JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 105, NO. D6, PAGES 7337–7356, MARCH 27, 2000

Statistical significance of trends and trend differences in layer-average atmospheric temperature time series

B. D. Santer,[1] T. M. L. Wigley,[2] J. S. Boyle,[1] D. J. Gaffen,[3] J. J. Hnilo,[1] D. Nychka,[2] D. E. Parker,[4] and K. E. Taylor[1]

Abstract. This paper examines trend uncertainties in layer-average free atmosphere temperatures arising from the use of different trend estimation methods. It also considers statistical issues that arise in assessing the significance of individual trends and of trend differences between data sets. Possible causes of these trends are not addressed. We use data from satellite and radiosonde measurements and from two reanalysis projects. To facilitate intercomparison, we compute from reanalyses and radiosonde data temperatures equivalent to those from the satellite-based Microwave Sounding Unit (MSU). We compare linear trends based on minimization of absolute deviations (LA) and minimization of squared deviations (LS). Differences are generally less than 0.05°C/decade over 1959–1996. Over 1979–1993, they exceed 0.10°C/decade for lower tropospheric time series and 0.15°C/decade for the lower stratosphere. Trend fitting by the LA method can degrade the lower-tropospheric trend agreement of 0.03°C/decade (over 1979–1996) previously reported for the MSU and radiosonde data. In assessing trend significance we employ two methods to account for temporal autocorrelation effects. With our preferred method, virtually none of the individual 1979–1993 trends in deep-layer temperatures are significantly different from zero. To examine trend differences between data sets we compute 95% confidence intervals for individual trends and show that these overlap for almost all data sets considered. Confidence intervals for lower-tropospheric trends encompass both zero and the model-projected trends due to anthropogenic effects. We also test the significance of a trend in $d(t)$, the time series of differences between a pair of data sets. Use of $d(t)$ removes variability common to both time series and facilitates identification of small trend differences. This more discerning test reveals that roughly 30% of the data set comparisons have significant differences in lower-tropospheric trends, primarily related to differences in measurement system. Our study gives empirical estimates of statistical uncertainties in recent atmospheric temperature trends. These estimates and the simple significance testing framework used here facilitate the interpretation of previous temperature trend comparisons involving satellite, radiosonde, and reanalysis data sets.

1. Introduction

Since 1979 the satellite-based Microwave Sounding Units (MSU) have measured the upwelling microwave radiation from oxygen molecules. These observations have been used to monitor the vertically weighted temperature of deep atmospheric layers [*Spencer and Christy*, 1992a, b; *Christy et al*., 1998]. In recent years, considerable scientific attention has focused on one specific MSU product, the 2_{LT} retrieval of lower-tropospheric temperatures.

Several studies have noted the close agreement (to within 0.03°C/decade) [*Christy et al*., 1997, 1998] between global-mean MSU 2_{LT} trends and lower-tropospheric temperature-change estimates derived from compilations of radiosonde data by *Angell* [1988] and *Parker et al*. [1997]. This agreement is frequently cited in discussions of the reliability of the MSU 2_{LT} temperature record [*Christy et al*., 1997, 1998]. As noted by *Santer et al*. [1999], however, such comparisons do not account for large spatial and temporal coverage differences between the satellite and radiosonde data sets. Accounting for these differences can degrade the previously reported MSU/radiosonde trend correspondence, which suggests that it may be partly fortuitous.

Santer et al. [1999] (henceforth S99) attempted to quantify some of the uncertainties that hamper interpretation of the previously reported MSU/radiosonde trend agreement. They identified four types of uncertainty. These were related to (1) residual inhomogeneities in both the radiosonde and the MSU data, (2) the procedures used in generating gridded radiosonde data sets from raw station data, (3) coverage differences between the MSU and radiosonde data sets, and (4) the method used in computing "equivalent" MSU temperatures from radiosonde data.

S99 focused on items 2, 3, and 4. They showed that two versions (HadRT1.1 and HadRT1.2) of the Hadley Centre radiosonde data set compiled by *Parker et al.* [1997] had mark-

[1]Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, California.
[2]National Center for Atmospheric Research, Boulder, Colorado.
[3]National Oceanic and Atmospheric Administration Air Resources Laboratory, Silver Spring, Maryland.
[4]Hadley Centre for Climate Prediction and Research, United Kingdom Meteorological Office, Bracknell.

Paper number 1999JD901105.
0148-0227/00/1999JD901105$09.00

edly different lower-tropospheric temperature trends over 1979–1996 (+0.040°C/decade and −0.037°C/decade, respectively). These were primarily due to large differences in spatial coverage, which in turn were related to different assumptions regarding the spatial representativeness of the raw radiosonde data. They also found that trends based on the globally complete MSU data and on the MSU data subsampled with HadRT1.1 coverage could diverge by up to 0.06°C/decade. This finding highlighted the importance of accounting for coverage differences in MSU/radiosonde comparisons. The choice of method for computing an equivalent MSU temperature was found to have a negligible effect on global-scale trends. Recent work by *Gaffen et al*. [2000] has explored trend uncertainties related to item 1 and shows that decisions made regarding adjustments for radiosonde inhomogeneities can have a significant impact on local trends and probably on resultant global-scale trends.

In the present study, we consider a fifth source of uncertainty, one introduced by the choice of statistical method used to estimate trends. To date, virtually all studies have described secular changes in layer-average atmospheric temperatures by fitting least squares linear trends to the data. An exception is the recent investigation by *Gaffen et al*. [2000], who demonstrate that trend estimates obtained with least squares linear regression differ by up to 0.03°C/decade from estimates based on the median of pairwise slopes. There is no reason a priori why a least squares linear fit should be preferable to alternative linear-fitting methods. Here we use both a least squares fit and a fit that minimizes the mean absolute (rather than the mean square) deviation between the data points and the trend line [*Press et al*., 1992]. We will show that over the relatively short MSU record, the two methods of obtaining linear fits can yield large trend differences.

Knowledge of the size of trend uncertainties arising from these sources provides some context for interpreting previous comparisons of trends in MSU and radiosonde data. Useful complementary information can be obtained by testing the formal statistical significance of the individual trends and the trend differences between data sets. The second main issue that we explore in this paper is how trend significance should be assessed. Our intention here is not to provide an exhaustive review of possible approaches for evaluating trend significance in the time domain [see, e.g., *Bartlett*, 1935; *Mitchell et al*., 1966; *Karl et al*., 1991] and frequency domain [*Bloomfield and Nychka*, 1992]. Rather, our aim is to consider the sensitivity of significance testing results to assumptions made regarding adjustments for temporal autocorrelation of the data.

The structure of the paper is as follows. In section 2 we briefly introduce the various data sets of layer-mean atmospheric temperature that we employ and describe how we compute equivalent MSU temperatures from radiosonde data and reanalyses. Section 3 considers the sensitivity of the trend value to the choice of method used to perform a linear fit to the data. The approaches that we use to determine the significance of individual trends and trend differences between data sets are outlined and applied in sections 4 and 5, respectively. A summary and conclusions are given in section 6.

2. Temperature Data

2.1. Satellite Data

We use versions "b," "c," and "d" of the actual MSU layer-mean temperature data, as supplied by John Christy (University of Alabama in Huntsville). These are referred to henceforth as MSUb, MSUc, and MSUd, respectively. Version "a" of the data set [*Spencer and Christy*, 1992a, b] utilized a simple procedure to merge data from the (currently nine) individual satellites that comprise the MSU record. Version b of the 2_{LT} retrieval attempted to account for a systematic bias in the sampling of the diurnal cycle related to an eastward drift of the NOAA 11 satellite. Corrections for eastward drift of NOAA 7 were implemented in version c, together with adjustments for intra-annual variations in instrument-body temperature. The most recent MSU 2_{LT} retrieval, version "d02," incorporates additional adjustments for east-west drift of satellites and uses improved calibration coefficients for the MSU instrument on NOAA 12 [*Christy et al*., 1999]. It also includes corrections for an orbital decay effect identified by *Wentz and Schabel* [1998] and for interannual variations in instrument-body temperature [*Christy et al*., 1999].

The MSUb and MSUc data spanned the periods 1979–1995 and 1979–1997, respectively, while MSUd was available for 1979–1998. All three versions of the MSU data were in the form of monthly means on a 2.5° × 2.5° latitude/longitude grid. For each version, data were available for the 2_{LT} retrieval and channels 2 and 4, which provide information on (vertically weighted) mean temperatures in the lower troposphere, midtroposphere and lower stratosphere, respectively. The nominal maxima of the weighting functions for these three channels are at 740, 595, and 74 hPa.

2.2. Reanalysis Data

Reanalysis projects use a numerical forecast model of the atmosphere with a fixed observational data assimilation system [*Trenberth*, 1995]. The model output is not a direct observation of the climatic state, since it is influenced by the data assimilation strategies and numerical models that are employed. However, it does yield internally consistent climate data uncontaminated by the changes in model physics that typically affect operational analyses [*Trenberth and Olson*, 1991].

We use data from two separate reanalyses. The first is that performed by the European Centre for Medium-Range Weather Forecasts (ECMWF) and is referred to henceforth as ERA (ECMWF Re-Analysis) [see *Gibson et al*., 1997]. The second is that conducted jointly by the National Center for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR). We refer to this as NCEP [see *Kalnay et al*., 1996]. The reanalyses are of different lengths: NCEP covers the period January 1958 through December 1997, while ERA data are available from January 1979 through February 1994. Monthly-mean reanalysis data were interpolated to a common 2.5° × 2.5° latitude/longitude grid to facilitate intercomparisons. Temperature data from NCEP and ERA were available on 17 discrete pressure levels.

The two reanalyses differ not only in terms of the physics and resolution of the numerical forecast models that they use but also in terms of the data assimilation strategies employed, particularly with regard to the assimilation of satellite data. It is therefore difficult to isolate the exact cause or causes of the differences in the climate changes that ERA and NCEP simulate (see S99).

2.3. Radiosonde Data

We consider temperature information from three different radiosonde data sets. The first two (HadRT1.1 and HadRT1.2) were compiled by *Parker et al*. [1997] and were based on

monthly CLIMAT reports. In HadRT1.1 (HadRT1.2), station data were gridded to 5° × 10° (10° × 20°) latitude/longitude boxes. The different gridding procedures result in a substantial coverage increase in HadRT1.2 relative to HadRT1.1 (see S99). Other differences between HadRT1.1 and HadRT1.2 are discussed by *Parker et al*. [1997].

While HadRT1.1 is available in the form of monthly-mean anomalies from January 1958 through December 1996, HadRT1.2 consists of seasonal-mean anomalies from March to May (MAM) 1958 through September to November (SON) 1996. In each case, anomalies were defined relative to a 1971–1990 base period. HadRT1.1 (HadRT1.2) has nine (eight) vertical levels.

The third radiosonde data set used here consists of virtual or "thickness" temperatures computed from the height differences between specific pressure levels [*Angell*, 1988]. We refer to this subsequently as "ANGELL." Thickness temperatures in ANGELL were estimated using individual (daily or twice daily) soundings from a network of 63 stations. Possible effects on global-average temperature estimates arising from this sparse coverage and from instrumental inhomogeneities have been discussed by *Trenberth and Olson* [1991], *Gaffen* [1994], and S99. *Elliott et al*. [1994] additionally consider the effect of both real and apparent humidity changes (the latter due to radiosonde humidity sensor changes) on ANGELL virtual temperatures. The ANGELL data are available in the form of global-mean seasonal-mean anomalies (relative to a 1958–1977 base period) from December to February (DJF) 1958 through DJF 1998.

2.4. Computation of Equivalent MSU Temperatures

To facilitate comparison with the actual MSU deep-layer temperatures for the 2_{LT} retrieval and channels 2 and 4, we computed equivalent MSU temperatures from NCEP, ERA, HadRT1.1, and HadRT1.2. This was not possible for the ANGELL data, since these exist in the form of layer-average temperatures only. Nevertheless, the ANGELL data were included in our study because they figure prominently in previous comparisons of MSU- and radiosonde-derived temperature trends [e.g., *Christy et al*., 1997, 1998].

We computed equivalent MSU temperatures in two ways, using both a radiative transfer code and a static weighting function [see *Spencer and Christy*, 1992a]. The former approach accounts for land/sea differences in surface emissivity and for variations in atmospheric moisture as a function of space and time, while the latter approach does not.

Information on both methods, henceforth referred to as "radiative transfer" (RT) and "weighting function" (WF), is provided in S99. The RT method requires actual temperatures and was not used for the HadRT1.1 and HadRT1.2 radiosonde data, since these are available as temperature anomalies only. Equivalent MSU temperatures from ERA and NCEP were computed with both RT and WF methods. The trend differences arising from the use of different methods of computing an equivalent MSU temperature were found to be generally <0.02°C/decade on global scales (see S99).

We also used the WF method to calculate equivalent MSU temperatures from "masked" versions of NCEP and ERA, as described in S99. The resulting "NCMASK" and "ERMASK" data sets mimic exactly the coverage changes in HadRT1.1. This provides useful information on the trend uncertainties resulting from coverage differences between data sets. Subsampling of the actual MSUc data with HadRT1.1 coverage ("MSUMASKc") was also performed.

Finally, we note that ANGELL's 850–300 hPa and 100–50 hPa virtual temperatures represent (weighted) averages over different layers than the actual and equivalent 2_{LT} and channel 4 temperatures obtained from MSU, reanalyses, and the HadRT data. While the midpoint of ANGELL's stratospheric layer (75 hPa) is very close to the peak of the MSU channel 4 weighting function (~74 hPa), MSU channel 4 samples a deeper atmospheric layer. In the tropics, where the tropopause is typically at ~100 hPa, the ANGELL layer is often completely in the stratosphere, whereas MSU channel 4 includes a substantial upper-tropospheric contribution.

