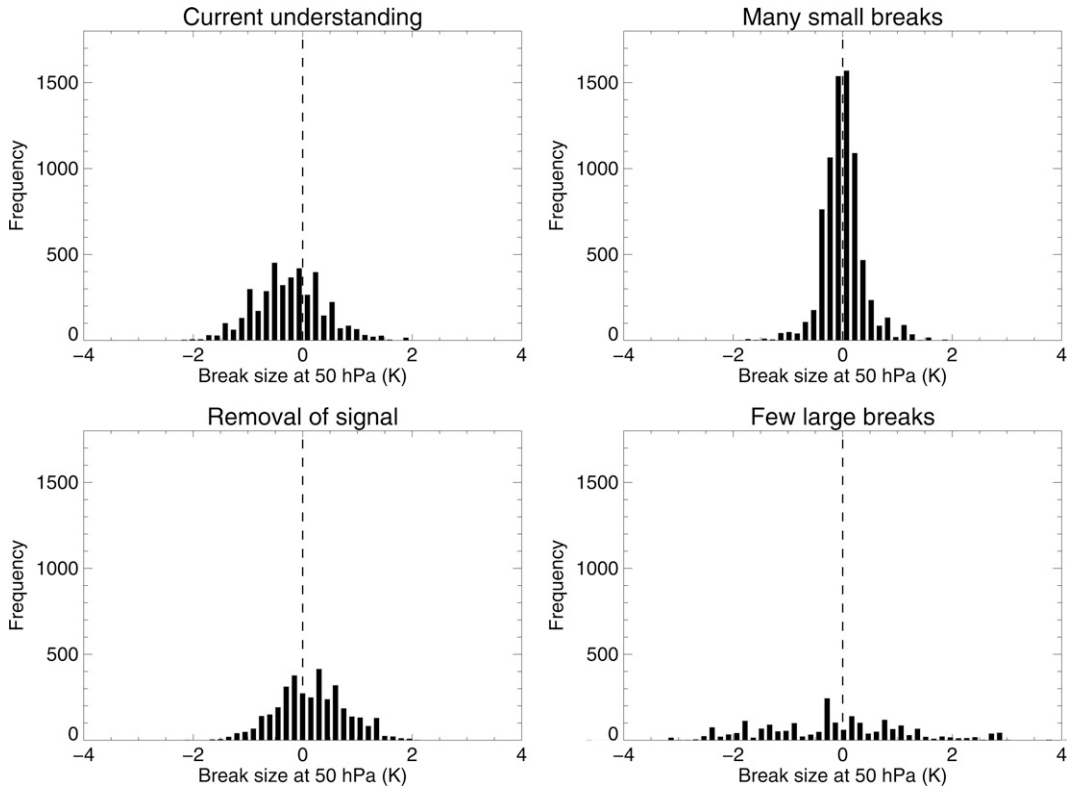

FIG. 2. Breakpoint distributions at 50 hPa for each daytime error model. A positive (negative) breakpoint is said to occur when there is an increase (decrease) in the mean of the later part time series compared to the mean of the earlier part of the time series. The number of breakpoints and the shape of the distributions vary between error models. The distribution also varies with height, with lower levels showing a similar shape but with a smaller absolute spread. The zero line is marked by the vertical dashed line, which highlights the negative bias in the *current understanding* error model and the positive bias in the *removal of signal* error model, at this level.

to remove most of the underlying climate change signal.

- *Few large breakpoints*. This error model contains fewer breakpoints than the other error models but with a larger standard deviation, resulting in a higher proportion of large breakpoints (Table 1, Fig. 2d).

Except in the case of the *current understanding* error model, where daytime breakpoints are given a negative offset in comparison to nighttime breakpoints, the differences between the daytime and nighttime error structures only arise from random differences in the generation of the breakpoint profiles. The differences in median biases (Table 1) are more evident at 50 hPa where the sample is smaller, as in the real observations. The main difference between the daytime and nighttime datasets is that nighttime has a much poorer spatial coverage, particularly outside the NH extratropics (Fig. 1a).

The error models can be used to assess the skill of a homogenization method on data containing different error structures, as the breakpoint locations and magnitudes, and climate change signal are known. They are also more complex than idealized test cases employed in many earlier tests of homogenization methods (e.g., McCarthy et al. 2008; Haimberger 2007; Sherwood et al. 2008a) and allow for a more comprehensive understanding of the trend uncertainties. The four error models used within this study were deliberately designed to be as different as possible so that the system is not tuned toward a given set of assumptions. They span a range of possible error structures, all of which may contain at least some characteristics (e.g., phasing, bias, magnitude, clustering) that exist within the true observations.

c. *Homogenization ensembles*

We perform an ensemble of 100 homogenization experiments on each dataset by randomly varying the 14 tunable system parameters listed in appendix B. Each homogenized output represents a different set of methodological choices and can be used to investigate uncertainty in trends resulting from the homogenization

(McCarthy et al. 2008). The same 100 random parameter configurations were used for each ensemble so that the only difference between each ensemble was the input dataset. The variations within a given ensemble are therefore entirely due to differences in the tuning of the homogenization method. Ensembles were created for each of the daytime and nighttime datasets: the observations, the control datasets, and the four error models. Ensembles using the original control datasets were produced simply to assess the impact of homogenizing breakpoint-free data. It is important to ascertain whether the system significantly alters these data as that would clearly be an undesirable characteristic.

d. Homogenization skill rankings

As well as assessing the absolute values of the trends produced from the homogenization system for error model ensembles, we also assess the relative skill within ensembles by ranking local error recovery and large-scale fidelity (see below). By comparing the rankings between each error model we can see whether particular experiments consistently do well (or poorly) and, therefore, determine how dependent the homogenization skill is on the underlying error structure. If there is little dependence, then we will be able to unambiguously tune our system toward a more reliable set of experiments and exclude the experiments we know to be ineffective. The two approaches we use to rank the experiments are summarized below.

- *Local skill*. Provides a measure of how well each experiment identifies breakpoint locations and magnitude. We compare the root-mean-square difference (RMSD) between the known error model time series and the control series for each station and level with the RMSD between the homogenized series and the control series. The values are summed over all stations and levels to yield a single value for each experiment within a given ensemble. The experiment values are then ranked between 1 and 100 (1 indicating closest agreement to the known error structure). A high local skill score therefore indicates an experiment that provides accurate local observations relative to the other experiments within the given ensemble.
- *Trend recovery*. Provides a measure of how successful each experiment is at capturing the ''true'' mean T2LT trend in the satellite era for a given large-scale region (global, tropical, NH extratropical, or SH extratropical). The experiments are ranked between 1 and 100 (1 indicating the best) based on how closely they estimate the trend in the unbiased control series.

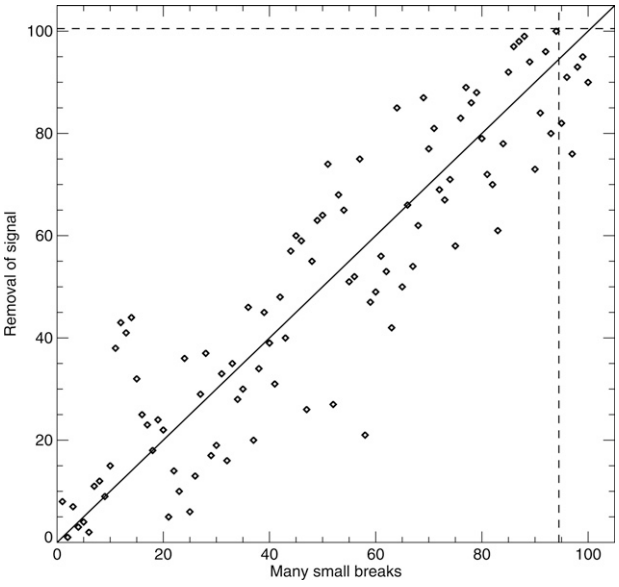

FIG. 3. A comparison of the local skill rankings for the *many small breaks* and the *removal of signal* daytime error models. Each point (100 in total) represents a homogenization experiment performed with a particular random parameter configuration using each of the error models. The clustering around the 1:1 line and the high correlation (0.90) of the ranks shows that if the homogenization system performed well (badly) for one error model, then it also performed well (badly) for the other. The dashed lines denote the rankings for each unhomogenized error model.

