Separating signal and noise in atmospheric temperature changes: The importance of timescale

B. D. Santer,[1] C. Mears,[2] C. Doutriaux,[1] P. Caldwell,[1] P. J. Gleckler,[1] T. M. L. Wigley,[3] S. Solomon,[4] N. P. Gillett,[5] D. Ivanova,[1] T. R. Karl,[6] J. R. Lanzante,[7] G. A. Meehl,[3] P. A. Stott,[8] K. E. Taylor,[1] P. W. Thorne,[6] M. F. Wehner,[9] and F. J. Wentz[2]

[1]   We compare global-scale changes in satellite estimates of the temperature of the lower troposphere (TLT) with model simulations of forced and unforced TLT changes. While previous work has focused on a single period of record, we select analysis timescales ranging from 10 to 32 years, and then compare all possible observed TLT trends on each timescale with corresponding multi-model distributions of forced and unforced trends. We use observed estimates of the signal component of TLT changes and model estimates of climate noise to calculate timescale-dependent signal-to-noise ratios (S/N). These ratios are small (less than 1) on the 10-year timescale, increasing to more than 3.9 for 32-year trends. This large change in S/N is primarily due to a decrease in the amplitude of internally generated variability with increasing trend length. Because of the pronounced effect of interannual noise on decadal trends, a multi-model ensemble of anthropogenically-forced simulations displays many 10-year periods with little warming. A single decade of observational TLT data is therefore inadequate for identifying a slowly evolving anthropogenic warming signal. Our results show that temperature records of at least 17 years in length are required for identifying human effects on global-mean tropospheric temperature.

1. Introduction

[2]   Since the late 1970s, it has been recognized that the identification of human effects on climate is inherently a signal-to-noise (S/N) problem [*Hasselmann*, 1979; *Madden and Ramanathan*, 1980; *Wigley and Jones*, 1981; *Wigley and Raper*, 1990; *Allen et al.*, 1994; *Santer et al.*, 1994, 1995]. The warming signal arising from slow, human-caused changes in atmospheric concentrations of greenhouse gases is embedded in the background 'noise' of natural climate variability. Yet much of the recent public discourse on the reality of a discernible human influence on global climate reflects pervasive confusion regarding the distinctions between short-term climate variability and long-term climate change. It is therefore important and timely to illuminate some basic issues related to the S/N behavior of atmospheric temperature data.

[3]   Signal identification requires the application of standard signal processing methods [*Hasselmann*, 1979], which exploit differences in the spatial pattern and/or the temporal evolution of signal and noise [e.g., *Hasselmann*, 1979, 1993; *Santer et al.*, 1994, 1995, 1996; *North et al.*, 1995; *Tett et al.*, 1996; *Hegerl et al.*, 1996, 2007; *Allen and Tett*, 1999; *Stott et al.*, 2000; *Barnett et al.*, 2001, 2005; *Gillett et al.*, 2002]. Most attempts to identify anthropogenic "fingerprints" in observational climate data rely on multi-decadal records. Shorter periods of record generally have small S/N ratios, making it difficult to identify an anthropogenic signal with high statistical confidence [*Allen et al.*, 1994; *Santer et al.*, 1995, 1996; *North et al.*, 1995; *Hegerl et al.*, 1996]. The timescale dependence of S/N ratios arises primarily because the climate noise in most meteorological and oceanographic time series is largest on short (daily to annual) timescales, and becomes smaller over longer averaging periods [*Santer et al.*, 1995, 1996; *Liebmann et al.*, 2010]. Fingerprint results show a similar dependence on the spatial scale of the analysis, with

[1]Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, California, USA.
[2]Remote Sensing Systems, Santa Rosa, California, USA.
[3]National Center for Atmospheric Research, Boulder, Colorado, USA.
[4]Department of Atmospheric and Oceanic Sciences, University of Colorado at Boulder, Boulder, Colorado, USA.
[5]Canadian Centre for Climate Modelling and Analysis, Environment Canada, Victoria, British Columbia, Canada.
[6]National Climatic Data Center, National Oceanic and Atmospheric Administration, Asheville, North Carolina, USA.
[7]Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Princeton, New Jersey, USA.
[8]Met Office Hadley Centre, Exeter, UK.
[9]Lawrence Berkeley National Laboratory, Berkeley, California, USA.

higher S/N ratios for global- to hemispheric-scale climate changes [*Stott and Tett*, 1998].

[4] Because interannual noise is large, it can have a pronounced impact on decadal trends, both in observations and climate models. Models are therefore capable of simulating decade-long periods with little or no surface warming, even under future anthropogenic forcing regimes. This has been shown in two recent investigations [*Easterling and Wehner*, 2009; *Knight et al.*, 2009]. These studies have not explicitly analyzed S/N ratios as a function of timescale, which is our primary concern here.

[5] The current investigation focuses on globally averaged changes in the temperature of the lower troposphere (TLT) rather than on detailed spatial (or spatio-temporal) fingerprints of surface or atmospheric temperature change. Previous fingerprint work with TLT has consistently identified an anthropogenic influence on satellite- and radiosonde-based estimates of tropospheric temperature change (for reviews of these studies, see *Santer et al.* [2006], *Hegerl et al.* [2007], *Hegerl and Zwiers* [2011], and *Thorne et al.* [2011a]). Our goal here is to use global-scale TLT changes for illustrating some basic features of the timescale-dependence of S/N ratios. Such information is difficult to obtain from "space-time" fingerprint studies, which combine spatial and temporal information into a single vector.