The midpoint of the ANGELL tropospheric layer is at 575 hPa, which is closer to the peak of the MSU channel 2 weighting function (~595 hPa) than to the peak of the 2_{LT} weighting function (~740 hPa). Historically, however, the ANGELL 850–300 hPa virtual temperatures have been compared with the MSU 2_{LT} retrieval rather than with MSU channel 2 temperatures, [*Christy*, 1995; *Christy et al*., 1997, 1998], probably to avoid the stratospheric influence on MSU 2 retrievals at middle and high latitudes. Noting this inconsistency, we nevertheless present comparisons between MSU 2_{LT} retrievals and ANGELL's 850–300 hPa data in order to shed light on the results of previous comparisons.

3. Linear Trend Sensitivity to Fitting Method

To investigate the sensitivity of linear trends to the choice of fitting method, we use global-mean seasonal-mean temperature anomalies from the data sets described in section 2. All anomalies are defined with respect to 1979–1993 climatological seasonal means. We consider sensitivities to fitting method for short-term trends over 1979–1993 (the period of overlap between the data sets used here) and for longer-term trends over 1959–1996 (the period of overlap between the NCEP, HadRT1.1, HadRT1.2, and ANGELL data sets). This yields time samples of $n_t = 60$ and $n_t = 152$, respectively.

Previous comparisons of linear trends in different temperature data sets have almost invariably used a least-squares estimator of the trend [e.g., *Parker et al*., 1997; *Christy et al*., 1998; S99]. Alternative linear trend estimators exist, which are less sensitive to outliers [see, e.g., *Lanzante*, 1996]. One such estimator involves minimization of the absolute deviations between the data and the linear fit [*Press et al*., 1992]. We refer to these two approaches subsequently as "LS" (least squares) and "LA" (least absolute deviation).

3.1. Channel 4

Over the period of the MSU record, lower stratospheric temperature anomalies typically show pronounced cooling (see Figure 1). In the radiosonde data this cooling is sustained over an even longer period of time (the reasons why long-term cooling is not evident in the 40-year NCEP reanalysis are discussed in S99). It is likely that some portion of this multidecadal cooling of the lower stratosphere is related to the combined anthropogenic effects of stratospheric ozone depletion and an increase in atmospheric CO_2 and other greenhouse gases [*Ramaswamy et al*., 1996; *Berntsen et al*., 1997; *Chanin et al*., 1999].

The short-term (1–2 year) stratospheric warming signatures of volcanic aerosols (e.g., from the eruptions of Mt. Agung in March 1963, Mt. El Chichón in April 1982, and Mt. Pinatubo in June 1991) constitute noise which hampers estimation of any

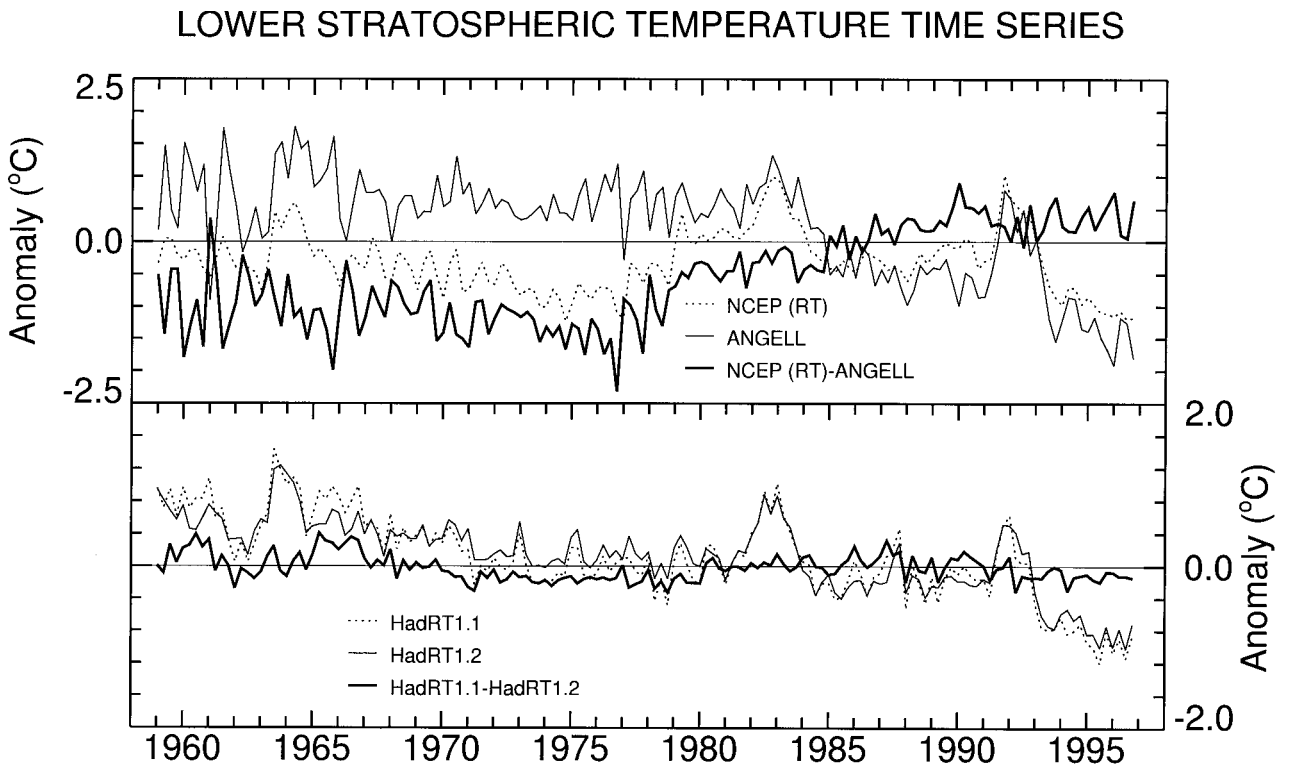

Figure 1. Time series of global-mean seasonal-mean temperature anomalies (°C) in the lower stratosphere. (top) Results for ANGELL's radiosonde-based 50–100 hPa thickness temperatures, the equivalent MSU channel 4 temperatures obtained from the NCEP reanalysis, and the difference between NCEP and ANGELL. (bottom) The equivalent channel 4 time series estimated from the HadRT1.1 and HadRT1.2 radiosonde data, together with their difference time series.

long-term anthropogenic signal. The weight that this noise is given is relatively greater in LS than in LA. We might therefore expect to find systematic differences between the overall channel 4 trends estimated with LA and LS. These differences are related to (1) the temporal distribution of volcanically induced noise within the time series (i.e., whether volcanic effects are symmetrically or asymmetrically distributed about the midpoint of the time series and how close they are to the midpoint), (2) the nonuniform response of the atmosphere to different volcanic eruptions, and (3) the asymmetrical nature of the temperature response to volcanic forcing (i.e., the rapid initial response and more gradual decay).

The noise induced by El Chichón and Pinatubo is not symmetrically distributed about the midpoint of the 1979–1993 channel 4 time series. Pinatubo's warming signature is closer to the endpoint of the time series and should therefore lead to LS trend estimates that are less negative than LA trend estimates (see Figure 2). Our results are in accord with this expectation in 16 out of 19 cases (13 LS/LA trend comparisons in Table 1a plus 6 in Table 1b).

We next investigated the sensitivity of LS and LA trend estimates to removal of the temperature signatures of El Chichón and Pinatubo from the channel 4 time series. After visual inspection of the MSUd channel 4 anomaly time series, we excluded the six seasons MAM 1982 through June to August (JJA) 1983 (El Chichón) and JJA 1991 through SON 1992 (Pinatubo) from each data set. We then recomputed LS and LA trend estimates for the reduced time sample (i.e., for $n_t = 48$ rather than $n_t = 60$).

Excluding volcanic effects from the 1979–1993 lower stratospheric temperature time series yields systematically larger cooling trends in virtually all cases, both for LS and LA trend estimates (see Table 1a). It also reduces differences between the LS and LA trend estimates (except for the NCEP (RT) results). This finding is relevant for comparisons of short-timescale lower stratospheric trends in models and data. Failure to account for volcanic forcing effects could easily yield large mismatches between observed and model-predicted trends over the period of the satellite record, even if anthropogenic forcing uncertainties and model errors were relatively small.

A comparison of Tables 1a and 1b indicates that the differences in channel 4 trends resulting from the linear fit method are larger for 1979–1993 than for the longer 1959–1996 period. This is primarily because 1979–1993 contains the two largest volcanically induced "warming outliers," and these outliers strongly influence the LS/LA trend differences.

For the 1979–1993 period, our results may be compared with results obtained by S99 (their Table 6). The latter study showed that channel 4 trend uncertainties arising from the version of the MSU or HadRT data used and the method used to compute an equivalent MSU temperature never exceeded 0.037°C/decade. The trend differences resulting from the choice of linear fitting method are much larger than this, exceeding 0.10°C/decade in 7 out of 13 cases (for the "Volcano Included" results in Table 1a).

3.2. Lower Tropospheric Retrieval and Channel 2

In the lower stratosphere, episodic volcanically induced warming is natural in origin and constitutes background noise which affects estimates of any putative anthropogenic signal

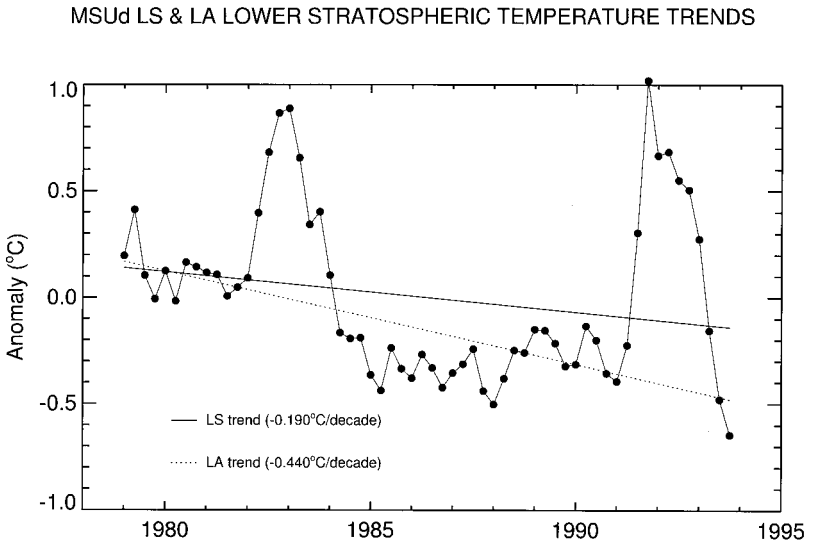

Figure 2. Time series of seasonal-mean lower stratospheric (channel 4) temperature anomalies (°C) over 1979–1993 in MSUd. Also shown are linear fits to the data using two approaches: least-squares (LS) and least-absolute deviations (LA).

trend. LA trend estimates are less sensitive to such noise than the more commonly used LS trend estimates. The former may therefore provide more reliable estimates of any underlying deterministic trend. In the lower and middle troposphere, however, it is much more difficult to partition time series into "signal" and "noise" components. Multidecadal trends in the 2_{LT} retrieval and channel 2 are strongly influenced by variability on 2 to 5-year El Niño–Southern Oscillation (ENSO) timescales and by the choice of endpoints relative to the phase of this quasi-periodicity (see Figure 3). There is also considerable evidence of longer-term ENSO variability (summarized by *Nicholls et al*. [1996]). Given the possibility that some component of this longer-term ENSO variability is associated with anthropogenic forcing [*Trenberth and Hoar*, 1996; *Timmermann et al*., 1999], we cannot easily partition signal from noise and do not know whether LA or LS trends provide a more reliable estimate of any underlying deterministic trend.

The 2_{LT} results bear certain similarities to those obtained for channel 4. First, there are systematic differences between the LS and LA trend estimates. In 16 out of 19 cases, the LS trends

Table 1a. Linear Trends Over 1979–1993 Estimated Using Least Squares and Least Absolute Deviation Approaches

	2_{LT} Retrieval			Channel 2			Channel 4 Volcanoes Included			Channel 4 Volcanoes Excluded		
	LS	LA	LS-LA	LS	LA	LS-LA	LS	LA	LS-LA	LS	LA	LS-LA
NCEP (RT)	−0.028	−0.127	+0.099	−0.044	−0.080	+0.036	−0.244	−0.315	+0.071	−0.391	−0.315	−0.076
NCEP (WF)	−0.034	−0.124	+0.090	−0.058	−0.062	+0.004	−0.246	−0.306	+0.060	−0.389	−0.364	−0.025
NCMASK (WF)	−0.001	−0.023	+0.023	−0.062	+0.022	−0.084	−0.276	+0.076	−0.353	−0.374	−0.334	−0.039
ERA (RT)	+0.106	+0.072	+0.033	+0.039	+0.010	+0.029	−0.256	−0.318	+0.062	−0.408	−0.410	+0.002
ERA (WF)	+0.101	+0.050	+0.051	+0.022	−0.026	+0.048	−0.263	−0.327	+0.064	−0.417	−0.400	−0.018
ERMASK (WF)	+0.093	+0.097	−0.004	+0.019	+0.004	+0.015	−0.300	−0.394	+0.094	−0.464	−0.396	−0.069
MSUb	−0.070	−0.172	+0.102	+0.007	−0.012	+0.019	−0.239	−0.394	+0.155	−0.427	−0.463	+0.036
MSUc	−0.049	−0.113	+0.064	+0.015	−0.014	+0.028	−0.240	−0.396	+0.156	−0.428	−0.463	+0.035
MSUd	−0.054	−0.139	+0.085	−0.074	−0.067	−0.007	−0.190	−0.440	+0.250	−0.371	−0.417	+0.046
MSUMASKc	+0.011	−0.037	+0.048	+0.052	+0.109	−0.056	−0.371	−0.494	+0.123	−0.539	−0.592	+0.053
HadRT1.1	+0.065	+0.042	+0.024	−0.005	+0.115	−0.120	−0.340	−0.185	−0.154	−0.393	−0.320	−0.074
HadRT1.2	−0.049	−0.075	+0.026	−0.098	−0.149	+0.051	−0.347	−0.384	+0.036	−0.447	−0.422	−0.025
ANGELL	−0.053	−0.053	0.000	—	—	—	−0.976	−1.143	+0.167	−1.216	−1.280	+0.064

Linear trends (°C/decade) in global-mean seasonal-mean temperature for three deep atmospheric layers, as estimated from reanalyses (NCEP, ERA), the satellite-based Microwave Sounding Unit (MSUb, MSUc, and MSUd), and radiosondes (HadRT1.1, HadRT1.2, and ANGELL). All trends were computed over 1979–1993. The atmospheric layers considered are the lower troposphere, midtroposphere, and lower stratosphere, as defined in terms of the characteristics of the MSU weighting functions for the 2_{LT} retrieval and channels 2 and 4, respectively. To facilitate comparison of trends in disparate data sets, "equivalent" MSU temperatures were computed from the NCEP and ERA data sets using two approaches: a global-mean weighting function (WF) and a radiative transfer code (RT; see section 3). Only the WF approach was used for computing equivalent MSU temperatures from the HadRT data sets. Trends estimated from the NCMASK, ERMASK, and MSUMASKc data sets were obtained after subsampling the globally complete reanalyses and MSUc with the actual coverage changes in HadRT1.1. Linear trends were fitted using both a conventional least squares approach (LS) and a method that minimizes the absolute deviations (LA). The trend differences arising from use of different fitting methods (LS minus LA) are also shown. The final three columns give trends computed after removing most of the effects of the El Chichón and Pinatubo eruptions from time series of lower-stratospheric temperatures.