4. Error model results

a. Ranking results

Using the method outlined in section 3d, Fig. 3 compares the local skill rankings for the *many small break*s and *removal of signal* daytime ensembles. The clustering around the 1:1 line and the high correlation between the rankings show that there is good agreement as to how the homogenization system performs on these two error models. The correlations between the other error models for both daytime and nighttime were also found to be high (Table 2). This is encouraging and indicates that the ability of the system to reduce RMSD is not very dependent on the underlying error structure. A total of 35 (32) experiments consistently rank in the top half for the daytime (nighttime) error model ensembles. Assuming that the error models are uncorrelated, if the system showed no skill, it would effectively become a random number generator. Under this assumption the degree of agreement actually found is considerably more than would be expected from a set of four random number populations ($100 \times 0.5^4 = 6$). We also produced separate rankings for the troposphere only and found very similar results (not shown).

The absolute RMSD values reveal that nearly all ex-

TABLE 2. Rank correlations between local error recovery skill for each error model based on the day (lower-left-hand triangle) and night (italic, upper-right-hand triangle) RMSD. Correlations between the day and night RMSD rankings for each individual error model are given on the diagonal (bold).

	Current understanding	Many small breaks	Removal of signal	Few large breaks
Current understanding	**0.80**	*0.89*	*0.95*	*0.66*
Many small breaks	0.68	**0.97**	*0.86*	*0.48*
Removal of signal	0.85	0.90	**0.98**	*0.78*
Few large breaks	0.84	0.90	0.97	**0.80**

periments improve local skill relative to making no adjustments (Table 3). Any breakpoints identified and adjusted for in the control experiments are false, and so are detrimental to the local skill and trend. However, the RMSD values for the control ensembles were found to be an order of magnitude smaller than the RMSD values for error models. The benefit (in terms of RMSD, i.e., the accuracy of local observations) of adjusting data in the presence of biases is therefore much greater than the cost if no errors exist.

Figure 4 compares the tropical trend recovery rankings for the *current understanding* and *many small breaks* daytime ensembles. The correlations (Table 4) show that these rankings are less consistent than the local skill rankings. Therefore, skill in recovering large-scale trends depends on the underlying error structure. Also, an experiment that is relatively skillful at minimizing RMSD error is not necessarily skillful at recovering the large-scale trends (Sherwood et al. 2008b). However, there are still 17 experiments that consistently rank in the top half for all daytime error models, which is more than expected by chance. A total of 12 daytime experiments rank in the top half for both tropical trend recovery and local skill (only $100 \times 0.5^8 < 1$ is expected by chance), 7 of which also rank in the top half for the global and NH extratropical trend recovery. We refer to these seven experiments as our ''top'' experiments throughout the rest of this paper. SH extratropical results are ignored as there is poor agreement between the error model trend rankings in this region (only three experiments consistently rank in the top half using these rankings alone) owing to the sparsity of SH extratropical stations (Fig. 1a). Although the tropical stations are equally sparse, their neighbor correlation regions are larger, so the homogenization performs better there, as reflected in the better agreement between error model rankings. We only consider the best performing experiments for the daytime data as there is extremely poor agreement between the nighttime error models: only 7 consistently rank in the top half using the tropical trend recovery rankings alone, again owing to data sparsity.

An examination of the parameter settings for the ''top'' experiments did reveal that they all used an ''adaptive'' adjustment method (i.e., all adjustments are recalculated during every iteration unlike the ''non-adaptive'' adjustment method; see appendix A and section 3 of McCarthy et al. 2008 for more details). No other parameter settings stand out as being optimal. However, 100 experiments is not a large enough sample

TABLE 3. The number of homogenized series that have a lower RMSD or trend error than each unadjusted error model for day and night.

	RMSD		Trend	
Error model	Day	Night	Day	Night
Current understanding	98	98	88	21
Many small breaks	94	84	70	2
Removal of signal	100	100	93	38
Few large breaks	100	100	98	94

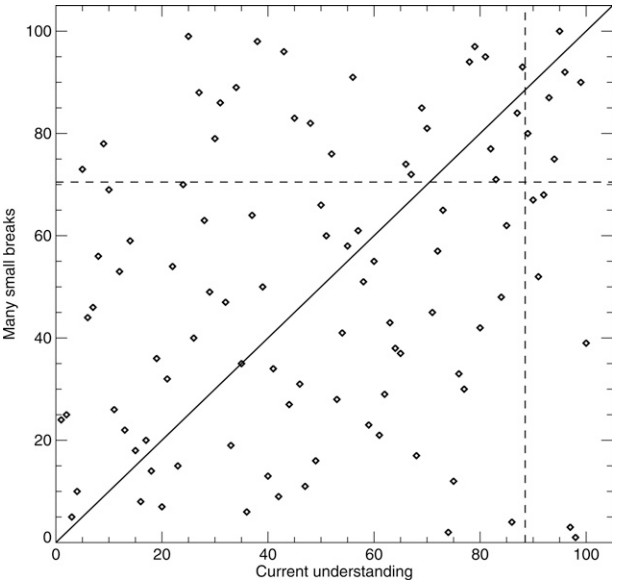

FIG. 4. As in Fig. 3 but using the tropical T2LT satellite-era trend recovery rankings for the *current understanding* and *many small breaks* daytime error models. The large departures from the 1:1 line and the low correlation (0.25) indicate that the relative tropical trend recovery skill depends on the underlying error structure.

TABLE 4. Rank correlations as in Table 2 except using the tropical T2LT trend recovery skill.

	Current understanding	Many small breaks	Removal of signal	Few large breaks
Current understanding	**0.02**	*−0.04*	*0.06*	*0.11*
Many small breaks	0.25	**0.12**	*0.25*	*−0.04*
Removal of signal	0.38	0.53	**0.14**	*−0.12*
Few large breaks	0.33	0.14	0.54	**0.25**

with purely random parameter setting choices to thoroughly investigate the effect of parameter values, as there are too many possible settings and many are likely to interact nonlinearly. A more detailed study with a larger number of experiments would be required to achieve this and is not the purpose of the current study.

b. Absolute trend results

We now consider the tropical T2LT satellite-era trend results from the daytime error model ensembles (Fig. 5a). By comparing the ensembles with the original control trend we can infer how well the system is likely to recover the true trend in the real observations. It can be seen that the spread in the daytime ensembles encompasses the original (control) trend for all error models except for the *removal of signal* case. Although the median trends underestimate the original trend, in all cases they reduce the trend bias. This is also reflected in the daytime results for trend error in Table 3, which indicates that most experiments improve the tropical trend recovery in comparison to each unadjusted error model. The underestimation of the bias is partly due to the known inclusion of poor homogenization experiments that are very conservative in the detection and adjustment of breakpoints and result in little or no change in trend compared to the unadjusted data. The trend in the unadjusted data does lie outside of or near the edge of the interquartile range for all error models. The medians of the top experiments are better at estimating the magnitude of the trend bias compared to the median of all experiments, and the original trend is almost recovered for the *few large breaks* error model, which is a priori the most tractable.