[6] We focus on TLT because the magnitude of observed changes in TLT and lapse rates – and the consistency of these changes with climate model results – has been the subject of considerable scientific debate [e.g., *National Research Council*, 2000; *Gaffen et al.*, 2000; *Karl et al.*, 2006; *Hegerl et al.*, 2007; *Thorne et al.*, 2007, 2011b; *Douglass et al.*, 2008; *Santer et al.*, 2008; *McKitrick et al.*, 2010; *Fu et al.*, 2011]. Most previous studies of the consistency between modeled and observed TLT trends have performed such comparisons over a single period of record only – often a period less than the total length of the observations. Typically, such work does not assess whether modeled and observed temperature trends are consistent on multiple timescales, and does not consider whether the selected segment of the observed record is representative of the longer-term statistical behavior of the time series (an exception is the work of *Thorne et al.* [2007]). Given recent interest in comparing modeled and observed temperature trends over both short periods of record (the last 10 to 15 years) and the full satellite era, we assess trend consistency over a range of timescales (from 10 to 32 years) rather than over a single period of record.

[7] We note that there is a related body of literature which seeks to determine the "detection time" – the time at which an anthropogenically-forced climate signal can be statistically identified relative to background noise [see, e.g., *Karl et al.*, 1991; *Allen et al.*, 1994; *Santer et al.*, 1994; *Weatherhead et al.*, 1998; *Henson et al.*, 2010]. Such work uses either statistical or physical models (or both) to estimate the structure and levels of the background noise against which an observed, model-predicted, or idealized climate-change signal must be detected. To date, "detection time" studies (which are inherently a form of S/N analysis) have not been performed with global-scale changes in TLT. Nor has such work typically utilized noise information from a large, multi-model archive (as we do here).

[8] The structure of this paper is as follows. Section 2 describes the observational and model TLT data sets used here. Section 3 compares simulated and observed trend distributions on different timescales, and then briefly outlines how we assess whether observed TLT trends are statistically unusual relative to TLT trends in model control runs and externally forced simulations. Section 4 analyzes S/N ratios as a function of timescale, and provides time-scale-dependent estimates of the probability that model unforced variability could plausibly explain observed TLT trends. In Section 5, we examine the consistency between TLT trends in observations and in simulations of externally forced climate change. The possibility of spurious inflation of our estimated S/N ratios by model variability errors is addressed in Section 6. A short summary is given in Section 7. Appendix A supplies detailed information on statistical notation and on the calculation of *p*-values and S/N ratios.

2. Observational and Model Temperature Data

[9] We compare simulated and observed global-scale TLT trends using three different observational data sets, each based on measurements of microwave emissions made by Microwave Sounding Units (MSUs) on polar-orbiting satellites. The three MSU TLT data sets analyzed here were developed by research groups at the University of Alabama at Huntsville (UAH) [*Christy et al.*, 2007] and Remote Sensing Systems in Santa Rosa, California (RSS) [*Wentz and Schabel*, 1998; *Mears and Wentz*, 2005]. Differences between the temperature-change estimates generated by UAH and RSS arise from different choices made in the data set construction process, particularly in the treatment of inter-satellite biases, drifts in instrument calibration, and the effects of orbital drift [*Karl et al.*, 2006]. Two versions of the RSS TLT data (versions 3.2 and 3.3) were available, which differ only in terms of the amount of information they incorporate from Advanced Microwave Sounding Units (AMSUs), and therefore differ only after 1998. All observed MSU data sets span the period from 1979 through to the end of 2010.

[10] To obtain model estimates of forced and unforced TLT changes, we use output from phase 3 of the Coupled Model Intercomparison Project (CMIP-3) [*Meehl et al.*, 2007; *Intergovernmental Panel on Climate Change*, 2007], which was an important scientific resource for the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). We analyze three different types of simulation in the CMIP-3 multi-model archive: (1) pre-industrial control runs with no changes in either anthropogenic or natural external influences on climate. These simulations provide information on internal climate noise; (2) 20th century (20CEN) runs with estimated historical changes in human and (in some cases) natural external forcings. Examples of such forcings include anthropogenic changes in well-mixed greenhouse gases, stratospheric ozone, and sulfate aerosols, as well as natural changes in volcanic aerosols and solar irradiance; and (3) simulations with 21st century changes in greenhouse gases and anthropogenic aerosols prescribed according to the A1B scenario

of the IPCC Special Report on Emissions Scenarios [*Nakićenović and Swart*, 2000].

[11] Most of the 20CEN simulations end in 1999 or 2000. The A1B simulations were initiated from atmospheric and ocean conditions at the end of the 20CEN runs, and then integrated to at least 2099. The A1B runs provide a reasonable estimate of GHG forcings over 2000 to 2010, but it is not clear how realistically they represent the true net aerosol forcing over this period.

[12] Assessing the realism of the A1B aerosol forcing is hampered by substantial uncertainty in observed estimates of the net forcing caused by anthropogenic aerosols [*Forster et al.*, 2007; T. M. L. Wigley and B. D. Santer, Quantifying the anthropogenic component of 20th century warming, manuscript in preparation, 2011]. In addition to this large observational uncertainty in aerosol forcing, there are at least several other reasons why we might expect to see differences between the simulated and observed aerosol forcing over the last decade. First, many of the CMIP-3 20CEN and A1B simulations omit both the indirect effects of aerosols on clouds and the direct effects of black carbon aerosols (see auxiliary material).[1] Second, the A1B scenario may have underestimated the recent rapid increase in anthropogenic sulfate emissions from China, and hence may have underestimated the direct radiative forcing caused by sulfate aerosols [*Kaufmann et al.*, 2011].

[13] Splicing together output from the 20CEN and A1B simulations facilitates the comparison of simulated and observed atmospheric temperature changes over the full observed MSU record (1979 to 2010). While splicing increases the length of model record available for comparison with observations, it can also introduce forcing discontinuities [*Arblaster et al.*, 2011], which in turn impart inhomogeneities to the simulated TLT changes.

[14] The synthetic MSU temperatures analyzed here are computed from the CMIP-3 pre-industrial control runs and the 'spliced' 20CEN/A1B runs (see Figures 1 and 2, respectively). In most previous calculations of this type, the average temperature of the lower troposphere was obtained by applying a global-mean TLT weighting function to the atmospheric temperature profiles at each model grid-point [*Douglass et al.*, 2008; *Santer et al.*, 2008; *McKitrick et al.*, 2010]. Here, we use a new method developed at RSS (C. Mears et al., Fast, simple methods for calculating synthetic microwave sounder brightness temperatures from discrete-level data, submitted to *Journal of Atmospheric and Oceanic Technology*, 2011), in which local weighting functions were applied to model temperature profiles. The local weights depend on the grid-point surface pressure, temperature, and type (land or ocean). This new method provides more accurate estimates of synthetic MSU temperatures, particularly over high elevation regions.