Table 1b. Linear Trends Over 1959–1996 Estimated Using Least Squares and Least Absolute Deviation Approaches

	2_{LT} Retrieval			Channel 2			Channel 4		
	LS	LA	LS-LA	LS	LA	LS-LA	LS	LA	LS-LA
NCEP (RT)	+0.142	+0.135	+0.007	+0.160	+0.135	+0.025	−0.018	−0.029	+0.011
NCEP (WF)	+0.160	+0.147	+0.013	+0.178	+0.154	+0.024	−0.002	−0.014	+0.011
NCMASK (WF)	+0.115	+0.120	−0.005	+0.101	+0.105	−0.004	−0.068	−0.072	+0.004
HadRT1.1	+0.098	+0.095	+0.003	+0.017	+0.020	−0.003	−0.350	−0.364	+0.014
HadRT1.2	+0.110	+0.091	+0.018	+0.025	+0.010	+0.015	−0.315	−0.309	−0.006
ANGELL	+0.096	+0.095	+0.001	—	—	—	−0.501	−0.547	+0.046

As for Table 1a, but for linear trends (°C/decade) over 1959–1996. The ERA and MSU data sets are not included here since they commence in 1979. The "volcanoes excluded" case considered in Table 1a is not shown here.

are larger than LA trends (Tables 1a and 1b). This is in part due to the lesser weight that LA trend estimates give to the large positive temperature anomalies associated with El Niño events, which are more prominent near the end of the record (Figure 3). Second, trend differences between the LA and LS methods (over 1979–1993) are larger than those arising from coverage differences, version of the MSU data, and the method used to compute an equivalent MSU temperature (see Table 1a above and Table 6 in S99). The two linear fitting methods give lower tropospheric trend differences ≥0.05°C/decade in 6 out of 13 cases and ≥0.10°C/decade in 1 case (Table 1a). The third similarity with the channel 4 results is that the trend uncertainties due to the linear fitting method are much smaller over the longer 1959–1996 period than over 1979–1993 (compare Tables 1a and 1b).

One interesting result relates to levels of trend agreement between MSU and radiosondes (Table 2). Previously published MSUc/ANGELL and MSUc/HadRT1.2 trend comparisons for lower tropospheric temperatures have relied on LS estimates of overall trends [e.g., *Christy et al*., 1997, 1998]. Over 1979–1993 the LS lower tropospheric trends in MSUc and ANGELL agree to within 0.004°C/decade, while MSUc/HadRT1.2 trends are identical. (However, note that MSUc/HadRT1.1 trends differ by 0.114°C/decade, for reasons primarily related to differences in coverage.) The use of LA trend estimates systematically degrades these correspondences. The LA trend difference between MSUc and HadRT1.2 is 0.038°C/decade, while differences between MSUc and ANGELL (0.060°C/decade) and MSUc and HadRT1.1 (0.155°C/decade) are even larger. A similar degradation in trend correspondence is obtained for all MSUd/radiosonde comparisons over 1979–1993 (Table 2) as well as for four out of six MSU/radiosonde trend comparisons over 1979–1996, a period frequently considered in previous work [*Parker et al*., 1997; *Christy et al*., 1998].

For channel 2, trend uncertainties resulting from the fitting method can be as large as 0.120°C/decade. The LS/LA trend differences are less systematic than those found for channel 4 and the 2_{LT} retrieval but again tend to be much smaller over the longer 1959–1996 period than over 1979–1993.

4. Statistical Significance of Individual Trends

In this section we consider how the significance of individual trends should be assessed when the data are strongly autocorrelated. We do this primarily within the framework of a model consisting of a linear trend plus noise, where the noise is assumed to have a lag-1 autocorrelation structure. Many alternative statistical models can be fitted to atmospheric temperature time series [see, e.g., *Karl et al*., 1991; *Bloomfield and Nychka*, 1992; *Woodward and Gray*, 1993, 1995]. The model that we use here is simple and has considerable empirical

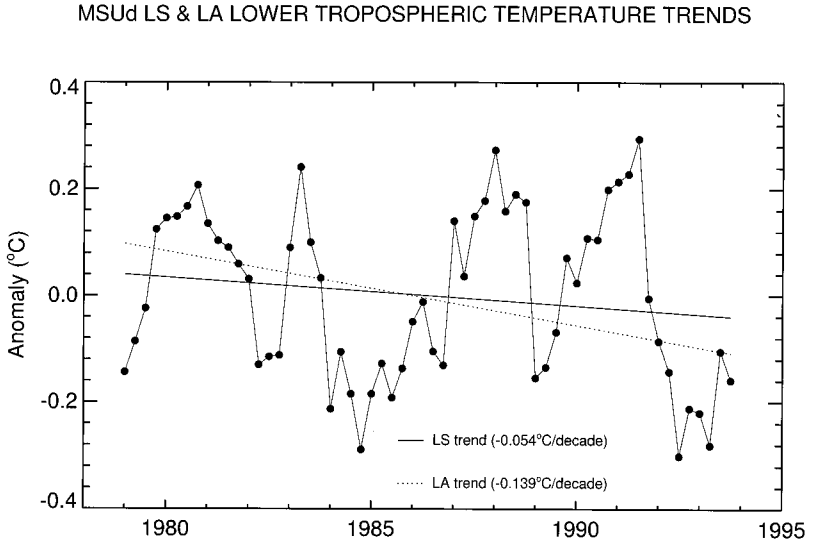

Figure 3. Time series of seasonal-mean lower tropospheric (2_{LT}) temperature anomalies (°C) over 1979–1993 in MSUd. Note that the LS and LA linear fits yield a trend difference of 0.085°C/decade.

Table 2. MSU/Radiosonde Lower Tropospheric Trends and Trend Differences Over 1979–1993 and 1979–1996

Comparison	1979–1993		1979–1996	
	LS	LA	LS	LA
MSUc	−0.049	−0.113	−0.033	−0.044
MSUd	−0.054	−0.139	−0.014	−0.045
HadRT1.1	+0.065	+0.042	+0.040	+0.020
HadRT1.2	−0.049	−0.075	−0.037	−0.076
ANGELL	−0.053	−0.053	−0.057	−0.084
MSUc minus HadRT1.1	−0.114	−0.155	−0.073	−0.064
MSUc minus HadRT1.2	0.000	−0.038	+0.004	+0.032
MSUc minus ANGELL	+0.004	−0.060	+0.024	+0.040
MSUd minus HadRT1.1	−0.119	−0.181	−0.054	−0.065
MSUd minus HadRT1.2	−0.005	−0.064	+0.023	+0.031
MSUd minus ANGELL	−0.001	−0.086	+0.043	+0.039

Lower tropospheric trend agreement between MSU and various radiosonde data sets. All trends (in °C/decade) were computed using global-mean seasonal-mean anomaly data and are given for two periods (1979–1993 and 1979–1996) and two methods of obtaining linear fits (LS and LA). The first five rows give the actual trends in MSUc, MSUd, HadRT1.1, HadRT1.2, and ANGELL. The last six rows indicate MSU/radiosonde trend differences.

justification based on results from extensive stochastic simulations (D. Nychka et al., manuscript in preparation, 2000).

We stress that our focus is on demonstrating the sensitivity of trend significance results to assumptions made in accounting for temporal autocorrelation. We do not address the possible causes of underlying trends in the atmospheric temperature series examined here and do not consider whether such trends are predominantly stochastic or deterministic in nature. Deducing cause and effect is hampered by (1) the short length (20 years or less) of the available deep-layer temperature time series, (2) forcing uncertainties and model errors, which lead to uncertainties in the climate-change signals associated with anthropogenic and natural external forcing, (3) inadequate knowledge of the statistical properties of such signals (i.e., a lack of ensembles of experiments with different forcing mechanisms), (4) large high-frequency noise contributions from natural modes of variability, such as El Niño, and (5) our poor understanding of possible linkages between anthropogenic forcing and changes in the frequency, intensity, and duration of El Niño [*Trenberth and Hoar*, 1996; *Timmermann et al*., 1999] and other natural modes of variability [*Corti et al*., 1999; *Hasselmann et al*., 1999]. For information on studies that specifically address possible causes of recent temperature changes in the free atmosphere, refer to *Karoly et al*. [1994], *Santer et al*. [1996a], *Tett et al*. [1996], *Hansen et al*. [1997, 1998], and *Bengtsson et al*. [1999].

4.1. Method

Consider a time series of global-mean seasonal-mean temperature anomalies, $x(t)$, for some specified atmospheric layer and data set. Here the time index t runs from DJF 1979 through SON 1993, so that the number of time samples in each series, n_t, is 60. The least squares linear regression estimate of the trend in $x(t)$, b, minimizes the squared differences between $x(t)$ and the regression line $\hat{x}(t)$

$$\hat{x}(t) = a + bt; \qquad t = 1, \cdots, n_t. \tag{1}$$

The regression residuals, $e(t)$, are defined as

$$e(t) = x(t) - \hat{x}(t); \qquad t = 1, \cdots, n_t. \tag{2}$$

For statistically independent values of $e(t)$, the standard error of b is defined as

$$s_b = \frac{s_e}{\left[\sum_{t=1}^{n_t} (t - \bar{t})^2\right]^{1/2}}, \tag{3}$$

where s_e^2, the variance of the residuals about the regression line, is given by

$$s_e^2 = \frac{1}{n_t - 2} \sum_{t=1}^{n_t} e(t)^2 \tag{4}$$

[see, e.g., *Wilks*, 1995]. Note that in some studies, it is implicitly (and often incorrectly) assumed that values of $e(t)$ are statistically independent [e.g., *Balling et al*., 1998].

Whether a trend in $x(t)$ is significantly different from zero is tested by computing the ratio between the estimated trend and its standard error

$$t_b = b/s_b. \tag{5}$$

Under the assumption that t_b is distributed as Student's t, the calculated t ratio is then compared with a critical t value, t_{crit}, for a stipulated significance level α and $n_t - 2$ degrees of freedom. If $e(t)$ is autocorrelated, this approach (henceforth referred to as "NAIVE") gives results that are too liberal; that is, it yields too frequent rejection of the null hypothesis $b = 0$ when compared with empirical expectations based on stochastic simulations (D. Nychka et al., manuscript in preparation, 2000).

If values of $e(t)$ are not statistically independent, as is often the case with temperature data (see Table 3), the NAIVE approach must be modified. There are various ways of accounting for temporal autocorrelation in $e(t)$ [see, e.g., *Wigley and Jones*, 1981; *Bloomfield and Nychka*, 1992; *Wilks*, 1995; *Ebisuzaki*, 1997; *Bretherton et al*., 1999]. The simplest way [*Bartlett*, 1935; *Mitchell et al*., 1966] uses an effective sample size n_e based on r_1, the lag-1 autocorrelation coefficient of $e(t)$:

$$n_e \approx n_t \frac{1 - r_1}{1 + r_1}. \tag{6}$$

Table 3. Lag-1 Autocorrelation Coefficients and Effective Sample Sizes

	2_{LT} Retrieval				Channel 2				Channel 4			
Data Set	SEAS (r_1)	ANN (r_1)	SEAS (n_e)	ANN (n_e)	SEAS (r_1)	ANN (r_1)	SEAS (n_e)	ANN (n_e)	SEAS (r_1)	ANN (r_1)	SEAS (n_e)	ANN (n_e)
NCEP (RT)	0.763	0.141	8	11	0.756	0.197	8	10	0.823	0.342	6	7
NCEP (WF)	0.766	0.155	8	11	0.759	0.216	8	10	0.821	0.339	6	7
NCMASK (WF)	0.668	0.193	12	10	0.700	0.334	11	7	0.728	0.213	9	10
ERA (RT)	0.756	0.245	8	9	0.787	0.273	7	9	0.825	0.377	6	7
ERA (WF)	0.764	0.237	8	9	0.793	0.271	7	9	0.823	0.369	6	7
ERMASK (WF)	0.683	0.221	11	10	0.730	0.357	9	7	0.667	0.095	12	12
MSUb	0.780	0.255	7	9	0.733	0.088	9	13	0.856	0.448	5	6
MSUc	0.796	0.284	7	8	0.755	0.126	8	12	0.856	0.446	5	6
MSUd	0.735	0.128	9	12	0.675	−0.050	12	17	0.852	0.399	5	6
MSUMASKc	0.712	0.332	10	8	0.681	0.284	11	8	0.775	0.325	8	8
HadRT1.1	0.688	0.211	11	10	0.696	0.354	11	7	0.645	0.080	13	13
HadRT1.2	0.652	0.096	13	12	0.677	0.147	12	11	0.781	0.120	7	12
ANGELL	0.597	0.108	15	12	—	—	—	—	0.733	0.442	9	6

Lag-1 autocorrelation coefficients (r_1) and effective sample sizes (n_e) for the global-mean seasonal-mean (SEAS) and annual-mean (ANN) anomaly data described in Table 1a. The actual sample size (n_t) is 60 for seasonal-mean data and 15 for annual-mean data. Effective sample sizes are reported to the nearest integer, but full precision was retained for calculating adjusted standard errors.