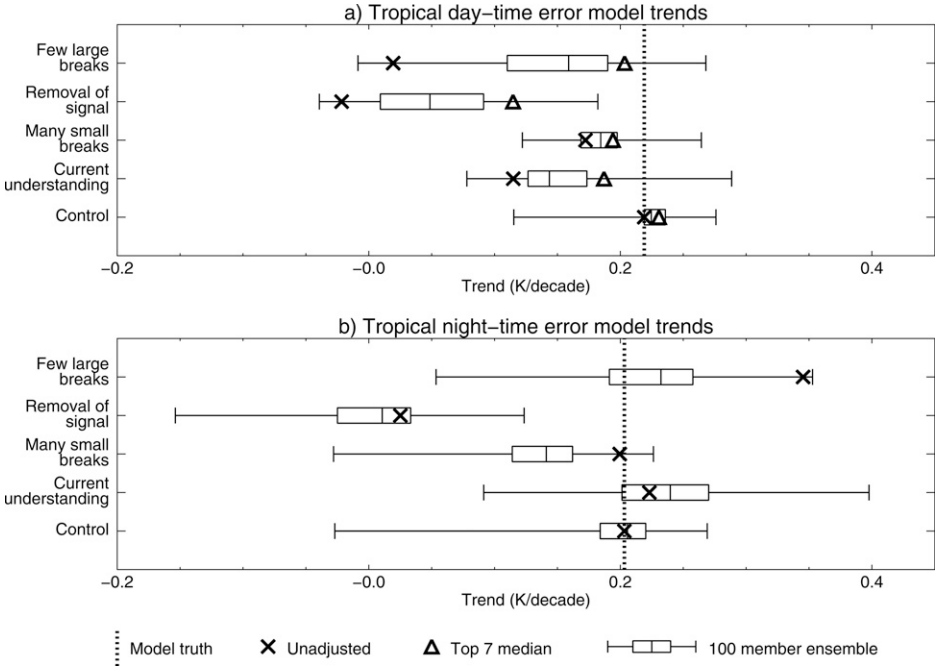

FIG. 5. Tropical (a) daytime and (b) nighttime T2LT equivalent trends for the satellite period for the four error models and the control. The black crosses denote the trends in the unadjusted error models and each box plot denotes the spread from the 100-member homogenized ensemble, where the central line gives the median trend, the box gives the interquartile range (25th to 75th percentiles) of the trends, and the whiskers give the full range of trends. The triangles denote the median trends from the 7 ''top'' experiments. The vertical dotted line shows the model ''truth.''

We infer that a median trend from a daytime ensemble is likely to correctly capture the sign of the trend bias, but is likely to underestimate the adjustment required (see also McCarthy et al. 2008). We can therefore infer the sign of the trend bias and place a strong lower limit on its magnitude from the median of the full ensemble. It is also likely that the true trend bias will be at least as large as the median of the top performing experiments, and possibly larger.

The nighttime tropical trend results are less encouraging, which is reflected in the larger spread of solutions (Fig. 5b). Thus, there is a larger dependence on the error structure. Adjusted trends from all error models except *removal of signal* do manage to encompass the original trend, but in many experiments there is a tendency for the adjustment to move the trend further from the truth. The bias in the median trend for all cases, except *few large breaks* (recall this is a priori the most tractable), is similar to or worse than the trend in the unadjusted data. These results are also seen in Table 3 and must be due to the data coverage, which we recall is much poorer than for the daytime data (Fig. 1a). We therefore have very little confidence that results from our system for the nighttime data in the tropics can provide a robust indication of the sign or magnitude of the systematic bias in the data.

Any breakpoints identified in the control experiments are false. In some control experiments this erroneously shifted the tropical T2LT trends up to 0.1 K decade^{-1} away from the already homogenous trends in the daytime data, and up to 0.2 K decade^{-1} away in the nighttime data. This highlights the risk of relying upon a single homogenization method for trend estimation. However, 96 (86) daytime (nighttime) experiments shifted the trends by less than 0.05 K, and the median trends changed by less than 3% from the original control trends. This suggests that false breakpoint detection is not likely to be the major failing of a majority of the homogenization members. Further analysis of the error model ensembles (not shown) supports this and revealed that the underestimation of the trend bias in the majority of experiments occurred mainly as a result of missed breakpoints or incorrect adjustments, and not from falsely identified breakpoints.

The global tropospheric results are similar to those for the tropics, although the spread in the trend estimates tends to be smaller (Figs. 6a and 7a). The same applies to other subregions (Figs. 6b,c and 7b,c), although the SH extratropical trends for the daytime *current understanding* error model (which are unbiased in the raw data) are shifted away from the truth, probably owing to sparsity of data. The NH extratropical nighttime data perform better than for the tropics and other subregions, as the median trend captures the right sign of the trend bias for all error models (albeit the median trend shifts very little for two of the error models). This is likely due to higher station density, better station management, and better quality metadata—all of which are important for the homogenization methodology.

5. Observation results

Figure 8 shows the tropical T2LT satellite-era trends for the adjusted observations produced from the same 100 experimental setups as those used for the error model ensembles. The unadjusted daytime trend is biased cold relative to the climate model expectation of 0.14–0.20 K decade^{-1} based upon amplification of surface trends (Santer et al. 2005). The median trend from the daytime ensemble shifts the trend closer, but the full spread, −0.01–0.14 K decade^{-1}, does not quite encompass the range of model estimates. The top experiments shift the median trend closer to the range based upon model expectation. The unadjusted daytime trend lies on the edge of the interquartile range, 0.03–0.07 K decade^{-1}. This is a relatively large spread compared to the daytime control ensemble, for example, which shifts the median trend very little.

The unadjusted nighttime trend, 0.19 K decade^{-1}, lies within the range of model expectation, although the median trend from the adjusted ensemble, 0.13 K decade^{-1}, is out of this range. The very large spread of the adjusted trends, 0.03–0.23 K decade^{-1}, even encompasses the raw daytime trend.

We can make inferences about these real world trends using the results from the model-based experiments, although we are unable to calculate exact probabilities. The pink bar in Fig. 8 represents the true trend implied from the findings. However, we must caveat that the results are based on four out of an infinite number of possible error structures (see section 7). The following statements can be made about the tropical tropospheric trends:

- The trend in the raw daytime observations (0.03 K decade^{-1}) is very likely biased cold (consistent with McCarthy et al. 2008, Randel and Wu 2006, and Sherwood et al. 2005), but we cannot comprehensively quantify the magnitude of the bias and its uncertainty.
- Our findings suggests that the true tropical trend is not only warmer than the median trend produced from the daytime ensemble (0.05 K decade^{-1}), but also as warm or warmer than the median trend from the top experiments (0.08 K decade^{-1}).