[15] Most of our trend calculations rely on a standard least-squares estimator of linear changes in tropospheric temperature. We also explore the sensitivity of our model-versus-observed trend consistency results to use of an alternative trend estimator which is less sensitive to outliers (see Section 6). All observed and simulated trends are computed with monthly-mean near-global TLT data. For

simplicity, however, we refer to analysis timescales in years rather than months.

3. Accounting for Climate Noise in Comparisons of Modeled and Observed Trends

[16] Figure 3a shows why it is important to account for the effects of climate noise in comparisons of modeled and observed trends. On interannual timescales, one of the most prominent manifestations of climate noise is the El Niño/Southern Oscillation (ENSO). The relatively small values of overlapping 10-year TLT trends during the period 1998 to 2010 are partly due to the fact that this period is bracketed (by chance) by a large El Niño (warm) event in 1997/98, and by several smaller La Niña (cool) events at the end of the MSU record (Figure 7b).

[17] To assess the effect of seasonal and interannual variability on 10-year TLT trends, we fit linear trends to overlapping 120-month segments of the 384-month MSU TLT record. For maximally overlapping 120-month intervals (i.e., for overlap by all but one month), this yields 265 samples of 120-month trends. We use maximally overlapping trends to guard against excluding the largest cooling or warming trends from our analysis (see Appendix A).

[18] Ten-year TLT trends in the RSS v3.3 data set range from −0.05 to +0.44°C/decade (Figure 3a). The impact of interannual variability on trends is markedly reduced by trend fitting over 20-year periods. All 145 of the 20-year RSS v3.3 TLT trends are positive, and the range of 20-year trends (0.15 to 0.25°C/decade) decreases by more than a factor of four relative to the 10-year trend case (Figure 3b). Comparable results are obtained for RSS v3.2 and UAH TLT data (not shown).

[19] Similar noise reduction also occurs in model TLT data (Figures 4a–4c). On 10-year timescales, distributions of unforced and forced TLT trends overlap markedly (Figure 4a). This overlap occurs because even under anthropogenic and natural external forcing, interannual climate noise has a large influence on short, decadal trends. When trends are computed over 20-year periods, there is a reduction in the amplitude of both the control run noise and the noise superimposed on the externally forced TLT signal in the 20CEN/A1B runs. Because of this noise reduction, the signal component of TLT trends becomes clearer, and the distributions of unforced and forced trends begin to separate (Figure 4b). Separation is virtually complete for 30-year trends (Figure 4c).

[20] Figure 5 shows how we account for the effects of interannual noise in assessing the consistency between observed and simulated trends. As described above, we first calculate maximally overlapping observed trends of length L months. For the $L = 120$ case, there is nothing statistically unusual about the small temperature trends with starting dates in the post-1998 period (Figure 5a). Similar periods with near-zero or negative decadal trends occurred at the start of the satellite record and for starting dates around 1987. The average observed 10-year TLT trend over 1979 to 2010 is positive in all cases (0.17, 0.15, and 0.15°C/decade for RSS v3.2, RSS v3.3, and UAH, respectively), with large sampling uncertainty.

[21] For each of the 265 observed 10-year trends, we can use control TLT data to calculate the probability $p_c(i)'$ that

Unforced TLT Variability in CMIP-3 Pre-Industrial Control Runs
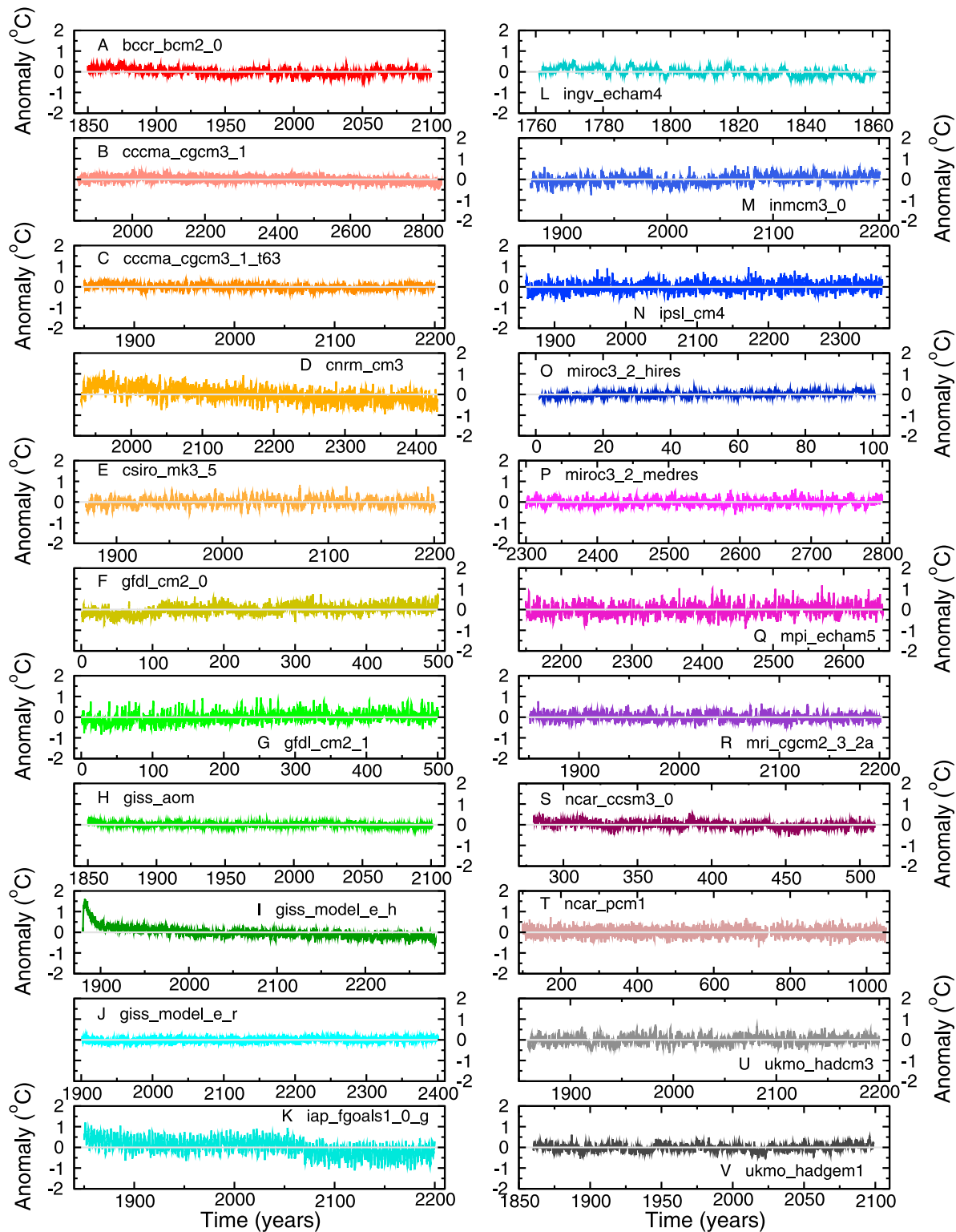
Near-global averages (70°S-82.5°N)