By substituting the estimated effective sample size n_e for n_t in (4), one obtains "adjusted" estimates of the standard deviation of regression residuals (s'_e) and hence of the standard error (s'_b) and t ratio (t'_b). We refer to this modification of the NAIVE approach as adjusted standard error (AdjSE). A third variant, AdjSE + Adjusted Degrees of Freedom (AdjSE + AdjDF), involves use of the effective sample size n_e not only in computation of the adjusted standard error but also in the indexing of the critical t value.

One interesting issue is whether r_1 should be estimated directly from $x(t)$ or from the regression residuals $e(t)$. In the presence of a large overall trend in $x(t)$, the former approach yields higher estimates of r_1, since the trend inflates the lag-1 autocorrelation. We examined the sensitivity of our significance test results (and of our adjusted confidence intervals; see section 5.1) to the choice of how r_1 is estimated and found this sensitivity to be small for the layer-average temperature time series used here. This reflects the fact that over the short period of the satellite record, the LS linear trends explain only a small portion of the overall variance of the time series. The large volcanic warming signatures (in the lower stratosphere) and the large amplitude variability associated with El Niño (in the troposphere) dominate the lag-1 autocorrelation, which is why we find that r_1 is not very different if estimated from $x(t)$ or $e(t)$. Here we have chosen to estimate r_1 from $e(t)$ and note that this choice leads to slightly smaller "adjusted" standard errors and a slightly more liberal test for the significance of the trend in $x(t)$.

4.2. Results

The effect of large positive values of r_1 is to inflate s_b in (3) and increase the width of the confidence interval about the estimated trend b. For the global-mean seasonal-mean anomaly data examined here, r_1 ranges from 0.597 (ANGELL, 2_{LT}) to 0.856 (MSUb and MSUc, channel 4) so that n_e ranges from 15 to 5 (1/4 to 1/12 of the actual sample size; see Table 3). Thus AdjSE is a more conservative test than NAIVE: Both have the same value of t_{crit}, but the former has a smaller calculated t ratio (if r_1 is nonzero). Temporal autocorrelation in $e(t)$ also has the consequence that AdjSE + AdjDF is a more conservative test than AdjSE: Both tests have the same calculated t value (t'_b), but the former has fewer degrees of freedom and hence a larger critical t value.

These systematic differences in computed significance levels are illustrated in Figure 4, which gives p values for tests of the null hypothesis of zero trend. Results are for global-mean seasonal-mean anomaly time series over 1979–1993 and are given for LS trends only. We first consider results for the lower stratosphere, where decisions on trend significance depend on the test assumptions. Using the NAIVE approach, all channel 4 trends that include volcanic effects (except MSUd) are significantly different from zero at the 10% level or better. The same trends fail to achieve significance at the 10% level with the AdjSE and AdjSE + AdjDF methods (the sole exception is the ANGELL result; see section 5.2). There are systematic differences between channel 4 results that include or exclude volcanic effects. Excluding volcanic influences generally enhances cooling of the lower stratosphere (see section 3.1) and markedly reduces the standard errors, so that channel 4 trends are systematically more significant than in the "volcanoes included" case.

In contrast, decisions on the significance of 2_{LT} and channel 2 trends are relatively insensitive to the significance testing method (Figure 4). This reflects the fact that tropospheric temperature trends over this short 15-year period are very small relative to the year-to-year variability, so that p values for all three significance testing methods are generally well above 0.10. Only three of the lower and midtropospheric trends over 1979–1993 (out of a possible 75) are significantly different from zero at the 10% level or better: the large negative channel 2 trends for MSUd and HadRT1.2 (−0.074 and −0.098°C/decade, respectively) and the positive ERA (RT) trend (+0.106°C/decade) for the 2_{LT} retrieval. In these three cases, trends are judged to be significantly different from zero with the NAIVE approach but not with AdjSE or AdjSE + AdjDF (see Figure 4).

The nonsignificance of the 2_{LT} trends is in most cases unaffected by the inclusion of more recent data. For example, while the inclusion of an additional 5 years of data increases the MSUd 2_{LT} trend from −0.054°C/decade over 1979–1993 to +0.061°C/decade over 1979–1998 (Table 4), the corresponding change in p value (estimated with the AdjSE + AdjDF approach) from 0.698 to 0.569 is relatively small. Neither trend is significant.

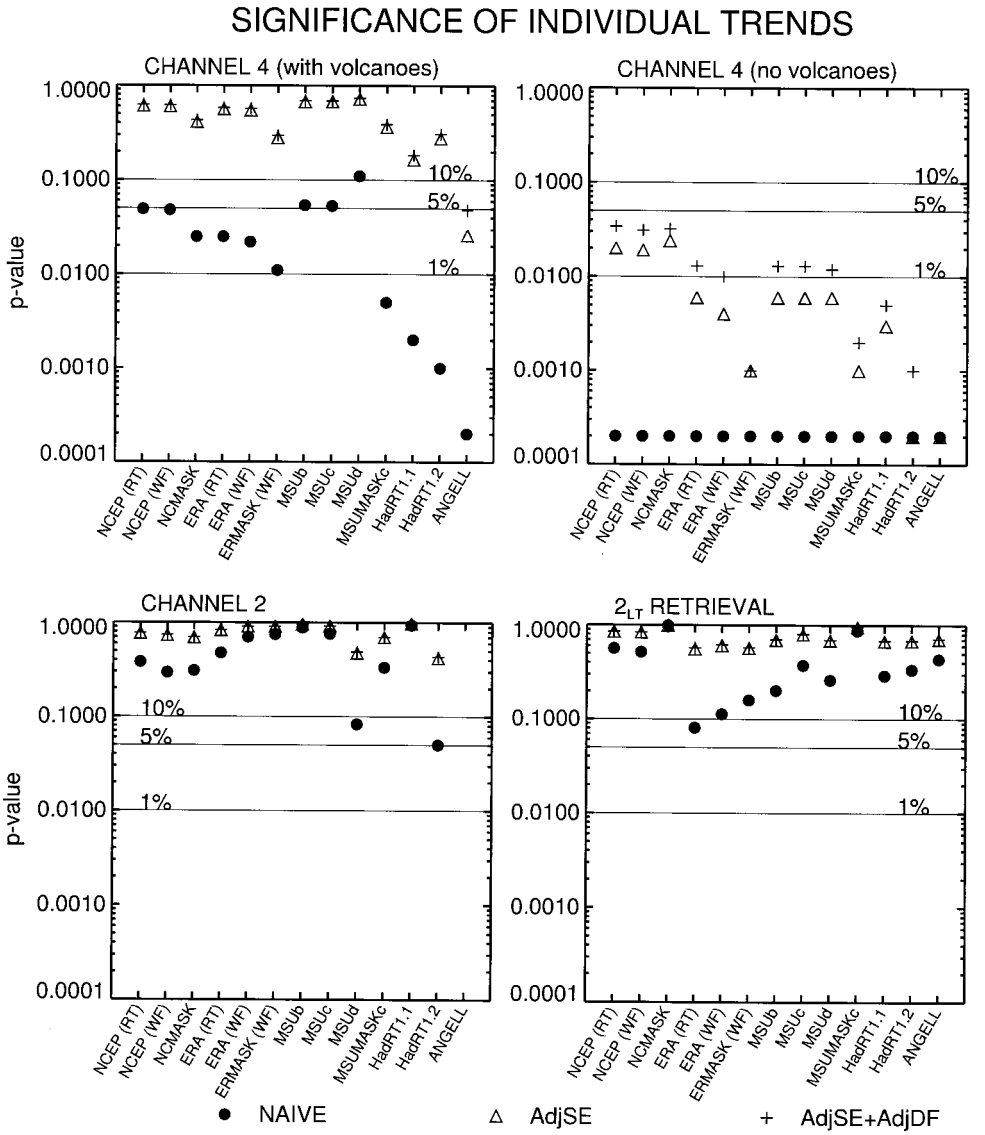

Figure 4. Significance of trends in individual time series of global-mean seasonal-mean temperature anomalies over 1979–1993. Results shown are p values for tests of the null hypothesis of zero trend, obtained using the NAIVE, AdjSE, and AdjSE + AdjDF approaches (see section 4.1). The nominal 1, 5, and 10% significance levels are indicated with thin horizontal lines. Note that for the channel 2 and 2_{LT} results, p values obtained with the AdjSE and AdjSE + AdjDF approaches are highly similar. In the case of channel 4 data that exclude volcanic effects, very small p values were set to 0.0002 to facilitate plotting.

Which of the three significance testing approaches outlined above yields results closest to theoretical expectations? This question is addressed by D. Nychka et al. (manuscript in preparation, 2000) using a stochastic simulation approach similar to that employed by *Zwiers and von Storch* [1995]. The latter study focused on accounting for temporal autocorrelation effects in the context of testing the significance of differences in overall means with one- and two-sample t tests. The work by D. Nychka et al. (manuscript in preparation, 2000) deals specifically with the issue of assessing trend significance in the presence of autocorrelated data. It indicates that AdjSE + AdjDF, while still liberal relative to empirically derived expectations, is nevertheless much closer to expected significance levels than either NAIVE (which performs worst) or AdjSE. We therefore concentrate on the discussion of AdjSE + AdjDF significance results in subsequent sections.

5. Statistical Significance of Trend Differences

We use two approaches to assess the significance of trend differences. Consider two time series, $x(t)$ and $y(t)$, with least squares linear trends b_x and b_y and estimated standard errors s_{bx} and s_{by}. In the first approach, we examine whether there is overlap between the regions defined by $b_x \pm s_{bx}$ and $b_y \pm s_{by}$ (or between the "adjusted" confidence intervals, $b_x \pm s'_{bx}$ and $b_y \pm s'_{by}$). The second method that we employ uses the difference time series $d(t) = x(t) - y(t)$ and then determines whether b_d, the trend in $d(t)$, is significantly different from zero. Operating on the difference time series reduces noise levels by subtracting variability common to $x(t)$ and $y(t)$. This facilitates identification of real trend differences that may exist between the two time series.

Note that different null hypotheses are being examined in

Table 4. Unadjusted and Adjusted 95% Confidence Intervals for MSUd Least Squares Trends in Lower Tropospheric Temperature

Period	Unadjusted		Adjusted	
	Seasonal	Annual	Seasonal	Annual
	LS Trend and 95% Confidence Interval			
1979–1993	−0.054 ± 0.093	−0.060 ± 0.169	−0.054 ± 0.304	−0.060 ± 0.219
1979–1997	−0.011 ± 0.062	−0.013 ± 0.108	−0.011 ± 0.171	−0.013 ± 0.138
1979–1998	+0.061 ± 0.069	+0.059 ± 0.127	+0.061 ± 0.229	+0.059 ± 0.156
	Lag-1 Autocorrelation, Actual or Effective Sample Size			
1979–1993	0.735 (60)	0.128 (15)	0.735 (9)	0.128 (12)
1979–1997	0.697 (76)	0.136 (19)	0.697 (14)	0.136 (14)
1979–1998	0.761 (80)	0.116 (20)	0.761 (11)	0.116 (16)
	Variance of Regression Residuals			
1979–1993	0.025	0.021	0.204	0.028
1979–1997	0.023	0.017	0.145	0.024
1979–1998	0.033	0.028	0.288	0.036

Sensitivity of unadjusted and adjusted 95% confidence intervals to length of record. Results are for MSUd least squares linear trends in lower-tropospheric temperature computed over three different intervals (1979–1993, 1979–1997, and 1979–1998). LS trends and confidence intervals are given in °C/decade and are based on both seasonal-mean and annual-mean anomaly data. Also shown for unadjusted and adjusted results are the lag-1 autocorrelation (r_1) of the regression residuals $e(t)$, the actual or effective sample sizes (n_t and n_e), and the variance of $e(t)$ (s_e^2 and $s_e^{2'}$).

these two approaches. In the first approach, we are testing whether the individual trends in $x(t)$ and $y(t)$ are drawn from the same population. In the second method, we are testing whether differences in data treatment (measurement methods, spatial coverage, the version of the dataset, or the methods used to compute an equivalent MSU temperature) have a significant effect on the trends.

5.1. Confidence Interval Method

Given the raw standard errors s_{bx} and s_{by}, the $P\%$ confidence intervals for b_x and b_y can be determined assuming that the sampling distributions of b_x and b_y are Gaussian. This is a reasonable assumption if the temporal sample size is large (>30), as in calculation of the unadjusted standard errors s_{bx} and s_{by} (where $n_t = 60$ seasons). In this case, the unadjusted 95% confidence interval is simply $b_x \pm 1.96\ (s_{bx})$, with the 95% confidence interval for b_y defined similarly.

However, for the seasonal-mean anomaly data considered here, values of n_e used for calculating the adjusted standard errors are invariably $\ll$30 (see Table 3). To determine the 95% confidence intervals for s'_{bx} and s'_{by}, it is more appropriate to assume that b_x and b_y are distributed as Student's t. Since the t distribution gives greater "weight" (i.e., assigns greater probability) to the tails than the normal distribution [see, e.g., *Wilks*, 1995], the small-sample confidence intervals estimated with the t distribution are wider than the corresponding confidence intervals estimated with the normal distribution. For $n_e = 5$, for example, (the smallest effective sample size in Table 3), the estimated 95% confidence interval is $b_x \pm 2.57\ (s'_{bx})$, which is nearly 30% larger than in the normal distribution case.