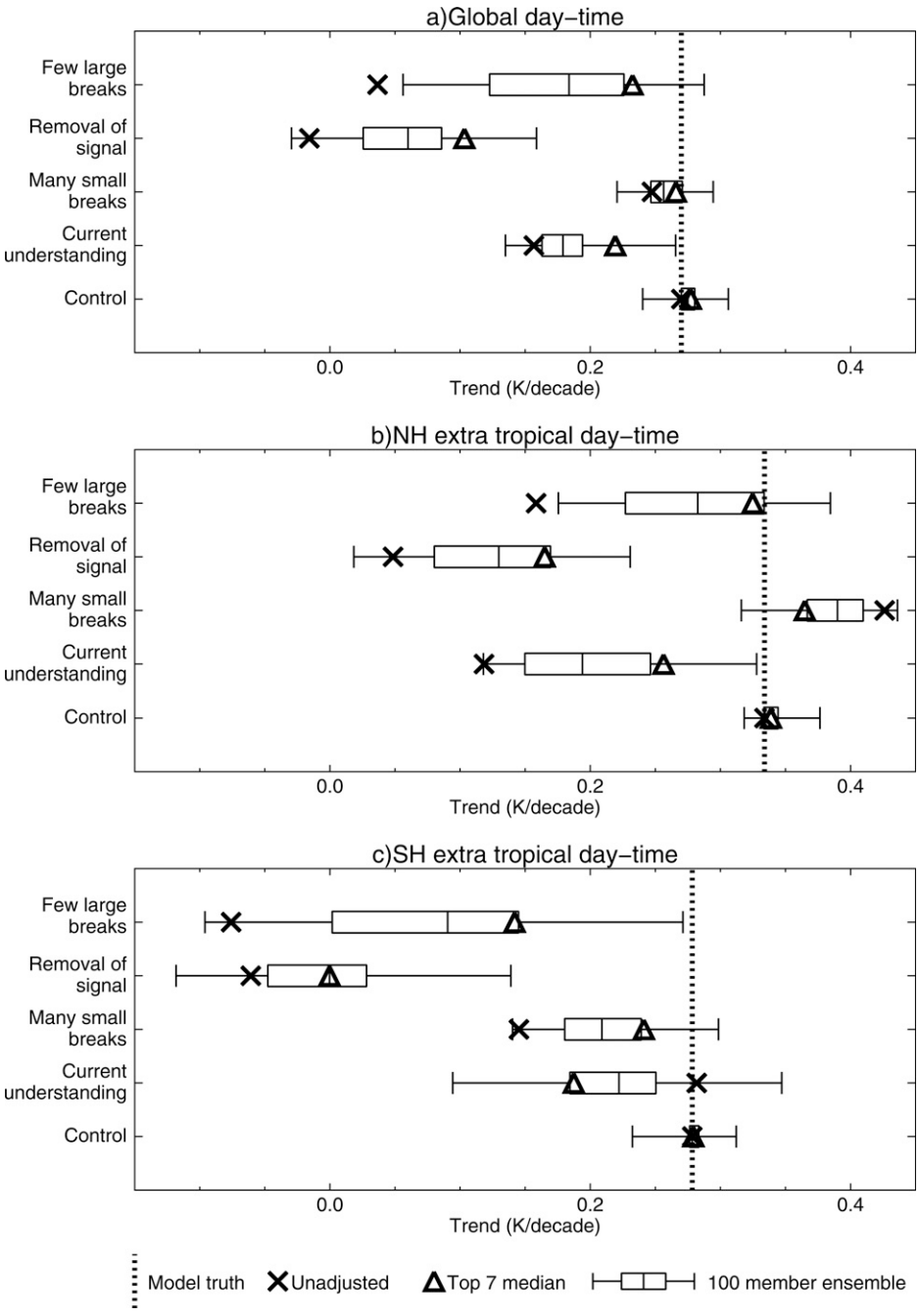

FIG. 6. As in Fig. 5 but for the (a) global, (b) NH extratropical, and (c) SH extratropical daytime trends.

- The tropical trends from many current datasets, including HadAT (0.05 K decade^{-1}), are likely to be biased cold.
- The true tropical trend may be warmer than the range estimated from the daytime ensemble. Therefore remaining biases may still account for the apparent tropical lapse rate discrepancy between the observations and climate models.

We now consider the T2LT satellite-era trends for the other large-scale regions (Fig. 9). The median ensemble trends indicate that the unadjusted daytime trends are biased cold globally and in the NH extratropics. In both cases the unadjusted trends also lie outside the interquartile range. We note that the median trends from the global nighttime ensembles are only able to capture the correct sign of the bias for two

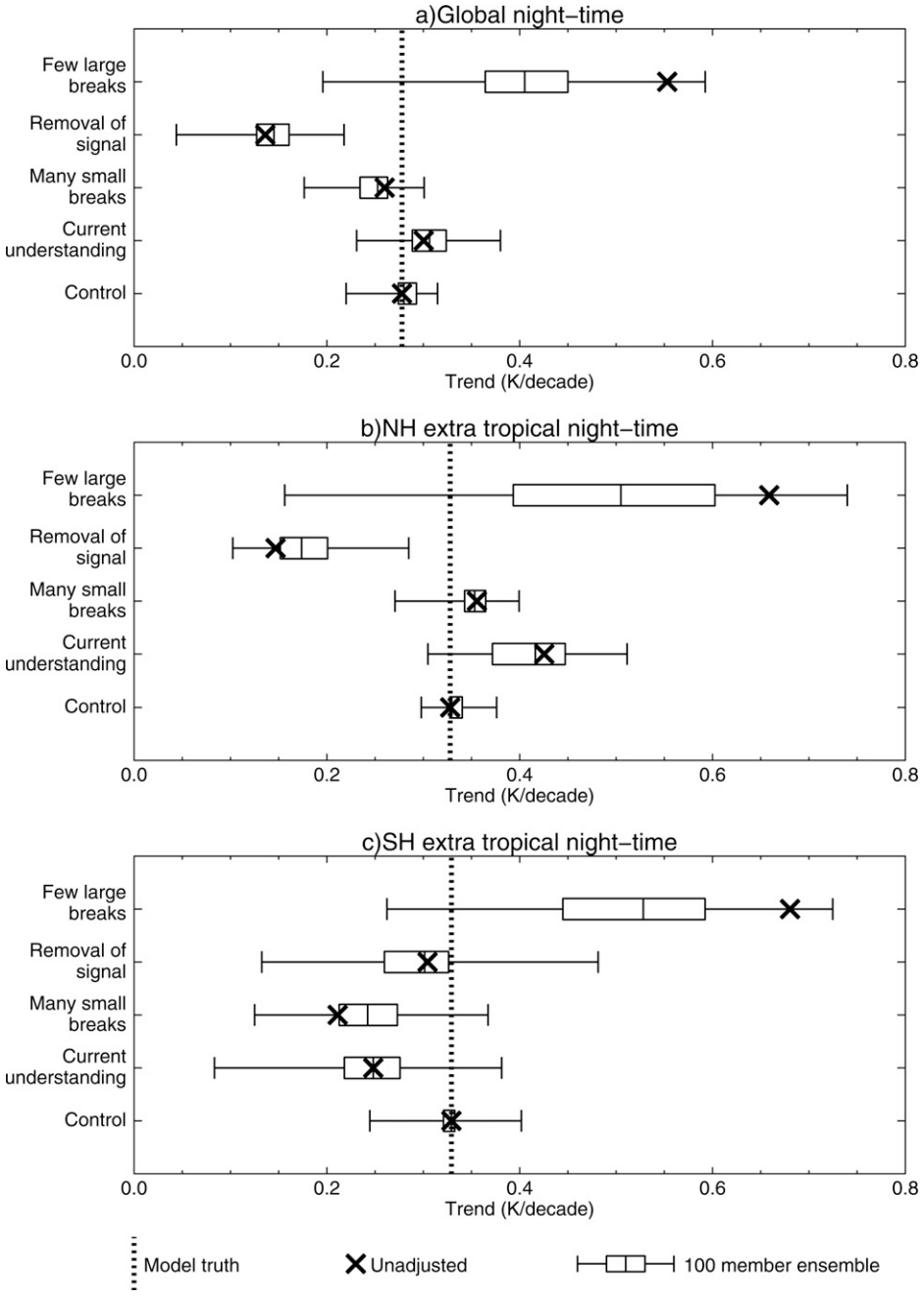

FIG. 7. As in Fig. 5 but for the (a) global, (b) NH extratropical, and (c) SH extratropical nighttime trends.

of the error models (Fig. 7a). However, in the other cases the median trend is shifted very little and the unadjusted trend falls well within the narrow interquartile range. The unadjusted trend from the night observations (Fig. 9a), however, falls above the upper quartile. There is, therefore, some evidence for a warm global nighttime bias, although this is not robust. The median nighttime trend for the NH extra tropics does capture the correct sign of the bias for all error models, even though the unadjusted trend sometimes lies within the interquartile range (Fig. 7b). Therefore Fig. 9b hints at a cold bias in the observed NH nighttime trends, although again this is not a robust result particularly as the unadjusted trend lies within a narrow interquartile range. Given the poor results from the validation experiments in the SH extratropics, we can say very little