Figure 1. Time series of monthly mean, near-global anomalies in lower tropospheric temperature (TLT) in the pre-industrial control runs performed with 22 different CMIP-3 models. Model anomalies are averaged over 82.5°N–70°S (the latitudinal extent of RSS TLT data), and are defined with respect to climatological monthly means over the entire control run. For information on the processing of control run data, refer to Table S3 in the auxiliary material.

TLT Changes in CMIP-3 Extended 20th Century Runs

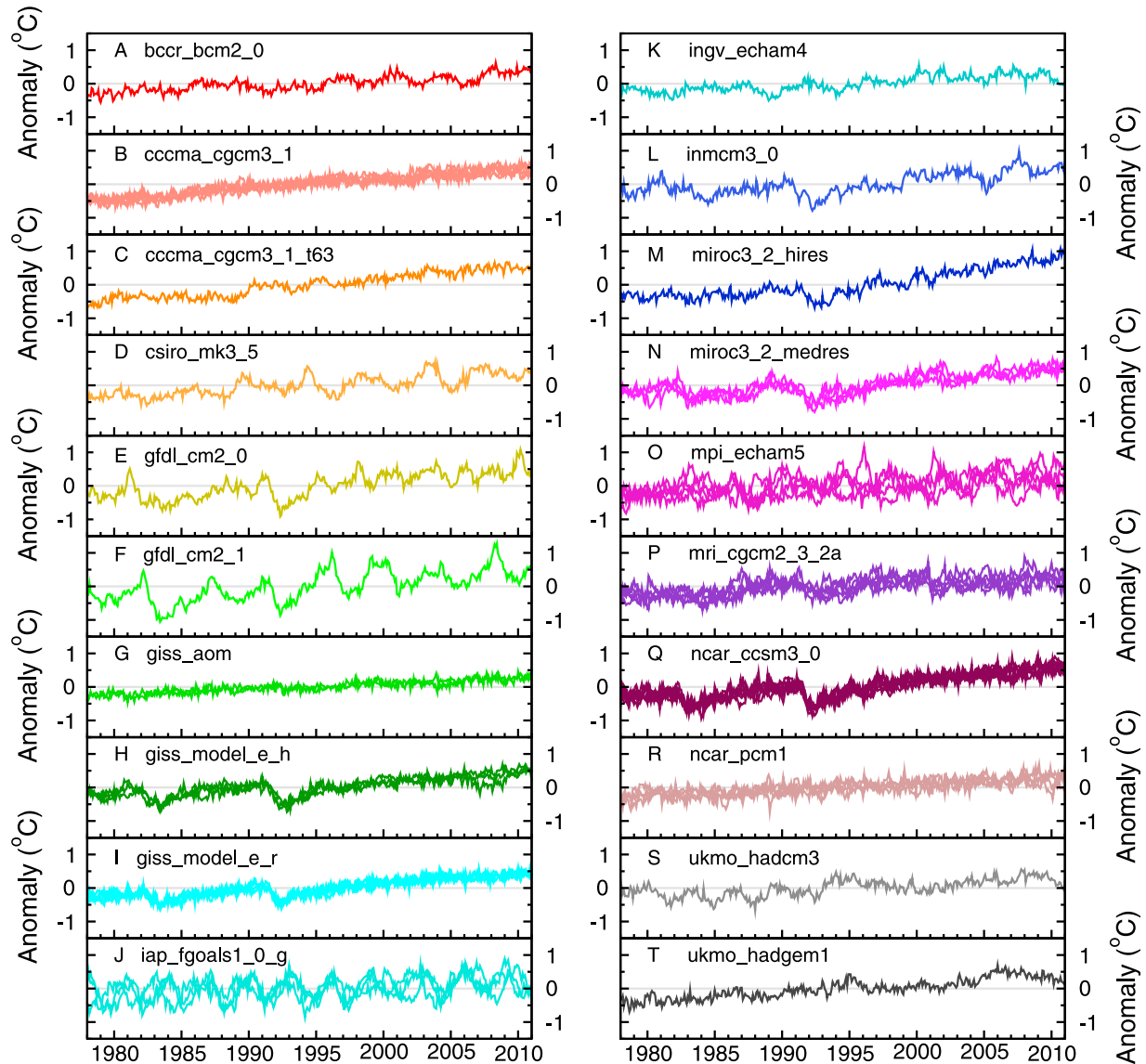## Near-global averages (70°S-82.5°N). Extension with SRES A1B runs