In the following, we assume that b_x and b_y are normally distributed for calculating unadjusted 95% confidence intervals. Adjusted 95% confidence intervals are calculated by inverting Student's t distribution to obtain t_{inv} for n_e degrees of freedom and $p = 0.975$ (two-tailed test). The adjusted 95% confidence interval is simply $b_x \pm t_{\text{inv}}\ (s'_{bx})$.

5.1.1. Confidence intervals for 2_{LT} retrieval and channel 2.

For the 2_{LT} retrieval the unadjusted 95% confidence intervals for LS trends over 1979–1993 range from ±0.093 (MSUd) to ±0.136°C/decade (MSUMASKc; see Figure 5). The adjusted intervals are much larger, by a factor of 2–4, and range from ±0.255 (HadRT1.2) to ±0.456°C/decade (MSUc). All of the unadjusted and adjusted 95% confidence intervals encompass zero and include positive and negative values (Figure 5). There is considerable overlap between the adjusted 95% confidence intervals for all 13 data sets. Even without performing a formal statistical test, it is evident that we cannot reject the null hypothesis that the individual 2_{LT} trends in the satellite, reanalysis, and radiosonde data are drawn from the same population.

However, it is important to note that the large adjusted 95% confidence intervals also include within them the expected trends due to anthropogenic forcing [see *Santer et al*., 1996b]. While we cannot reject the hypothesis of no trend in the 2_{LT} time series, neither can we claim that the observed trends over 1979–1993 differ significantly from model projections.

Similar results are obtained for midtropospheric temperature trends (Figure 6). Again, even the (smaller) unadjusted 95% confidence intervals overlap for all 12 time series. Confidence intervals are very similar to those obtained for the 2_{LT} retrieval and range from ±0.082 (MSUd) to ±0.118°C/decade (NCMASK) for unadjusted intervals and from ±0.224 (MSUd) to ±0.472°C/decade (ERA (WF)) for adjusted intervals.

These results show the need for caution in interpreting the previously reported MSUc/ANGELL and MSUc/HadRT1.2 agreement of a few hundredths of a degree C/decade for lower tropospheric trends [*Christy et al*., 1997, 1998]. While such agreement may indicate common low-frequency behavior in the MSUc, ANGELL, and HadRT1.2 data sets, it may also be fortuitous. The large confidence intervals found here highlight the significant uncertainties associated with all of these trend estimates.

The converse of this is that a 2_{LT} trend difference of

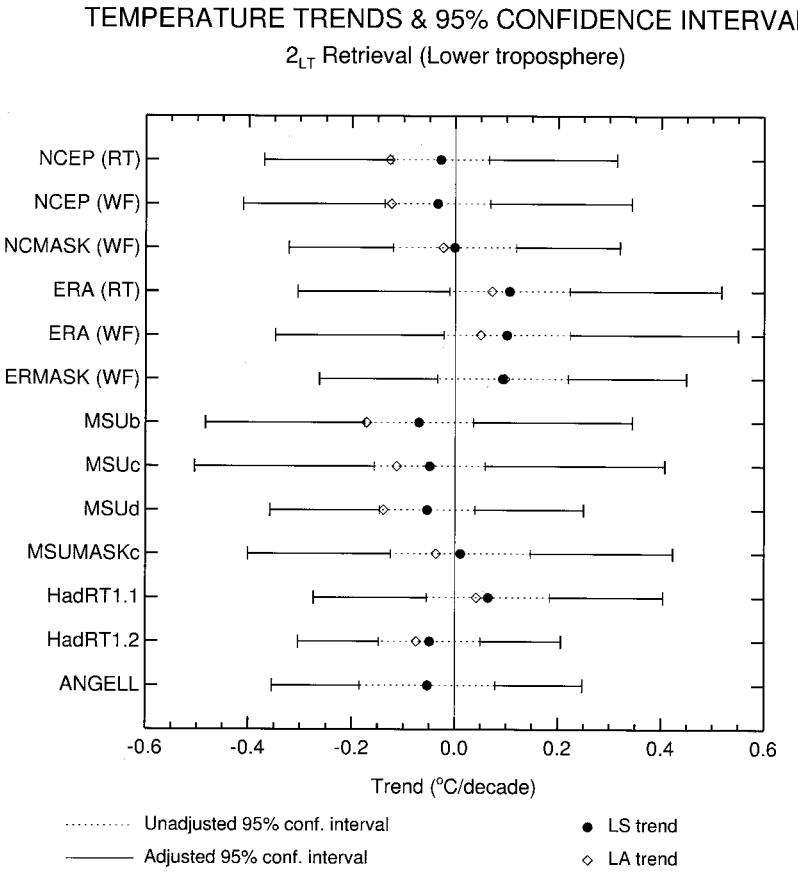

Figure 5. Unadjusted and adjusted 95% confidence intervals for least squares linear trends in lower tropospheric temperature (2_{LT} retrieval). All confidence intervals were computed with global-mean seasonal-mean anomaly data spanning the period 1979–1993. The smaller, unadjusted confidence intervals do not account for temporal autocorrelation and are estimated with the normal distribution. The adjusted 95% confidence intervals (shown here as extensions to the unadjusted intervals) account for temporal autocorrelation in the data and are estimated with the Student's t distribution (see section 5.1). Trends computed with the LS and LA approaches are also shown.

~0.12°C/decade, such as the difference between LS trend estimates for MSUd and HadRT1.1 (Table 1a), is still well within the adjusted 95% confidence intervals of the individual MSUd and HadRT1.1 trends. While there may be physical reasons for concluding that the two trend estimates are inconsistent, we could not reach this conclusion by examination of their standard errors alone.

5.1.2. Confidence intervals for channel 4. Confidence interval ranges for lower stratospheric temperature trends over 1979–1993 are considerably larger than those obtained for deep-layer temperatures in the lower and midtroposphere (compare Figure 7 and Figures 5 and 6). This is due in part to the large lower stratospheric warming signatures of El Chichón and Pinatubo (see Figure 1 and section 3.1). Values range from ±0.190 (HadRT1.2) to ±0.297°C/decade (ANGELL) and from ±0.523 (HadRT1.1) to ±1.495°C/decade (MSUb) for unadjusted and adjusted intervals, respectively. In one case (ANGELL), the unadjusted 95% confidence interval does not overlap with the unadjusted intervals from the other 12 time series, although the adjusted intervals do overlap (see Figure 7 and section 3.1). For all data sets, both adjusted and unadjusted confidence intervals decrease when volcanic effects are excluded (see section 3.1).

5.1.3. Sensitivity to sampling interval. Are our confidence interval estimates sensitive to the selected sampling interval? To address this issue, we calculated adjusted 95% confidence intervals from annual-mean 2_{LT} anomaly data over 1979–1993 and then compared these with results based on seasonal-mean anomalies (Figure 8). The choice of sampling interval does not alter our primary conclusion that the adjusted 95% confidence intervals overlap in all 13 data sets considered here and consistently encompass both zero and model projections. There is, however, a systematic difference between the adjusted 95% confidence intervals based on seasonal-mean and annual-mean data, with the latter smaller in 11 out of 13 cases. The maximum difference is ~30% for the NCEP (RT) data. Note that the difference in sampling interval has only a very small effect (a few hundredths of a degree Celsius or less) on LS trend estimates.

5.1.4. Sensitivity to inclusion of recent data. The preceeding discussion focused on estimating confidence intervals for LS trends over one specific 15-year period (1979–1993). We next examine the sensitivity of confidence interval estimates to the incorporation of more recent data. We do this only for one data set (MSUd) and one atmospheric layer (the 2_{LT} lower tropospheric retrieval), comparing confidence inter-

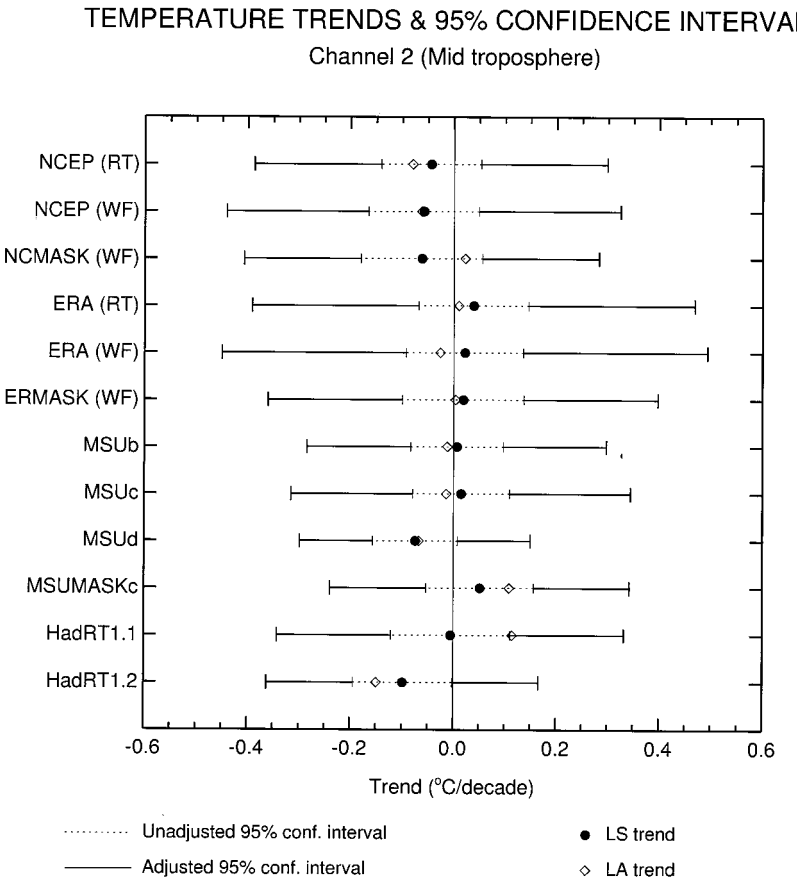

Figure 6. As for Figure 5, but for the midtroposphere (channel 2).

vals computed over three different periods: 1979–1993, 1979–1997, and 1979–1998 (Table 4).

In all four cases (i.e., for unadjusted and adjusted 95% confidence intervals based on seasonal- and annual-mean data), the width of the confidence intervals decreases from 1979–1993 to 1979–1997. This is due to an increase in actual and effective sample size and a decrease in the unadjusted and adjusted variance of the regression residuals. When data for 1998 are included, there is an increase in both the variance of the regression residuals and the width of confidence intervals (relative to results for 1979–1997; see Table 4). This is largely due to the strong El Niño event in 1998.

Sensitivity of the estimated confidence intervals to the length of record is comparatively small (~20–30%), at least for the MSUd 2_{LT} data. Our previous conclusion that the adjusted 95% confidence intervals for short timescale 2_{LT} trends are large and encompass both zero and model predictions is therefore robust.

5.2. Difference Series Method

An alternative method to identify small trend differences embedded in noisy time series is to examine the difference time series $d(t) = x(t) - y(t)$. In our case, it is meaningful to consider pairwise differences in $x(t)$ and $y(t)$, since both purportedly represent temperature fluctuations in the same atmospheric layer and over the same time period. Differencing facilitates identification of overall trend differences by removing variability that is common to both time series. An analysis of this kind is typical of statistical assessments of the effects of physical or chemical treatments [e.g., *Dixon and Massey*, 1983]. Here we consider whether there are significant trend differences that may be related to differences in the system used to estimate temperature, in the version of the data set, in spatial coverage, and in the method used to compute an equivalent MSU temperature.

To determine whether any trend b_d in $d(t)$ is significantly different from zero, we proceed as in eqs. (1)–(6) but now substituting $d(t)$ for $x(t)$. To assess the significance of b_d we use the same three approaches outlined in section 4.1: NAIVE, AdjSE, and AdjSE + AdjDF.

As noted in section 4.1 for tests of the significance of individual trends, these three methods yield systematic differences in computed significance levels, with NAIVE the most and AdjSE + AdjDF the least liberal approach. The same systematic differences are found in tests of the significance of trends in $d(t)$. This is evident from Figure 9, which shows numerous instances where decisions on the significance of a trend difference (at a stipulated significance level of $\alpha = 0.01$ or $\alpha = 0.05$) depend on the choice of significance testing method. We concentrate here on the AdjSE + AdjDF significance results, which are in closest accord with empirical expectations from stochastic simulations (D. Nychka et al., manuscript in preparation, 2000).

Tables 5a–5c give the least squares linear trends in all possible $d(t)$ pairs, together with p values for the null hypothesis of zero trend in $d(t)$. All results are for trends over 1979–1993 computed with global-mean seasonal-mean anomaly data. First consider results for the 2_{LT} retrieval. We showed in sec-

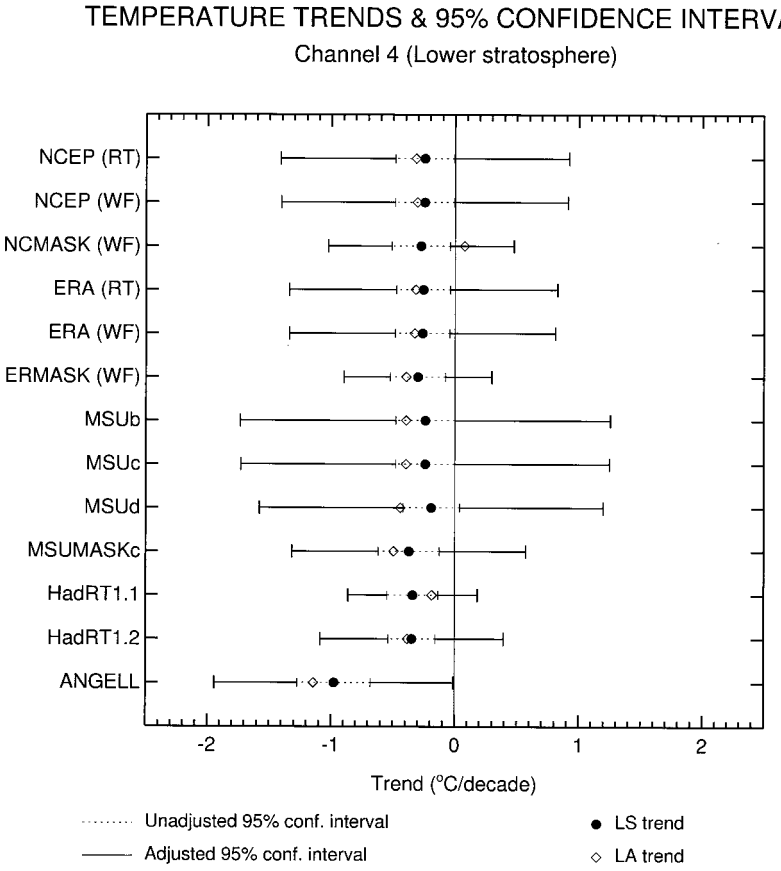

Figure 7. As for Figure 5, but for the lower stratosphere (channel 4).

tion 5.1 that the least squares linear trends in all 13 individual data sets had strongly overlapping 95% confidence intervals. Tests of the trend in $d(t)$, however, reveal that there are significant trend differences between ERA and all other data sets except HadRT1.1 and MSUMASKc (which like ERA have a positive 2_{LT} trend over 1979–1993). Thus the use of paired differences has enabled us to discern small but significant trend differences that were not obvious when the individual time series were considered.