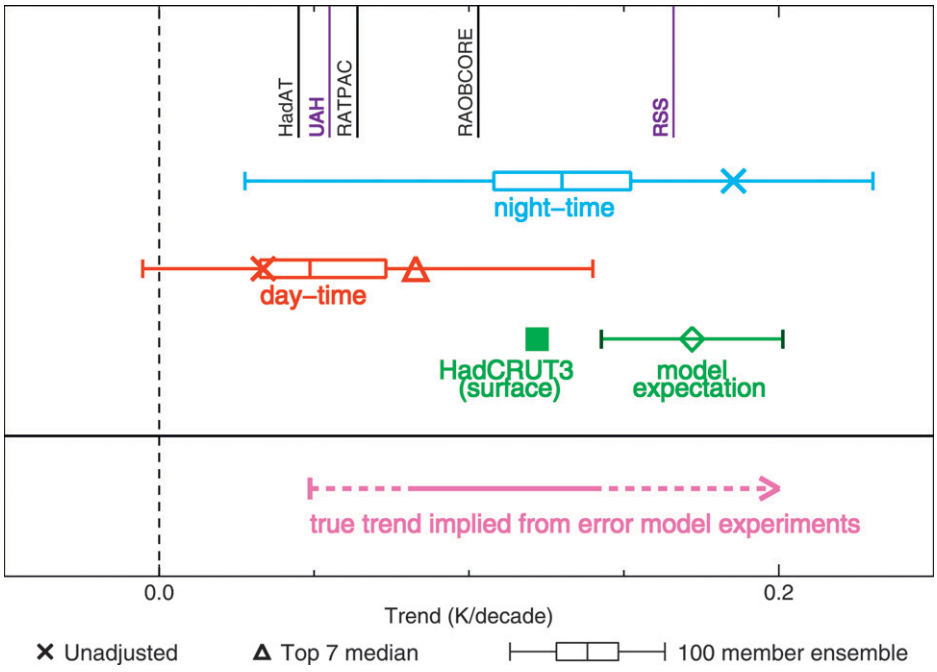

FIG. 8. Tropical T2LT MSU equivalent trends for the satellite period. The red (daytime) and blue (nighttime) crosses denote the trends in the unadjusted observations and the box plots denote the spread from each 100-member ensemble. The triangle shows the median trend from the "top" daytime experiments. The green diamond and horizontal bar denote the range of expected trends based on an ensemble of transient simulations using 19 different climate models (Santer et al. 2005, 2006). These were derived using the model tropospheric amplification estimates assuming that the Hadley Centre Climatic Research Unit, version 3 (HadCRUT3) surface trend (green square; Brohan et al. 2006) is perfect. Trend estimates from other radiosonde datasets are given in black (RAOBCORE version 1.4; Haimberger 2007); HadAT2 (Thorne et al. 2005b), Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC; Lanzante et al. 2003; Free et al. 2005), and MSU datasets are given in purple [UAH version 5.2 (Christy and Norris 2004, 2006); RSS version 2.1 (Mears et al. 2003; Mears and Wentz 2005)]. The pink bar denotes the true trend implied from the error model experiments in section 4a (the arrow represents our inability to place an upper bound on the range). The solid section denotes the range of estimates for which we have a higher confidence based on our findings.

about the SH extratropical daytime and nighttime trends. The homogenization system therefore has little or no skill in this region.

We can only make the following inferences regarding Fig. 9 using the results from the validation experiments that are shown in Fig. 7:

- The unadjusted global daytime trend is very likely biased cold, but we cannot comprehensively quantify the magnitude of the bias and its uncertainty.
- Our findings suggest that the true global trend is not only warmer than the median trend from the daytime ensemble (0.11 K decade^{-1}), but also warmer than the median trend from the top experiments (0.12 K decade^{-1}).
- The unadjusted NH extratropical daytime trend is very likely biased cold, but we cannot comprehensively quantify the magnitude of the bias and its uncertainty.
- Our findings suggest that the true NH extratropical trend is not only warmer than the median trend produced from the daytime ensemble (0.27 K decade^{-1}), but also warmer than the median trend from the top experiments (0.28 K decade^{-1}).

6. Assessing the value of the GUAN network

The error models can also be used to understand the effects of systematic changes to either the method employed or the data used. The latter can guide us on future possible changes to radiosonde network. We perform two additional sets of experiments using the error models to assess the value of the Global Climate

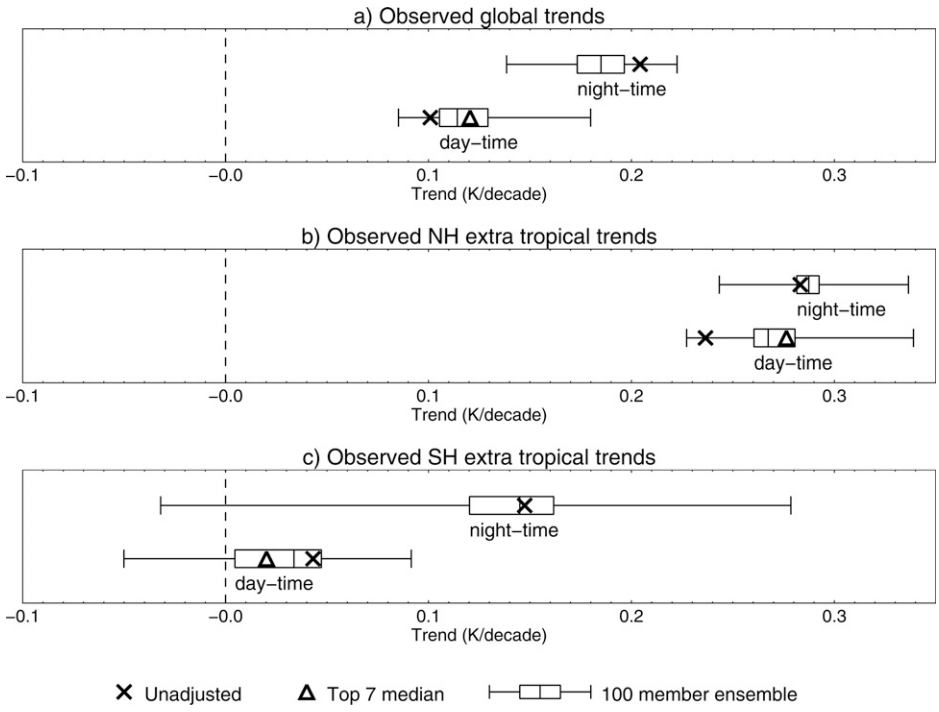

FIG. 9. (a) Global, (b) NH extratropical, and (c) SH extratropical observed T2LT equivalent trends. The box plots denote the spread in the daytime and nighttime 100-member ensembles and the crosses denote the trend in the unadjusted observations.

Observing System (GCOS) Upper Air Network (GUAN). This is a small network of stations with a worldwide coverage sufficient for the detection of global-mean trends (McCarthy 2008).

a. Obtaining perfect metadata for GUAN stations

First, we assess the impact of accurately recording known instrumental or observational changes for the GUAN stations. We subsampled the GUAN stations (Fig. 1b) in the daytime (giving a total of 123 stations) and nighttime (giving a total of 77 stations) error model datasets and passed them through the system with no metadata. We then used the same data but this time with perfect metadata containing the correct timings of all breakpoints. The system was forced to apply adjustments at these breakpoint timings, although in some cases where breakpoints were close together only one adjustment may have been applied (see section 3 of McCarthy et al. 2008 for more details). We did this using the seven top experimental setups on the daytime and nighttime datasets for each error model.