Figure 2. Time series of monthly mean, near-global TLT anomalies in the spliced 20CEN/A1B runs performed with 20 different CMIP-3 models. Spatial averaging and anomaly definition is as in Figure 1. For detailed information regarding splicing of 20CEN and A1B runs, see Table S2 in the auxiliary material. The y-axis range is identical in each panel.

an internally generated trend could exceed the current observed TLT trend. In our notation (see Appendix A), the subscripts o, c, and f denote use of data from observations, control integrations, and the forced 20CEN/A1B runs, respectively. The index i is over the number of overlapping L-month trends in the observations. The prime notation denotes 'weighted' p-values, which account for inter-model differences in either the length of control integrations or in the number of realizations of forced climate change. We use these weighted p-values solely to evaluate whether observed trends are unusual relative to model distributions of

forced and unforced trends (rather than as a basis for conducting formal statistical significance tests).

[22] Interannual noise has a large effect on $p_c(i)'$ (Figure 5b). As expected, the smallest $p_c(i)'$ values are obtained for the largest observed 10-year TLT trends, which end at the time of the 1997/1998 El Niño. For RSS v3.2, v3.3, and UAH TLT data, the mean $p_c(i)'$ value of the 265 individual 10-year trends, $\overline{p_c}'$, is 0.20, 0.24, and 0.22 (respectively). On average, therefore, roughly 20% to 24% of the unforced model trends are larger than the observed 10-year trends. Note that the average of individual $p_c(i)'$

Effect of Interannual Variability on Estimates of Observed TLT Trends
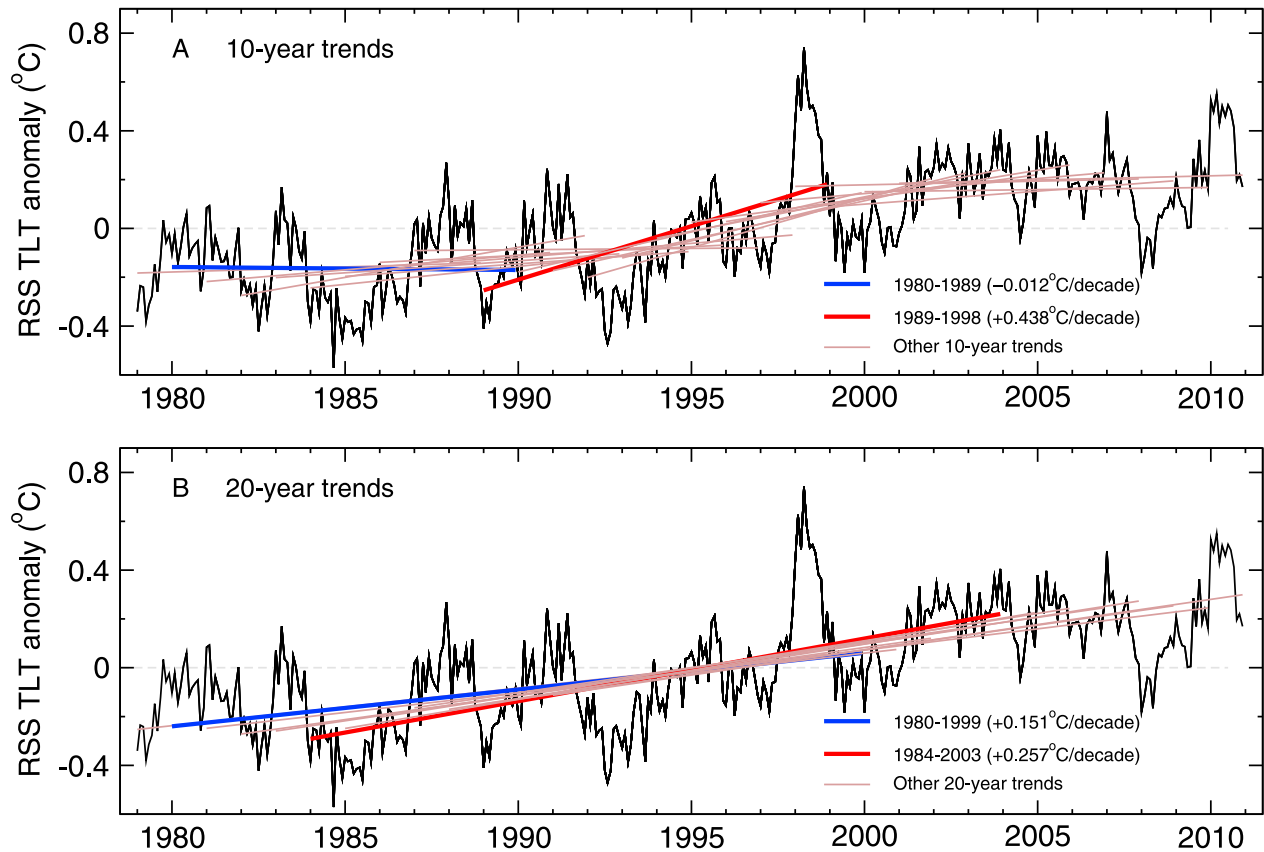### RSS near-global TLT trends (82.5°N-70°S)

Figure 3. Satellite-based estimates of monthly mean, near-global TLT anomalies. Observations are from RSS version 3.2 [*Mears and Wentz*, 2005]. Least squares linear trends were fitted to overlapping (a) 120- and (b) 240-month segments of the RSS time series. For visual display purposes, the trend overlap shown here is by all but 12 months (rather than the one-month overlap used for calculating p-values; see Appendix A). The largest negative and positive trends are highlighted in blue and red. Spatial averaging and anomaly definition is as in Figure 2.

values is very similar to the p_c value calculated using the average of individual observed trends.

[23] Our results show that observed 10-year TLT trends are not statistically unusual relative to unforced control run trends. This does not mean that there is no underlying externally forced signal in the observed data – merely that the noise level on this timescale is large, which precludes reliable identification of such a signal.