We infer from this result that differences in measurement systems can have a significant impact on lower tropospheric temperature trends. Other factors do not yield significant differences in 2_{LT} trends, although the trend difference between HadRT1.1 and HadRT1.2 (largely related to coverage differences; see S99) is significant at the 5% level.

For channel 2, most of the trend differences significant at the 5% level or better involve HadRT1.2 and MSUMASKc (Table 5b). These have the largest negative and positive midtropospheric trends over 1979–1993 (−0.098 and +0.052°C/decade, respectively; see Table 1a). The channel 2 trend differences between MSUb (+0.007°C/decade) and MSUd (−0.074°C/decade) and between MSUc (+0.015°C/decade) and MSUd are significant at the 5% level. The large trend differences between the latest MSU version and the earlier two are due to adjustments made to MSUd for changes in instrument body temperature and east-west drift of satellites. These adjustments are thought to have had a net cooling effect (J. Christy, personal communication, 1999). In the MSUd 2_{LT} data they are offset by a correction for the orbital decay effect identified by *Wentz and Schabel* [1998]. Since this decay effect does not influence channel 2, there is no compensating adjustment in the MSUd channel 2 data.

For channel 4, all difference series tests involving the ANGELL data set yield trend differences that are significant at the 1% level (Table 5c). Possible explanations for the much larger lower-stratospheric cooling in ANGELL are reviewed by *Angell* [1999] and S99. These include limited and spatially nonuniform coverage of the 63-station ANGELL network [*Trenberth and Olson*, 1991], instrumental inhomogeneities in the ANGELL data [*Gaffen*, 1994], and differences in the atmospheric layers sampled by the channel 4 weighting function and ANGELL's 100–50 hPa layer-mean virtual temperature (see section 2.4).

Subsampling the MSUc channel 4 data with the HadRT1.1 data mask leads to a large (+0.131°C/decade) and marginally significant trend in the difference relative to the spatially complete MSUc data (see MSUc-MSUMASKc comparison in Table 5c). Significant differences are not evident in NCEP-NCMASK or ERA-ERMASK comparisons. Coverage differences have a relatively greater effect in MSUc than in NCEP and ERA since MSUc channel 4 has larger (and spatially more coherent) warming in the tropics than either ERA or NCEP (Figure 10). The HadRT1.1 coverage mask (see S99) removes much of this warming, hence the large change in the lower-stratospheric trend in MSUc, from −0.241 to −0.371°C/decade (MSUc versus MSUMASKc; see Table 1a) and the much smaller decreases in ERA and NCEP.

The trends for differences between MSUd on the one hand

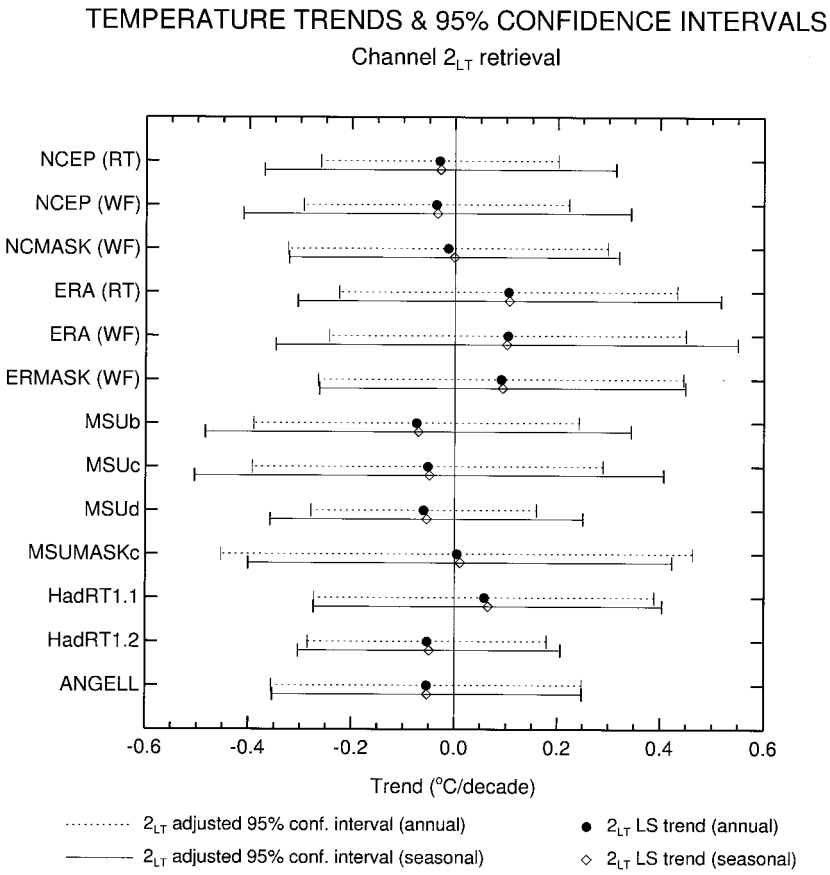

Figure 8. Sensitivity of confidence intervals for 2_{LT} least squares linear trends to sampling interval. Trends and adjusted 95% confidence intervals (see section 5.1) were computed with both seasonal-mean and annual-mean anomaly data spanning the period 1979–1993.

and MSUb and MSUc on the other hand are also significant at the 5% level. As noted above for the channel 2 results, these differences are probably related to the adjustments made to MSUd for changes in instrument body temperature and east-west drift effects. Another interesting and curious aspect of the channel 4 results is that the trend in the MSUb-MSUc difference time series is significant at the 1% level, even though it is very small (+0.001°C/decade; see Table 5c). The fact that we judge this trend in $d(t)$ to be statistically significant indicates that there is a systematic component to MSUb-MSUc differences and that the adjustments made to MSUc did have an effect on its lower-stratospheric trends. However, a trend in $d(t)$ of 0.001°C/decade is of no practical importance when compared with trend uncertainties arising from other sources (see S99).

6. Conclusions

Several recent investigations have attempted to improve our understanding of observational uncertainties of temperatures close to the Earth's surface [*Jones et al*., 1997] and in the free atmosphere [*Gaffen et al*., 2000; *Santer et al*., 1999]. The latter study (S99) focused on uncertainties arising from the system used to monitor temperature (satellites, radiosondes, and reanalysis), the method used to compute an equivalent MSU temperature, the adjustments made to individual versions of a specific data set, and from differences in the coverage of the data sets.

The present work complements the earlier study by S99. It addresses two main issues. The first relates to the sensitivity of linear trends to the selected fitting method. The second deals with the significance of trends and trend differences and with the question of which procedures one might use in order to determine significance.

The first issue, trend sensitivity to the fitting method, has also been considered by *Gaffen et al*. [2000] in the context of radiosonde-derived temperature records. They found a relatively small sensitivity (0.03°C/decade or less) to the use of two different methods to compute radiosonde-based trends over 1959–1995 (least squares and "median pairwise slopes"). Our results, obtained using both a least squares approach (LS) and minimization of absolute deviations (LA), are consistent with those of *Gaffen et al*. [2000] when we consider trends over a comparable period. When we compute trends over a much shorter period (1979–1993) than Gaffen et al. examined, we find that the LS and LA methods can yield trend differences exceeding 0.10°C/decade.

Furthermore, we find that there are systematic differences between the 1979–1993 trends estimated with LS and LA approaches, particularly for trends in the lower stratosphere (channel 4) and lower troposphere (2_{LT} retrieval), where LA trends are generally more negative or less positive, respectively. These systematic differences are most likely related to the uneven temporal distribution of volcanically induced warming events in the lower stratosphere and El Niño events in

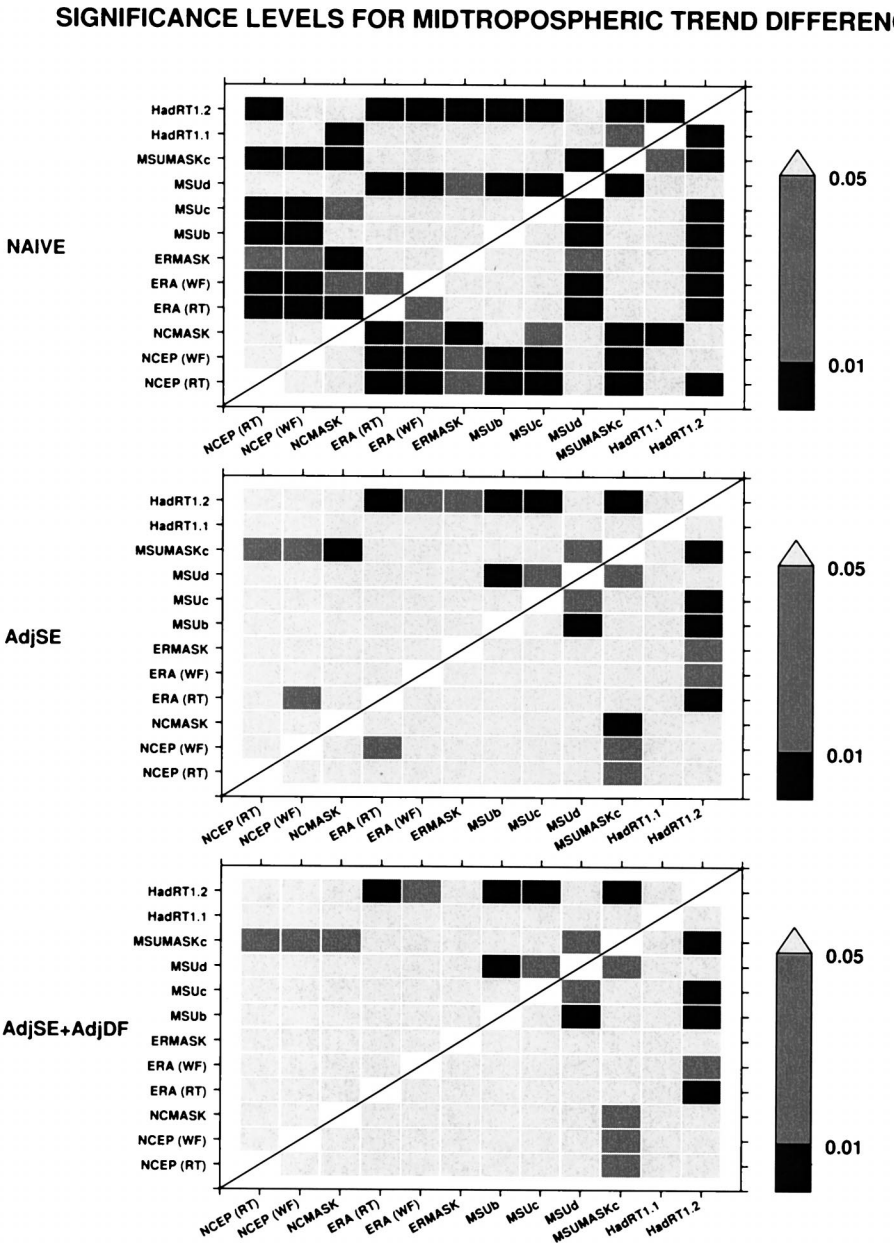

Figure 9. Significance of midtropospheric (channel 2) trend differences between various data sets, as assessed by the "NAIVE," "AdjSE," and "AdjSE + AdjDF" approaches (see section 4). All three test the significance of the trend in the difference time series $d(t)$. Tests involve global-mean seasonal-mean anomaly data spanning the period 1979–1993. Matrices are symmetrical about the diagonal line.

the lower troposphere. In the lower stratosphere, LA/LS trend differences are strongly reduced by removal of the temperature effects of El Chichón and Pinatubo.

Such issues are highly relevant in interpreting the results of previously published MSU/radiosonde trend comparisons [e.g., *Christy et al*., 1997, 1998], which have noted a close correspondence between the 2_{LT} trends in MSUc and in several radiosonde data sets (HadRT1.2 and ANGELL). These previous comparison relied solely on LS trend estimates. We find here that the LA approach systematically degrades the lower-tropospheric trend correspondence between MSUc and radiosonde data. A similar result was obtained by S99, who found that accounting for coverage differences degraded MSU/radiosonde trend correspondence in the midtroposphere. That study and the current investigation point toward the need for some caution in interpreting the results of satellite/radiosonde trend comparisons.