High quality metadata do not have a beneficial impact on the T2LT tropical trends compared to using no metadata at all (Fig. 10). The median trend of each daytime ensemble changes very little. At night, performance with metadata is worse in two ensembles. This may seem surprising given that the system is given perfect knowledge of breakpoint locations, but there is still a requirement for the system to provide accurate adjustments, which is particularly a challenge when only a few bad quality neighbors are used. It is likely that the GUAN nighttime coverage (Fig. 1b) is too sparse to create a sufficiently homogeneous neighbor composite series when only GUAN stations are used. It has already been seen in section 4b that the poor coverage in the nighttime data inhibits the ability of our system to recover the large-scale trends even when the full network is used (albeit incomplete metadata were used). Sparsity of the GUAN nighttime stations may also cause more of a problem when given perfect metadata because the system is being forced to adjust all breakpoints, including the very small ones, in the early iterations because knowledge of metadata in these experiments all but guarantees breakpoint identification.

b. Non-GUAN stations using perfect GUAN stations as neighbors

We now investigate the impact of using a high quality set of GUAN stations (Fig. 1b) for homogenizing the rest of the network (Fig. 1c). This assesses the potential benefits of maintaining the GUAN network according to the GCOS climate monitoring principles. First, we

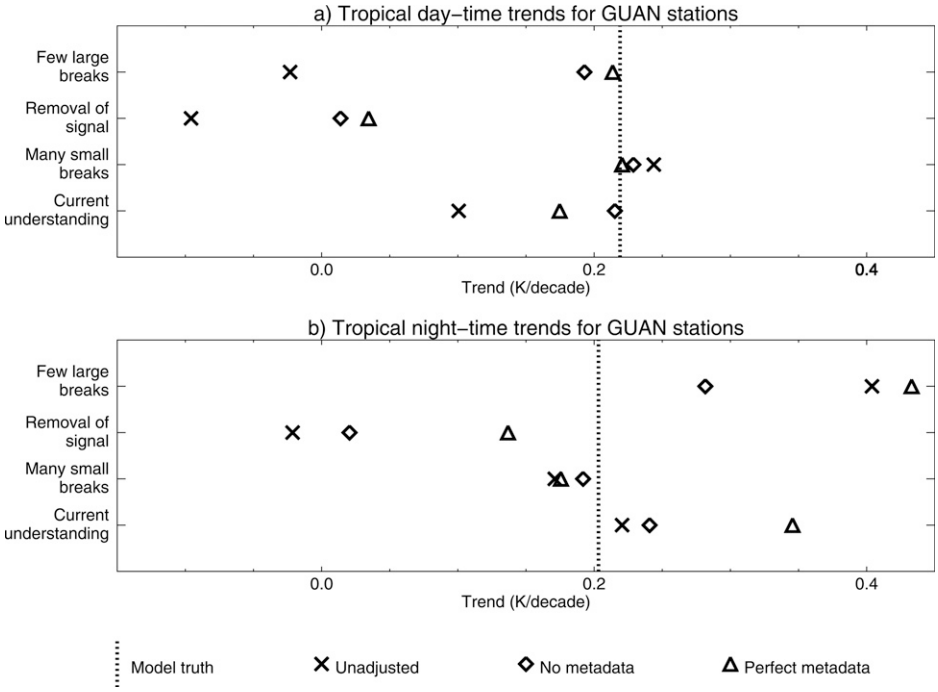

FIG. 10. Error model tropical T2LT trends using GUAN stations only for the (a) daytime and (b) nighttime data. Median trends for the ''top'' seven homogenization experimental setups are given using no metadata (diamonds) and perfect metadata (triangles). The trends in the unadjusted data are denoted by the black crosses and the original control trends are marked by the vertical dotted lines.

used the biased GUAN stations within each given error model to create the neighbor composite series, and then we used the perfect breakpoint-free GUAN stations from the original control dataset. Again we used the seven top experimental setups for each of the daytime and nighttime error model datasets.

The system is always better at capturing the true control trend in tropical T2LT when the perfect GUAN stations are used as neighbors (Fig. 11). The improvements in the *removal of signal* ensembles are particularly noticeable, which is encouraging as this error model was the hardest to homogenize in the previous experiments (section 4b), though only seven experiments were used here. These results indicate a high quality reference series is important for trend recovery, and this may be a problem when a biased sparse network is used. If we could gain a high quality GUAN or similar-sized network, then it is very likely we would be able to adequately constrain the uncertainties in the trends for the rest of the global network.

7. Conclusions and discussion

The main aim of this study was to assess whether the tropical tropospheric lapse rate discrepancy between climate models and some radiosonde datasets could conceivably be accounted for by uncertainty in the radiosonde records. To assess this uncertainty we developed four substantially different error models. These error models contained artificial breakpoint profiles based on different assumptions about the underlying error structure. They were applied to HadAM3 climate model data, which were subsampled to the daytime and nighttime observations. These biased data were adjusted using the automatic radiosonde homogenization system developed by McCarthy et al. (2008). We then assessed the ability of the system to recover the original large-scale trends, enabling us to make inferences about the biases and uncertainties in the real world observations.

The homogenization system produces a number of different realizations based on different methodological assumptions. A 100-member ensemble was created using each daytime and nighttime error model dataset, as well as the observations. Changing the parameter settings influenced the breakpoints identified and the adjustments calculated during the homogenization. This in turn affected the large-scale trends produced from each adjusted dataset.

Our prior knowledge of the original model trends

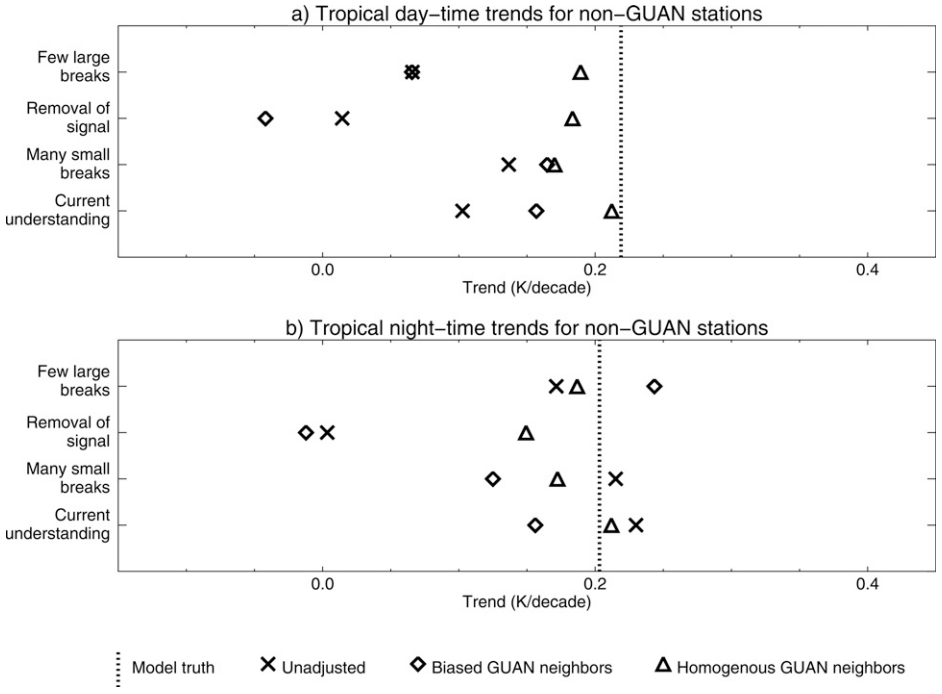

FIG. 11. As in Fig. 10 except using non-GUAN stations homogenized with error model biased GUAN stations as neighbors (diamonds) and homogeneous control GUAN stations as neighbors (triangles).

and the breakpoints within each error model enabled us to assess the homogenization skill of the system. Most experiments exhibited skill on a local observation basis relative to undertaking no homogenization at all. It was found that the performance of each member was fairly independent of the underlying error structure when ranked according to this local skill, but more dependent when ranked according to the large-scale tropospheric trend recovery skill (also see Sherwood et al. 2008b).