[24] Values of $p_f(i)'$ (Figure 5c) are defined in an analogous way to $p_c(i)'$, and provide information on whether observed TLT trends are unusually small relative to the multi-model distribution of forced TLT trends in the spliced 20CEN/A1B runs (see Appendix A). These simulations provide 51 realizations of forced TLT changes over 1979 to 2010, generated with 20 different climate models. For maximally overlapping 120-month trends, there are 13515 (265×51) samples of forced trends. On the 10-year timescale, $\overline{p_f}'$ values are 0.39, 0.37, and 0.37 for RSS v3.2, v3.3, and UAH (respectively). On average, therefore, between 37% and 39% of the 10-year TLT trends in the forced

20CEN/A1B runs are actually smaller than the observed 10-year trends.

[25] Even in the post-1998 period with muted observed warming, almost all individual $p_f(i)'$ values exceed 0.10 (Figure 5c). This indicates that more than 10% of the model trends are smaller than the post-1998 10-year trends in the RSS and UAH data sets. We conclude from these results that there is no statistical inconsistency between the recent small observed 10-year TLT trends and the trends obtained from the externally forced model simulations.

4. Signal-to-Noise Ratios

[26] We now examine the relative sizes of signal and noise estimates on different timescales. Consider the signal estimates first. In both the observations and the 20CEN/A1B runs, individual TLT time series consist of two components: an externally forced signal, and internally generated noise [*Santer et al.*, 2008]. The underlying signal component may have complex time evolution, but is approximated here by a
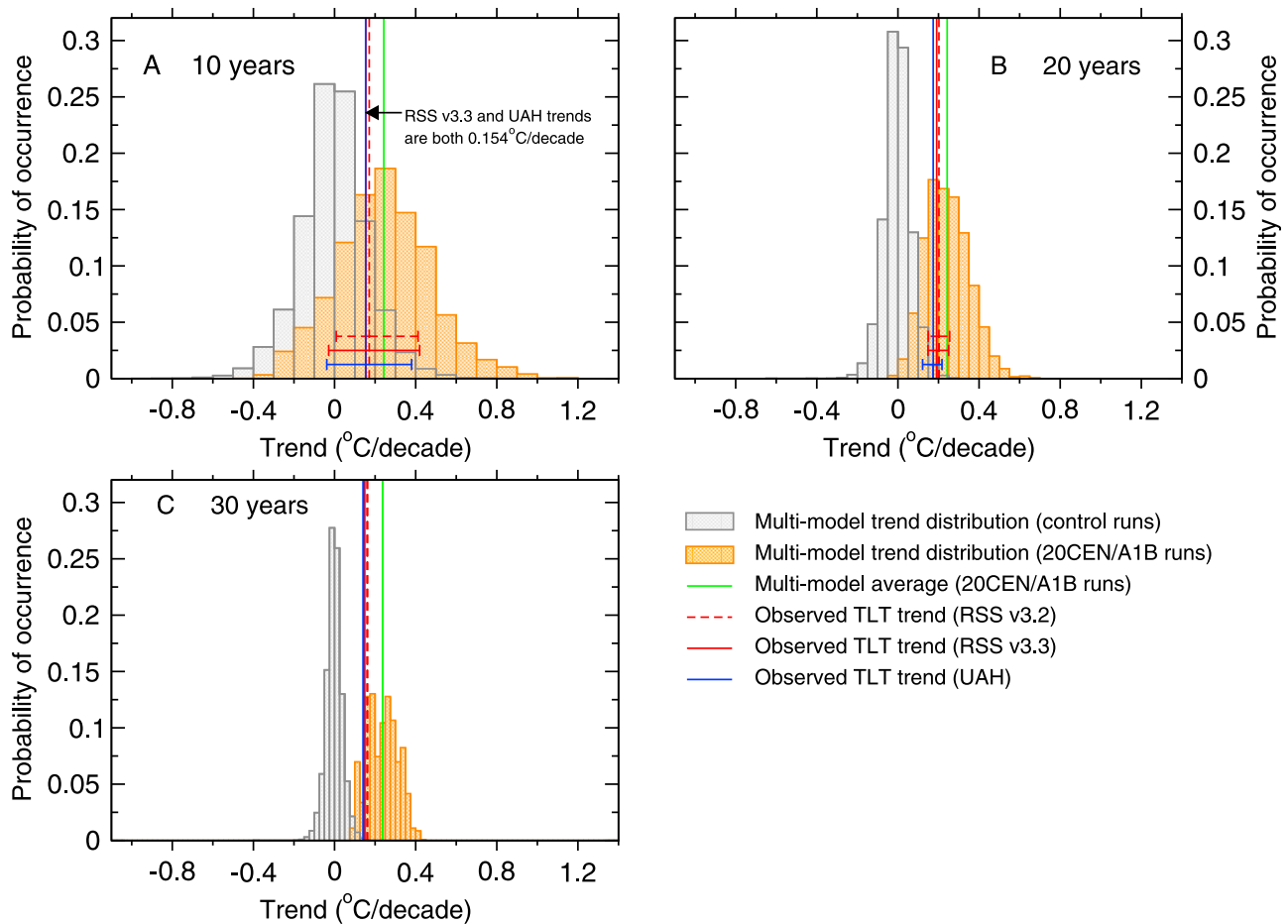
Figure 4. Sampling distributions of forced and unforced temperature trends on (a) 10-, (b) 20-, and (c) 30-year timescales. Results are for near-global changes in TLT. Model output is from the CMIP-3 control runs and spliced 20CEN/A1B runs [*Meehl et al.*, 2007]. Some models have multiple control runs and multiple realizations of forced climate change (see auxiliary material). All available model output was used in generating trend sampling distributions. Least squares linear trends of length 120, 240 and 360 months were fitted to maximally overlapping segments of each model's control- and perturbed-run synthetic TLT data (see Appendix A). Trends from individual models were then pooled to form multimodel distributions. The analysis period for the 20CEN/A1B runs (1979 to 2010) is identical to that used for the observations. Observed trend results are the averages of overlapping 120-, 240- and 360-month trends fitted to the RSS v3.2, v3.3, and UAH TLT data. The 5-95 percentiles of the observed trend distributions are shown for 120- and 240-month trends only. Here and elsewhere, multimodel average trends were calculated using the ensemble-mean trends of individual models (if multiple 20CEN/A1B realizations were available).

linear trend. To obtain estimates of the signal, we calculate averages of maximally overlapping trends, thereby reducing the impact of seasonal and interannual climate noise.