Is a trend in data set $x(t)$ significantly different from zero or from that in data set $y(t)$? This is the next issue that we have addressed. We used three different methods to assess the significance of individual trends and trend differences. These methods differ in terms of how they account for temporal autocorrelation effects. The first of these, NAIVE, does not account for temporal autocorrelation of the data being tested. The second, Adjusted Standard Error (AdjSE), uses an estimate of the lag-1 autocorrelation of the data to determine an effective sample size n_e. This in turn is used to adjust estimates of the standard error and calculated t value. The third ap-

Table 5a. Difference Time Series, 2_{LT} Retrieval: Linear Trends and Trend Significance

	2_{LT} Retrieval (1979–1993)											
	NCEP (WF)	NCMASK (WF)	ERA (RT)	ERA (WF)	ERMASK (WF)	MSUb	MSUc	MSUd	MSUMASK (c)	HadRT1.1	HadRT1.2	ANGELL
NCEP (RT)	+0.006 (0.621)	−0.027 (0.578)	−0.133 (0.005)*	−0.129 (0.019)†	−0.120 (0.033)†	+0.042 (0.120)	+0.022 (0.558)	+0.026 (0.244)	−0.039 (0.606)	−0.093 (0.112)	+0.021 (0.517)	+0.025 (0.511)
NCEP (WF)		−0.034 (0.493)	−0.140 (0.002)*	−0.135 (0.006)*	−0.127 (0.022)†	+0.036 (0.146)	+0.015 (0.652)	+0.020 (0.489)	−0.045 (0.543)	−0.099 (0.091)‡	+0.015 (0.687)	+0.019 (0.567)
NCMASK (WF)			−0.106 (0.098)‡	−0.102 (0.141)	−0.093 (0.108)	+0.069 (0.136)	+0.049 (0.342)	+0.053 (0.284)	−0.012 (0.789)	−0.066 (0.006)*	+0.048 (0.329)	+0.052 (0.326)
ERA (RT)				+0.004 (0.705)	+0.013 (0.748)	+0.175 (0.000)*	+0.155 (0.000)*	+0.160 (0.006)*	+0.094 (0.179)	+0.040 (0.583)	+0.155 (0.010)*	+0.159 (0.001)*
ERA (WF)					+0.009 (0.842)	+0.171 (0.000)*	+0.151 (0.001)*	+0.155 (0.021)†	+0.090 (0.215)	+0.036 (0.646)	+0.150 (0.025)†	+0.154 (0.001)*
ERMASK (WF)						+0.162 (0.001)*	+0.142 (0.003)*	+0.146 (0.013)†	+0.081 (0.013)†	+0.027 (0.680)	+0.141 (0.025)†	+0.145 (0.013)†
MSUb							−0.020 (0.557)	−0.016 (0.649)	−0.081 (0.186)	−0.135 (0.013)†	−0.021 (0.507)	−0.017 (0.638)
MSUc								+0.004 (0.909)	−0.061 (0.299)	−0.115 (0.063)‡	−0.001 (0.989)	+0.003 (0.930)
MSUd									−0.065 (0.363)	−0.119 (0.044)†	−0.005 (0.842)	−0.001 (0.981)
MSUMASK (c)										−0.054 (0.291)	+0.060 (0.414)	+0.064 (0.363)
HadRT1.1											+0.114 (0.027)†	+0.118 (0.036)†
HadRT1.2												+0.004 (0.924)

Significance of lower tropospheric (2_{LT}) trend differences between various data sets. The test uses $d(t)$, the time series of differences between individual pairs of data sets (e.g., between NCEP (RT) minus NCEP (WF), NCEP (RT) minus NC-MASK (WF), etc.). The test procedure, AdjSE + AdjDF, involves a one-sample t test, modified to account for autocorrelation in $d(t)$ (see sections 4.1 and 5.2). Least squares linear trends in $d(t)$, computed with the 1979–1993 global-mean seasonal-mean anomaly data described in Table 1a, are given (in °C/decade). The numbers in parentheses are the p values for the null hypothesis that the trend in $d(t)$ is not significantly different from zero.

*Trends in $d(t)$ achieving significance at 1% level.

†Trends in $d(t)$ achieving significance at 5% level.

‡Trends in $d(t)$ achieving significance at 10% level.

proach, AdjSE + Adjusted Degrees of Freedom (AdjSE + AdjDF), involves use of the effective sample size n_e not only in computation of the adjusted standard error and calculated t value but also in the indexing of the critical t value.

There are systematic differences in the significance levels yielded by these three approaches, with NAIVE the least conservative and AdjSE + AdjDF the most conservative test. We find that decisions on trend significance can depend critically on the choice of test, particularly for individual trends in lower-stratospheric temperatures. The AdjSE + AdjDF test is our preferred method and gives significance results that are in closest accord with empirical expectations based on stochastic simulations (D. Nychka et al., manuscript in preparation, 2000). Using this test, we find that none of the individual 1979–1993 trends in deep-layer temperatures is significantly different from zero. This result holds for virtually all data sets and atmospheric regions that we consider. In all data sets, individual (cooling) trends in lower-stratospheric temperatures become significant if volcanic effects are first removed from the time series.

For assessing the significance of trend differences, we used two complementary approaches, the "confidence interval" and "difference series" methods. In the former, we compute the "unadjusted" and "adjusted" 95% confidence intervals for LS linear trend estimates. The unadjusted intervals are based on a large-sample normal approximation, while the larger adjusted intervals account for temporal autocorrelation effects (through the effective sample size) and rely on a small-sample t distribution approximation. For the adjusted 95% confidence intervals, there is always overlap between the intervals estimated for different data sets. This holds for all three atmospheric regions considered here. It also holds, in all cases except ANGELL's lower-stratospheric trend over 1979–1993, for the unadjusted 95% confidence intervals. In virtually all cases, therefore, we cannot reject the null hypothesis that the trends in the individual satellite, radiosonde, and reanalysis data sets are drawn from the same population. Our results also show that for the 2_{LT} trends over 1979–1993, the large adjusted 95% confidence intervals for all data sets encompass both zero and the model-projected trends due to anthropogenic effects. This conclusion does not depend on whether adjusted confidence intervals are computed with seasonal-mean or annual-mean data.

The "confidence interval" test is not an efficient way of discerning relatively small trend differences that are embedded

Table 5b. Difference Time Series, Channel 2: Linear Trends and Trend Significance

	Channel 2 (1979–1993)										
	NCEP (WF)	NCMASK (WF)	ERA (RT)	ERA (WF)	ERMASK (WF)	MSUb	MSUc	MSUd	MSUMASK (c)	HadRT1.1	HadRT1.2
NCEP (RT)	+0.014 (0.618)	+0.018 (0.661)	−0.083 (0.074)‡	−0.066 (0.270)	−0.063 (0.215)	−0.050 (0.103)	−0.058 (0.079)†	+0.030 (0.577)	−0.096 (0.018)†	−0.039 (0.449)	+0.054 (0.055)‡
NCEP (WF)		+0.004 (0.917)	−0.097 (0.053)‡	−0.080 (0.140)	−0.077 (0.162)	−0.064 (0.116)	−0.072 (0.069)†	+0.016 (0.803)	−0.110 (0.020)†	−0.053 (0.380)	+0.041 (0.292)
NCMASK (WF)			−0.101 (0.118)	−0.084 (0.232)	−0.081 (0.324)	−0.068 (0.309)	−0.076 (0.267)	+0.012 (0.883)	−0.114 (0.013)†	−0.057 (0.233)	+0.036 (0.523)
ERA (RT)				+0.017 (0.393)	+0.020 (0.540)	+0.032 (0.532)	+0.024 (0.546)	+0.113 (0.213)	−0.013 (0.756)	+0.044 (0.568)	+0.137 (0.002)*
ERA (WF)					+0.003 (0.927)	+0.016 (0.808)	+0.008 (0.873)	+0.096 (0.362)	−0.030 (0.552)	+0.027 (0.762)	+0.121 (0.024)†
ERMASK (WF)						+0.012 (0.856)	+0.005 (0.942)	+0.093 (0.294)	−0.033 (0.377)	+0.024 (0.824)	+0.117 (0.061)‡
MSUb							−0.008 (0.799)	+0.080 (0.002)*	−0.046 (0.368)	+0.011 (0.875)	+0.105 (0.000)*
MSUc								+0.088 (0.025)†	−0.038 (0.433)	+0.019 (0.800)	+0.113 (0.000)*
MSUd									−0.126 (0.048)†	−0.069 (0.444)	+0.025 (0.481)
MSUMASK (c)									+0.057	+0.057 (0.307)	+0.150 (0.001)*
HadRT1.1											+0.094 (0.085)‡

As for Table 5a, but for midtropospheric (channel 2) trend differences between various data sets.

Table 5c. Difference Time Series, Channel 4: Linear Trends and Trend Significance

	Channel 4 (1979–1993)											
	NCEP (WF)	NCMASK (WF)	ERA (RT)	ERA (WF)	ERMASK (WF)	MSUb	MSUc	MSUd	MSUMASK (c)	HadRT1.1	HadRT1.2	ANGELL
NCEP (RT)	+0.002 (0.731)	+0.032 (0.701)	+0.012 (0.854)	+0.019 (0.767)	+0.056 (0.614)	−0.005 (0.957)	−0.004 (0.967)	−0.054 (0.493)	+0.127 (0.166)	+0.096 (0.478)	+0.103 (0.349)	+0.732 (0.000)*
NCEP (WF)		+0.030 (0.709)	+0.009 (0.884)	+0.017 (0.796)	+0.054 (0.622)	−0.007 (0.941)	−0.006 (0.950)	−0.056 (0.492)	+0.124 (0.175)	+0.093 (0.474)	+0.101 (0.344)	+0.730 (0.000)*
NCMASK (WF)			−0.021 (0.832)	−0.013 (0.892)	+0.024 (0.783)	−0.037 (0.787)	−0.036 (0.793)	−0.087 (0.480)	+0.094 (0.451)	+0.063 (0.541)	+0.071 (0.317)	+0.700 (0.000)*
ERA (RT)				+0.008 (0.081)‡	+0.044 (0.643)	−0.017 (0.803)	−0.015 (0.815)	−0.066 (0.139)	+0.115 (0.138)	+0.084 (0.527)	+0.092 (0.333)	+0.720 (0.000)*
ERA (WF)					+0.037 (0.693)	−0.024 (0.731)	−0.023 (0.741)	−0.073 (0.120)	+0.107 (0.165)	+0.076 (0.558)	+0.084 (0.362)	+0.713 (0.000)*
ERMASK (WF)						−0.061 (0.652)	−0.060 (0.657)	−0.110 (0.342)	+0.071 (0.568)	+0.040 (0.667)	+0.047 (0.348)	+0.676 (0.000)*
MSUb							+0.001 (0.006)*	−0.049 (0.017)†	+0.132 (0.078)‡	+0.101 (0.577)	+0.108 (0.447)	+0.737 (0.000)*
MSUc								−0.050 (0.013)†	+0.131 (0.081)‡	+0.099 (0.580)	+0.107 (0.450)	+0.736 (0.000)*
MSUd									+0.181 (0.012)†	+0.150 (0.353)	+0.158 (0.204)	+0.786 (0.000)*
MSUMASK (c)										−0.031 (0.894)	−0.023 (0.853)	+0.605 (0.000)*
HadRT1.1											+0.008 (0.891)	+0.636 (0.004)*
HadRT1.2												+0.629 (0.001)*

As for Table 5a, but for lower-stratospheric (channel 4) trend differences between various data sets.

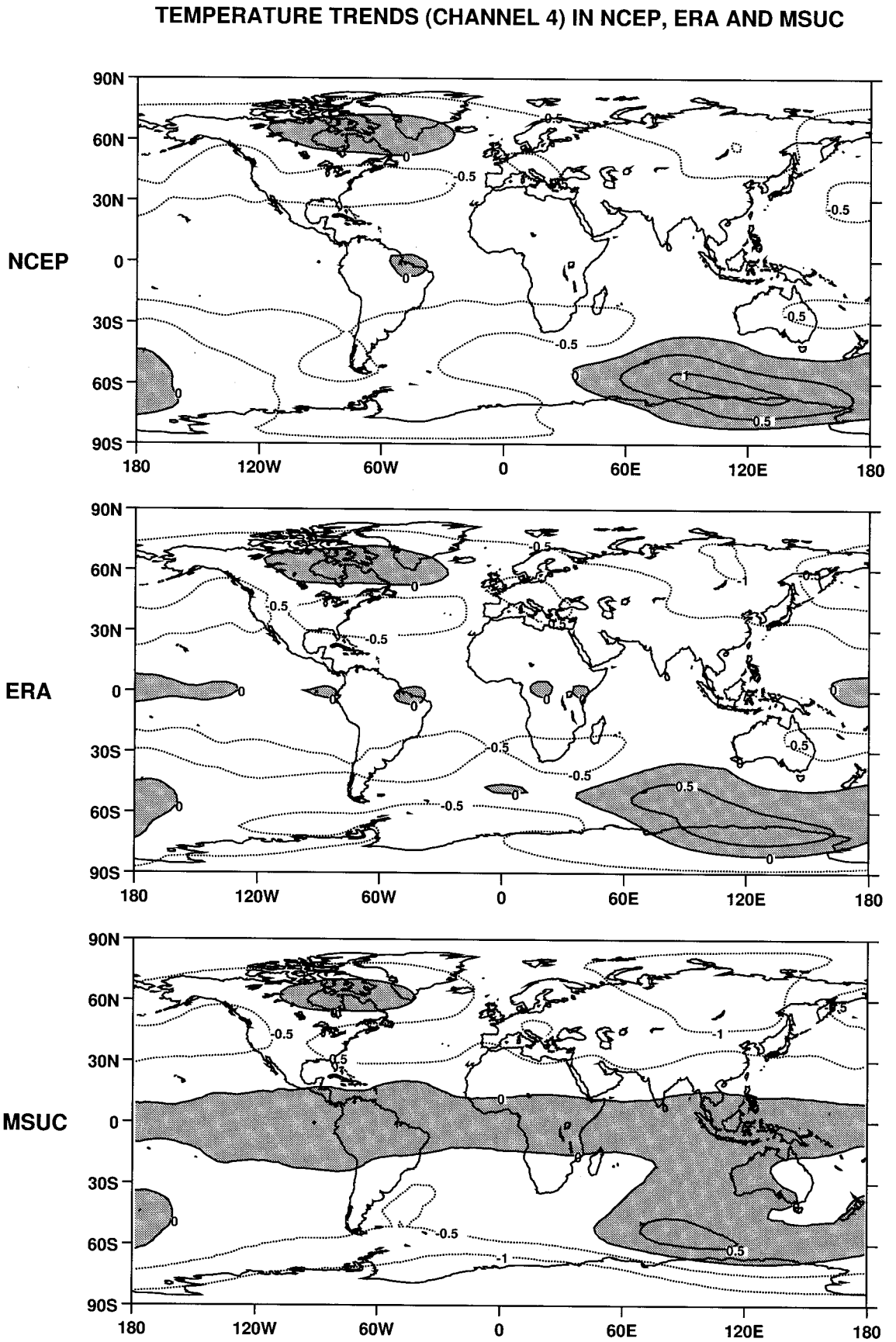

Figure 10. Lower-stratospheric temperature trends over 1979–1993 (in °C/decade) in the NCEP, ERA, and MSUc data. Equivalent channel 4 temperatures from NCEP and ERA were computed with a radiative transfer code. The contour interval is 0.5°C, and areas with positive changes are shaded.

in noisy time series. In the "difference series" approach we use $d(t)$, the time series of paired differences between $x(t)$ and $y(t)$. This markedly reduces noise levels by subtracting variability components common to $x(t)$ and $y(t)$ and facilitates the identification of trend differences arising from different data treatment methods. We then test whether the LS trend in $d(t)$ is significantly different from zero, using the same three approaches employed for testing significance of individual trends.