Our results indicate that the bias in the daytime tropical tropospheric trends was underestimated by the majority of error models experiments. One of the four daytime error model ensembles (the *removal of signal* ensemble) was unable to capture the original model trend at all. We are therefore unable to guarantee that the spread in daytime observation ensemble encompasses the real world trend. The change in the sign of the trend bias between the troposphere and stratosphere is unlikely to have caused the poor performance using the *removal of signal* error model, as the system's breakpoint identification and adjustment methodology should not be affected by such a change (see section 3a and McCarthy et al. 2008 for methodology details). However, a close examination of the *removal of signal* error model revealed that it contained randomly generated clustering of breakpoints at some times throughout the series. This could cause problems during the homogenization and inhibit the system's ability to recover the large-scale trends. It is possible that such clustering may occur in the real world observations. The cause could also be more complex still and relate to regional clustering and cancellation of errors, for example. The bottom line is that we have no unimpeachable basis on which to reject at least some aspects of each error model being prevalent in the poorly understood real world raw data.

However, the bias in each unadjusted error model trend was correctly reduced to some extent both in the median ensemble member trend and even more so in an identified set of optimal experimental setups. We are therefore confident that the system correctly identifies a cooling bias in the daytime observations, although we cannot robustly estimate the magnitude of the bias. Our analysis provides evidence that a lower bound of 0.08 K decade^{-1} can be placed on our real world tropical T2LT trend uncertainty estimate, but does not provide an upper bound. This lower bound indicates that many current upper-air datasets, such as HadAT, are biased cold. Unfortunately, the tropical nighttime observation ensemble results are unable to provide an upper or lower bound, as many of the error model experiments were unable to reduce the nighttime trend bias owing to the sparsity of data. Hence,

our analysis using realistic validation experiments is unable to discount or confirm the presence of a tropical tropospheric lapse rate discrepancy between the radiosonde observations and climate model expectations.

The daytime observations indicate that the unadjusted NH extratropical and global daytime trends are biased cold, although again we are unable to quantify the magnitude of these biases. There is some evidence that the unadjusted nighttime NH extratropical and global trends contain a cold and warm bias respectively, but this is not a robust result. Additional experiments would be required to investigate these biases further. The homogenization system performed particularly poorly in the SH extra tropics for some error models in both daytime and nighttime (likely owing to data sparsity), therefore it has little or no skill in this region and we have no confidence in the observational results.

The results from the error models, and hence the implications for our current understanding, depend on a number of factors. There are an infinite number of error structures that could be created by further varying the different assumptions. There are a large number of assumptions that we have not varied, such as the data used to create our control dataset. The effect of using a particular set of randomly generated breakpoints for each error model has also not been properly investigated (although this has been varied to some extent between the different error models). Although there is uncertainty in our results related to these influences, the four error models used within this work spanned a sufficiently large range to provide useful results regarding the performance of the homogenization system and the true observational trends.

Results are also dependent on the actual homogenization system being applied, therefore we will make our error model data freely available online at http://www.hadobs.org/ and encourage others to use them to critically reassess their systems also. The range of trend estimates produced from each ensemble highlights the risk of relying upon a single homogenization method for trend estimation, particularly when there is no or little knowledge of optimal parameter settings or methodological choices. We therefore believe that much would be gained from a coordinated comparison of independent radiosonde homogenization methods, particularly if realistic validation experiments are performed. The error model data have proven useful already in examining the new dataset produced by Sherwood et al. (2008a).

Sparsity of the network is likely to be accountable for the poor performance in some ensembles, particularly when compared to the ensembles that used a more comprehensive network, although the bias was consistently underestimated even in these cases. One explanation may be that the system is unable to construct a sufficiently high quality neighbor reference series. The ability of the system to recover the trends in some or all ensembles may be improved if alternative methodological decisions are made in addition to those already tested. We are therefore undertaking further experiments on the error models by adding additional flexibility to the system to try to ascertain whether we can better constrain our trend estimates. These include removal of data around assigned breakpoints in the neighbor series, an assessment of sensitivity to the time interval of the input data, and use of different breakpoint statistical identification tests.

This study has not only given us a better understanding of the trend uncertainties, but it also allowed us to investigate the impacts of possible targeted changes to the radiosonde network. A number of further validation experiments indicated that perfect metadata were unable to constrain the tropical tropospheric trend uncertainties using only GUAN stations. Results were much more encouraging when a GUAN network consisting of perfect station records was used as a reference series to homogenize the rest of the available stations. We therefore recommend that the GUAN network is maintained to a high standard, including adherence to the GCOS monitoring principles (GCOS 2004) so that we can have a better understanding of future trends in the free atmosphere.

Acknowledgments. Thanks to Steven Sherwood for comments on an earlier draft. Met Office Hadley Centre authors (and Simon Tett during his time at the Met Office Hadley Centre) were supported by the Joint Defra and MoD Integrated Climate Programme—GA01101, CBC/2B/0417_Annex C5. The work of Leo Haimberger was supported by Contract P18120-N10 of the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF).

APPENDIX A

Error Model Assumptions

The four error models were based on substantially different assumptions regarding breakpoint numbers, locations (i.e., station and date), sizes, and overall impacts on large-scale means. The average number of breakpoints per station (and therefore the total number of breakpoints) assigned to each error model (Table 1) was broadly based on the numbers found in existing literature (e.g., Thorne et al. 2005b; Haimberger 2007). Within each individual error model, the same breakpoint locations were assigned for both daytime and

TABLE A1. Summary of the different assumptions made during the derivation of the error models. Breakpoints corresponding to the real-world metadata record were initially imposed within each error model (columns 2 to 5). A number of extra breakpoint times were then randomly derived (columns 6–8). See text in appendix A for further details.

Error model	Percent of metadata events used to assign a break location	Standard deviation (σ) used in Eq. (A1) for the metadata breaks	Mean (μ_z) used in Eq. (A1) for the metadata breaks	The same metadata class of change given the same break profile	Standard deviation (σ) used in Eq. (A1) for the extra breaks	Mean (μ_z) used in Eq. (A1) for the extra breaks	Percent of extra breaks given the same profile at the same time within a given country
Current understanding	100% of known sonde change events, 66% of ground equipment changes, 33% of all other events	0.3 K	For daytime: decreasing with height from −0.04 to −0.15 K. For nighttime: 0.0 K at and below 500 hPa, decreasing with height above this from −0.03 to −0.07 K.	Yes	0.3 K	0.0 K at and below 500 hPa, decreasing with height above this so as to add a negative bias. A greater rate of decrease was used for the daytime profiles.	50%
Many small breaks	100%	0.3 K for 50% of breaks, 0.1 K for 50% of breaks	0.0 K	Yes	0.3 K until an average of 5 breaks per station were applied, 0.1 K until all remaining breaks were applied	0.0 K until an average of 5 breaks per station were applied, 0.0 K for 50% of remaining breaks, −0.01 K for 50% of remaining breaks	50% until an average of 5 breaks per station were applied, 0% for remaining breaks
Removal of signal	50%	0.3 K	Increasing with height from −0.08 to 0.12 K in order to remove most of the large-scale trends.	No	0.3 K	Increasing with height from −0.08 to 0.12 K in order to remove most of the large-scale trends.	50%
Few large breaks	75%	0.6 K	0.0 K	Yes	0.6 K	0.0 K	0%

TABLE B1. Possible settings for system parameters used in the ensemble experiments. Refer to McCarthy et al. (2008) for more information.