[27] We emphasize that such trend averages (which we refer to subsequently as 'signals') are only estimates of the true (but unknown) signal. Since the observations have far fewer trend samples than the multi-model ensemble of 20CEN/A1B simulations, we expect observational estimates of the underlying signal to be more contaminated by internally generated variability. In addition to the purely statistical uncertainty inherent in estimating trends from noisy data, model signal estimates incorporate uncertainties arising from inter-model differences in: (1) the applied human and natural external forcings; (2) the responses to these forcings; and (3) the size and spatio-temporal structure of internal climate variability superimposed on the signal [*Karl et al.*, 2006].

[28] In both models and observations, the signal is always positive (≥0.14°C/decade) over the timescale range examined here (Figure 6a). To first order, the signal is timescale-invariant. Average forced trends in the model 20CEN/A1B simulations vary by less than 10% (from 0.24 to 0.26°C/
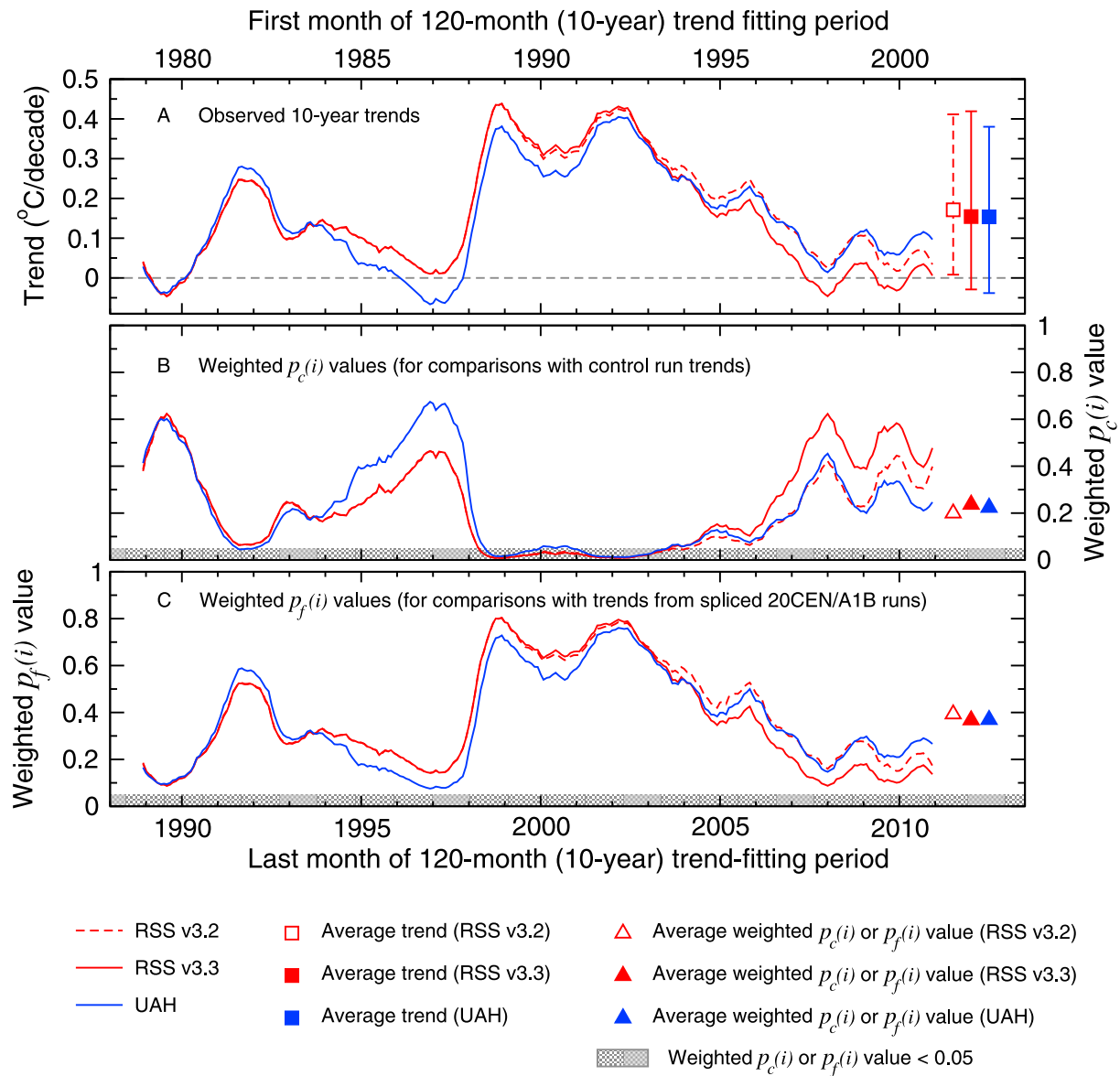
Figure 5. (a) Comparison of modeled and observed trends on the 10-year timescale. All results are for modeled and observed 120-month trends in near-global TLT. For each observed data set (RSS v3.2, v3.3, and UAH), there are 265 overlapping 120-month trends for the period 1979 to 2010. The average of these trend values, $\overline{b_o}$, is plotted on the right side of Figure 5a, together with the 5–95 percentiles of the observed trend distributions. Each observed trend was compared with multimodel sampling distributions of (b) unforced and (c) forced 120-month trends, allowing calculation of $p_c(i)'$ and $p_f(i)'$ values (respectively; see Appendix A). The averages of the 265 individual $p_c(i)'$ and $p_f(i)'$ values, $\overline{p_c}'$ and $\overline{p_f}'$, are indicated by squares in Figures 5b and 5c. The regions where $p_c(i)'$ and $p_f(i)'$ are less than 0.05 (i.e., where observed trends are statistically unusual relative to the multimodel sampling distributions of unforced and forced trends) are marked in grey. Note that two different sets of x-axis labels are provided, identifying both the start month (top axis) and the end month (bottom axis) of the trend-fitting period.

decade) over 10- to 32-year timescales. Average observational TLT trends show slightly greater percentage variation with timescale, and range from 0.16 to 0.21°C/decade (0.15 to 0.19°C/decade) for RSS v3.2 (v3.3) data and from 0.14 to 0.18°C/decade for UAH results.