For the lower tropospheric trends, our preferred approach (AdjSE + AdjDF) indicates that significant trend differences exist between ERA and all other data sets except HadRT1.1 and MSUMASKc (which have a positive 2_{LT} trend over 1979–1993, like ERA). In the midtroposphere, most of the trend differences significant at the 5% level or better involve HadRT1.2 and MSUMASKc, the two data sets with the largest negative and positive midtropospheric trends, respectively, over 1979–1993. Results for the lower stratosphere indicate that ANGELL's trend over 1979–1993 differs significantly from that in all other data sets. In most cases, the only factor that produces significant trend differences is the difference in the system used to monitor temperature (i.e., radiosondes, satellites, and reanalysis models). In a few instances, however, differences in the version of the MSU data (for channel 2 and 4) and the HadRT radiosonde data (for the 2_{LT} retrieval) were also found to be important.

In summary, it is difficult to obtain reliable estimates of short-timescale trends embedded in noisy time series and assess their statistical significance. Trend uncertainties arising from the choice of linear fitting method can be large. The high noise levels and strong temporal autocorrelation of the deep-layer temperature data used here lead to broad confidence bands about the trend estimates. Because of this, for virtually all data sets considered here, one cannot conclude that the observed trends differ from zero nor that they differ from model estimates of what these trends should be in response to anthropogenic perturbations. Claims that we know the observed global-mean lower-tropospheric temperature trend over the satellite era to within a few hundredths of a degree C/decade should therefore be treated with caution.

Acknowledgments. Work at Lawrence Livermore National Laboratory was performed under the auspices of the U.S. Department of Energy, Environmental Sciences Division, under contract W-7405-ENG-48. Tom Wigley received support from the NOAA Office of Global Programs ("Climate Change Data and Detection") under grant NA87GP0105. The MSU data and static MSU weighting functions were kindly provided by John Christy (University of Alabama in Huntsville). The HadRT1.1 and HadRT1.2 data sets were developed under United Kingdom Department of the Environment, Regions, and Transport Contract PECD/7/12/37 and Public Meteorological Service Contract 2/97. Jim Angell (NOAA) provided an updated version of the ANGELL radiosonde data. Useful comments and suggestions supplied by Kevin Trenberth, Jim Hurrell (NCAR), Dick Jones (Geophysical Statistics Project, NCAR, and University of Colorado Health Sciences Center), Jim Angell, John Christy, Francis Zwiers (Canadian Centre for Climate Modeling and Analysis), Richard Smith (North Carolina State University), and Greg Markowski (Texas A&M University) substantially improved the manuscript.

References

Angell, J. K., Variations and trends in tropospheric and stratospheric global temperatures, 1958–87, *J. Clim.*, *1*, 1296–1313, 1988.

Angell, J. K., Variation with height and latitude of radiosonde temperature trends in North America, 1975–94, *J. Clim.*, *12*, 2551–2561, 1999.

Balling, R. C., P. J. Michaels, and P. C. Knappenberger, Analysis of winter and summer warming rates in gridded temperature time series, *Clim. Res.*, *9*, 175–181, 1998.

Bartlett, M. S., Some aspects of the time-correlation problem in regard to tests of significance, *J. R. Stat. Soc.*, *98*, 536–543, 1935.

Bengtsson, L., E. Roeckner, and M. Stendel, Why is the global warming proceeding much slower than expected?, *J. Geophys. Res.*, *104*, 3865–3876, 1999.

Berntsen, T. K., I. S. A. Isaksen, G. Myhre, G. S. Fuglestvedt, F. Stordal, T. Alsvik Larsen, R. S. Freckleton, and K. P. Shine, Effects of anthropogenic emissions on tropospheric ozone and its radiative forcing, *J. Geophys. Res.*, *102*, 28,101–28,126, 1997.

Bloomfield, P., and D. Nychka, Climate spectra and detecting climate change, *Clim. Change*, *21*, 275–287, 1992.

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and Ileana Bladé, The effective number of spatial degrees of freedom in a spatial field, *J. Clim.*, *12*, 1990–2009, 1999.

Chanin, M.-L., V. Ramaswamy, D. J. Gaffen, W. J. Randel, R. B. Rood, and M. Shiotani, Trends in stratospheric temperatures, in *Scientific Assessment of Ozone Depletion, 1998*, pp. 5.1–5.59, World Meteorol. Org., Geneva, 1999.

Christy, J. R., Temperature above the surface layer, *Clim. Change*, *31*, 455–474, 1995.

Christy, J. R., R. W. Spencer, and W. D. Braswell, How accurate are satellite 'thermometers'? *Nature*, *389*, 342–343, 1997.

Christy, J. R., R. W. Spencer, and E. S. Lobl, Analysis of the merging procedure for the MSU daily temperature time series, *J. Clim.*, *11*, 2016–2041, 1998.

Christy, J. R., R. W. Spencer, and W. D. Braswell, Global temperature variations since 1979, paper presented at 10th Symposium on Global Change Studies, Am. Meteorol. Soc., Dallas, Texas, January 10–15, 1999.

Corti, S., F. Molteni, and T. N. Palmer, Signature of recent climate change in frequencies of natural atmospheric circulation regimes, *Nature*, *398*, 799–802, 1999.

Dixon, W. J., and F. J. Massey, *Introduction to Statistical Analysis*, 678 pp., McGraw-Hill, New York, 1983.

Ebisuzaki, W., A method to estimate the statistical significance of a correlation when the data are serially correlated, *J. Clim.*, *10*, 2147–2153, 1997.

Elliott, W. P., D. J. Gaffen, J. D. Kahl, and J. K. Angell, The effect of moisture on layer thicknesses used to monitor global temperatures, *J. Clim.*, *7*, 304–308, 1994.

Gaffen, D. J., Temporal inhomogeneities in radiosonde temperature records, *J. Geophys. Res.*, *99*, 3667–3676, 1994.

Gaffen, D. J., M. A. Sargent, R. E. Habermann and J. R. Lanzante, 1999: Sensitivity of tropospheric and stratospheric temperature trends to radiosonde data quality. *J. Clim.*, in press, 2000.

Gibson, J. K., P. Kållberg, S. Uppala, A. Hernandez, A. Nomura, and E. Serrano, ERA Description, *ECMWF Re-Analysis Project Rep. Ser.* 1, 66 pp., Eur. Cent. for Medium-Range Weather Forecasts, Reading, England, 1997.

Hansen, J. E., H. Wilson, M. Sato, R. Ruedy, K. P. Shah, and E. Hansen, Satellite and surface temperature data at odds?, *Clim. Change*, *30*, 103–117, 1995.

Hansen, J. E., et al., Forcing and chaos in interannual to decadal climate change, *J. Geophys. Res.*, *102*, 25,679–25,720, 1997.

Hansen, J. E., M. Sato, R. Ruedy, A. Lacis, and J. Glascoe, Global climate data and models: A reconciliation, *Science*, *281*, 930–932, 1998.

Hasselmann, K., Linear and nonlinear signatures, *Nature*, *398*, 755–756, 1999.

Jones, P. D., T. J. Osborn, and K. R. Briffa, Estimating sampling errors in large-scale temperature averages, *J. Clim.*, *10*, 2548–2568, 1997.

Kalnay, E., et al., The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471, 1996.

Karl, T. R., R. R. Heim Jr., and R. G. Quayle, The greenhouse effect in central North America: If not now, when?, *Science*, *251*, 1058–1061, 1991.

Karoly, D. J., J. A. Cohen, G. A. Meehl, J. F. B. Mitchell, A. H. Oort, R. J. Stouffer, and R. T. Wetherald, An example of fingerprint detection of greenhouse climate change, *Clim. Dyn.*, *10*, 97–105, 1994.

Lanzante, J. R., Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including appli-

cations to historical radiosonde station data, *Int. J. Climatol*., *16*, 1197–1226, 1996.

Mitchell, J. M., Jr., B. Dzerdzeevskii, H. Flohn, W. L. Hofmeyr, H. H. Lamb, K. N. Rao, and C. C. Wallén, *Climatic Change*, *Techn. Note 79*, 79 pp., World Meteorol. Org., Geneva, 1966.

Nicholls, N., G. V. Gruza, J. Jouzel, T. R. Karl, L. A. Ogallo, and D. E. Parker, Observed climate variability and change, in *Climate Change 1995: The Science of Climate Change*, edited by J. T. Houghton et al., pp. 133–192, Cambridge Univ. Press, New York, 1996.

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

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 963 pp., Cambridge Univ. Press, New York, 1992.

Ramaswamy, V., M. D. Schwarzkopf, and W. J. Randel, Fingerprint of ozone depletion in the spatial and temporal pattern of recent lower-stratospheric cooling, *Nature*, *382*, 616–618, 1996.

Santer, B. D., et al., A search for human influences on the thermal structure of the atmosphere, *Nature*, *382*, 39–46, 1996a.

Santer, B. D., T. M. L. Wigley, T. P. Barnett, and E. Anyamba, 1996b: Detection of climate change, and attribution of causes, in *Climate Change 1995: The Science of Climate Change*, edited by J. T. Houghton et al., pp. 409–443, Cambridge Univ. Press, New York, 1996b.

Santer, B. D., J. J. Hnilo, T. M. L. Wigley, J. S. Boyle, C. Doutriaux, M. Fiorino, D. E. Parker, and K. E. Taylor, Uncertainties in observationally based estimates of temperature change in the free atmosphere, *J. Geophys. Res*., *104*, 6305–6333, 1999.

Spencer, R. W., and J. R. Christy, Precision and radiosonde validation of satellite gridpoint temperature anomalies, Part I, MSU channel 2, *J. Clim*., *5*, 847–857, 1992a.

Spencer, R. W., and J. R. Christy, Precision and radiosonde validation of satellite gridpoint temperature anomalies, Part II, A tropospheric retrieval and trends during 1979–90, *J. Clim*., *5*, 858–866, 1992b.

Tett, S. F. B., J. F. B. Mitchell, D. E. Parker and M. R. Allen, Human influence on the atmospheric vertical temperature structure: Detection and observations, *Science*, *274*, 1170–1173, 1996.

Timmermann, A., J. M. Oberhuber, A. Bacher, M. Esch, M. Latif, and E. Roeckner, Increased El Nino frequency in a climate model forced by future greenhouse warming, *Nature*, *398*, 694–696, 1999.

Trenberth, K. E., Atmospheric circulation climate changes, *Clim. Change*, *31*, 427–453, 1995.

Trenberth, K. E., and T. J. Hoar, The 1990–1995 El Niño Southern Oscillation event: Longest on record, *Geophys. Res. Lett*., *23*, 57–60, 1996.

Trenberth, K. E., and J. G. Olson, Representativeness of a 63-station network for depicting climate changes, in *Greenhouse-Gas-Induced Climatic Change: A Critical Appraisal of Simulations and Observations*, edited by M. E. Schlesinger, pp. 249–259, Elsevier Sci., New York, 1991.

Wentz, F. J., and M. Schabel, Effects of orbital decay on satellite-derived lower-tropospheric temperature trends, *Nature*, *394*, 661–664, 1998.

Wigley, T. M. L., and P. D. Jones, Detecting CO_2-induced climate change, *Nature*, *292*, 205–208, 1981.

Wilks, D. S., *Statistical Methods in the Atmospheric Sciences*, 467 pp., Academic, San Diego, Calif., 1995.

Woodward, W. A., and H. L. Gray, Global warming and the problem of testing for trend in time series data, *J. Clim*., *6*, 953–962, 1993.

Woodward, W. A., and H. L. Gray, Selecting a model for detecting the presence of a trend, *J. Clim*., *8*, 1929–1937, 1995.

Zwiers, F. W., and H. von Storch, Taking serial correlation into account in tests of the mean, *J. Clim*., *8*, 336–351, 1995.

J. S. Boyle, J. J. Hnilo, B. D. Santer, and K. E. Taylor, Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, P. O. Box 808, Mail Stop L-264, Livermore, CA 94550. (bsanter@pcmdi.llnl.gov)

D. J. Gaffen, National Oceanic and Atmospheric Administration, Air Resources Laboratory, Silver Spring, MD 20910.

D. Nychka and T. M. L. Wigley, National Center for Atmospheric Research, Boulder, CO 80303.

D. E. Parker, Hadley Centre for Climate Prediction and Research, United Kingdom Meteorological Office, Bracknell, UK.

(Received April 15, 1999; revised September 29, 1999;
accepted October 19, 1999.)