Parameter name	Description	Possible settings
Neighbor weighting coefficients	Weighting coefficients for possible neighbor stations derived from reanalysis fields	NCEP or ERA-40
Country/metadata	Excludes any neighbor stations within the same country/with similar metadata records as the target station	Either both on or both off
K–S window width	Number of seasons used for the K–S test used to assign breakpoints	8–20 seasons
Metadata weighting	Weighting given to metadata events during the breakpoint identification procedure (0 = no weight, 1 = breakpoint at every metadata event)	0.–1.
Metadata_function	Shape of inversion in the metadata statistic series at known events	Exponential or step
Vary metadata background	Alters the background value of the metadata probability series for each station based on the number of metadata events (i.e., penalizes stations with poor metadata records)	On or off
Range	Minimum number of seasons required between each breakpoint	6–20 seasons
Critical value	Initial critical threshold used to identify breakpoints in the first iteration	0.005, 0.02, or 0.05
Max iteration	Number of iterations performed	3, 6, or 9
Iteration step	Increment that the critical value is increased by with each iteration	0.005, 0.01, or 0.02
Adjustment method	Adjustment method used. Adaptive recalculates all adjustments at each iteration. Semiadaptive recalculates only if the breakpoint is found again at a later iteration. Nonadaptive calculates adjustment only when the breakpoint is first found	Adaptive or nonadaptive
Adjustment_period	Number of seasons either side of each breakpoint used to calculate an adjustment factor	5–20 or 40–55 seasons
Adjustment_threshold	Thresholds for determining whether an adjustment should be applied or not based on a points scoring system (appendix B; McCarthy et al. 2008)	[1, 1], [5, 8], [5, 11], or [7, 11]

nighttime series under the assumption that discontinuities will occur at both times with any change in practice. However, breakpoint profiles applied differed between day and night. There is strong quantitative evidence for such behavior to be associated with radiosonde changes—at least through the series of WMO intercomparison projects (Nash et al. 2005 and references therein). Differences between the daytime and nighttime series within each individual error model therefore relate to these differing assumptions as to average day and night error structures as well as to coverage differences and to random differences in the breakpoint magnitudes.

The metadata record containing known breakpoints in the real world (Gaffen 1996, and subsequent updates) was used to assign some breakpoint locations within each error model. In some of these cases the same breakpoint profiles were applied to all breakpoints associated with a particular class of change. For example all changes from Vaisala RS80 to Vaisala RS90 had identical day breakpoints and identical night breakpoints applied.

Extra breakpoints were chosen by randomly selecting stations and dates until the chosen total number of breakpoints was reached. In some error models a proportion of these extra breakpoints were assigned at the same date to all stations within the same country as a randomly selected station. In this case common daytime breakpoint profiles and common nighttime breakpoint profiles were applied to all stations within the country.

Vertically correlated breakpoint sizes were derived in order to create each breakpoint profile:

$$\text{break}_z = (0.9 \times \text{break}_{z-1}) + \nu(\mu_z, \sigma), \qquad \text{(A1)}$$

where break_z is the breakpoint size at a given pressure level z (numbered 1 to 10, increasing height from 850 to 50 hPa), and ν is an offset derived from a random distribution with mean μ_z (at level z) and standard deviation σ. Here $\text{break}_{Z-1} = 0$ for $z = 1$ (i.e., at 850 hPa); μ_z and σ were varied between each error model, and sometimes between different breakpoint profiles within an individual error model. Once each breakpoint profile was generated it was added to the daytime/nighttime control series (section 2b) at and before the breakpoint date.

Table A1 summarizes the different assumptions made during the derivation of the error models and includes the values used in Eq. (A1). See Table 1 for a summary of the resulting set of breakpoints applied to each error model.

APPENDIX B

System Parameter Settings

The automated homogenization system contains a number of parameters that affect various components

of the homogenization process. These parameters are outlined in appendix A of McCarthy et al. (2008). Table B1 gives the range of settings used for each of the parameters in the 100 experiments used within this study. For each of the 100 experiments the parameter values were randomly selected using a random number generator on each range of settings.

REFERENCES

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

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 data set from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

Christy, J. R., and W. B. Norris, 2004: What may we conclude about global tropospheric temperature trends? *Geophys. Res. Lett.,* **31,** L06211, doi:10.1029/2003GL019361.

——, and ——, 2006: Satellite and VIZ–radiosonde intercomparisons for diagnosis of nonclimatic influences. *J. Atmos. Oceanic Technol.,* **23,** 1181–1194.

Durre, I., R. S. Vose, and D. B. Wuertz, 2006: Overview of the Integrated Global Radiosonde Archive. *J. Climate,* **19,** 53–68.

Free, M., and D. J. Seidel, 2005: Causes of differing temperature trends in radiosonde upper air datasets. *J. Geophys. Res.,* **110,** D07101, doi:10.1029/2004JD005481.

——, ——, 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.

Gaffen, D., 1996: A digitized metadata set of global upper-air station histories. NOAA Tech. Memo. ERL ARL-211, 38 pp.

GCOS, 2004: Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC. GCOS-92, WMO/TD 1219, World Meteorological Organization, 136 pp.

Haimberger, L., 2007: Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate,* **20,** 1377–1403.

——, C., Tavolato and S. Sperka, 2008: Toward elimination of the warm bias in historic radiosonde records—Some new results from a comprehensive intercomparison of upper-air data. *J. Climate,* **21,** 4587–4606.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77,** 437–471.

Karl, T. R., S. J. Hassol, C. D. Miller, and W. L. Murray, 2006: 2w?>Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences. U.S. Climate Change Science Program and the Subcommittee on Global Change Research Rep., Synthesis and Assessment Product 1.1, 166 pp.

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.

——, S. A. Klein, and D. J. Seidel, 2003: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. *J. Climate,* **16,** 224–240.

McCarthy, M. P., 2008: Spatial sampling requirements for monitoring upper air climate change with radiosondes. *Int. J. Climatol.,* **28,** 985–993.

——, 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. Climate,* **21,** 817–832.

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

——, M. C. Schabel, and F. J. Wentz, 2003: A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Climate,* **16,** 3650–3664.

Nash, J., R. Smout, T. Oakley, B. Pathack, and S. Kurnosenko, 2005: The WMO intercomparison of radiosonde systems. WMO/TD 1303, 109 pp.

Parker, D. E., M. Gordon, D. P. N. Cullum, D. M. H. Sexton, C. K. Folland, and N. Rayner, 1997: A new global gridded radiosonde temperature data base and recent temperature trends. *Geophys. Res. Lett.,* **24,** 1499–1502.

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. *Climate Dyn.,* **16,** 123–146.

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. Cambridge University Press, 963 pp.

Randel, W. J., and F. Wu, 2006: Biases in stratospheric and tropospheric temperature trends derived from historical radiosonde data. *J. Climate,* **19,** 2094–2104.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of SST, sea ice and night marine air temperature since the late nineteenth century. *J. Geophys. Res.,* **108,** 4407, doi:10.1029/2002JD002670.

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

——, J. E. Penner, and P. W. Thorne, 2006: How well can the observed vertical temperature changes be reconciled with our understanding of the causes of these changes? Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences, U.S. Climate Change Science Program and the Subcommittee on Global Change Research Rep., Synthesis and Assessment Product 1.1, 89–118.

——, and Coauthors, 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.

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.

——, C. L. Meyer, R. J. Allen, and H. A. Titchner, 2008a: Robust tropospheric warming revealed by iteratively homogenized radiosonde data. *J. Climate,* **21,** 5336–5350.

——, H. A. Titchner, P. W. Thorne, and M. P. McCarthy, 2008b: How do we tell which estimates of past climate change are correct? *Int. J. Climatol.,* in press.

Spencer, R. W., and J. R. Christy, 1990: Precise monitoring of global temperature trends from satellites. *Science,* **247,** 1558–1562.

Tett, S. F. B., and Coauthors, 2007: The impact of natural and anthropogenic forcings on climate and hydrology since 1550. *Climate Dyn.*, **28,** 3–34.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears, 2005a: Uncertainties in climate trends: Lessons from upper-air temperature records. *Bull. Amer. Meteor. Soc.,* **86,** 1437–1442.

——, ——, S. Tett, P. Jones, M. McCarthy, H. Coleman, and P. Brohan, 2005b: Revisiting radiosonde upper-air temperatures from 1958 to 2002. *J. Geophys. Res.,* **110,** D18105, doi:10.1029/2004JD005753.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. 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.