[29] In contrast, the standard deviation of the multi-model sampling distribution of unforced trends decreases by more than a factor of four over the 10- to 32-year timescale range (Figure 6b). This is the primary reason for the increase in S/N with increasing trend fitting period (Figure 6c). For all three
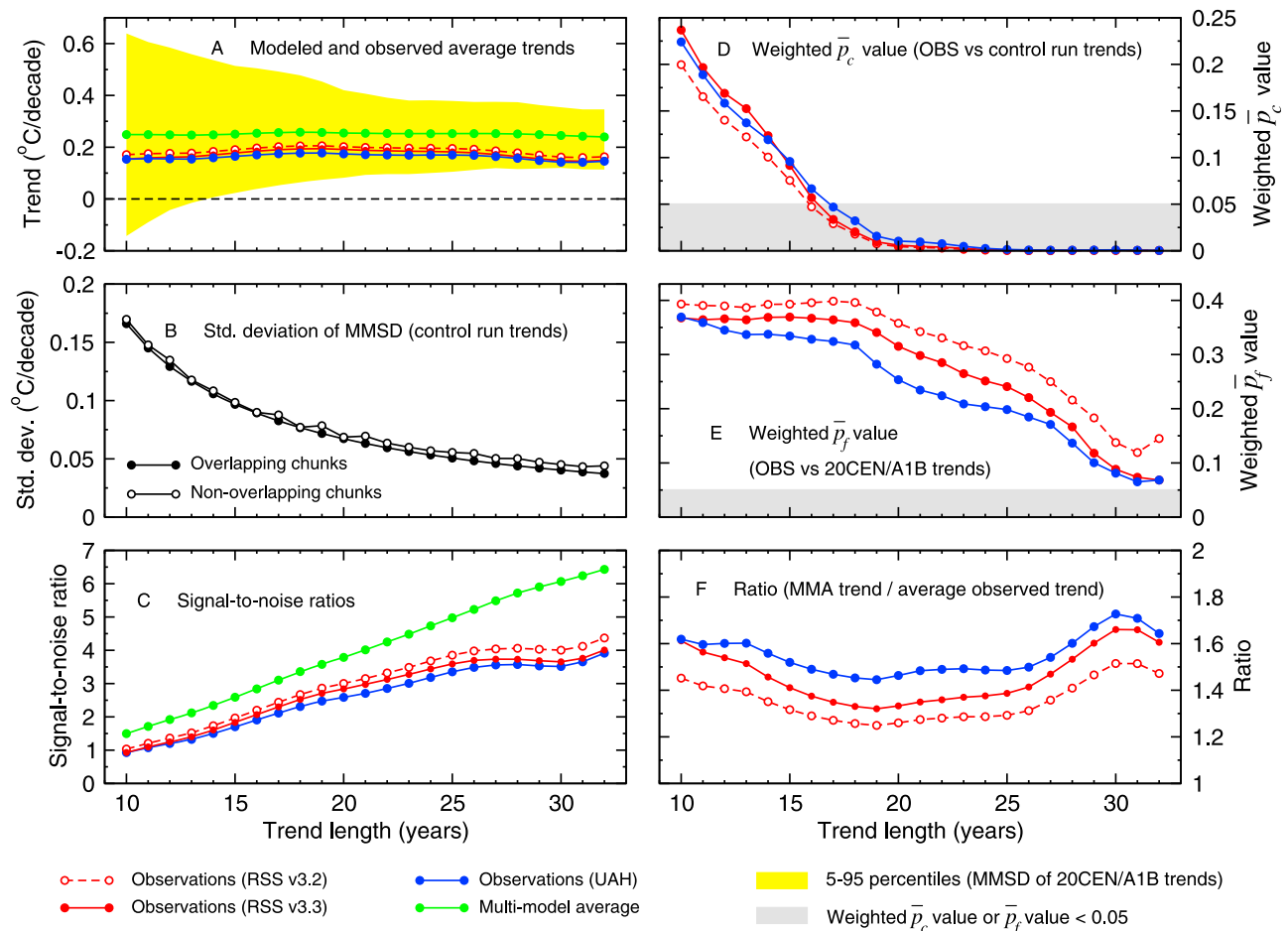
Figure 6. Behavior of signal-to-noise ratios (S/N) as a function of increasing trend length. All results are for modeled and observed trends in near-global TLT. The analysis is for trends on timescales ranging from 10 to 32 years. Results are for (a) the estimated signal component of trends; (b) noise trends; (c) S/N ratios; (d) values of $\overline{p_c}'$ and (e) $\overline{p_f}'$; and (f) the ratio between the multimodel average TLT trend and the average observed TLT trend. Observed trends in Figure 6a are values of $\overline{b_o}$, and represent the averages of RSS v3.2, v3.3, and UAH trend distributions (see Appendix A). Model results in Figure 6a are the multimodel average trends, calculated from sampling distributions of forced TLT trends obtained from the spliced 20CEN/A1B runs. Ensemble-mean results were used in estimating multimodel average trends. The 5–95 percentile range (based on the full multimodel trend sampling distribution) is shaded in yellow. Results in Figure 6b are the standard deviations of the multimodel sampling distributions of unforced trends, calculated using both overlapping and non-overlapping time series segments. S/N ratios in Figure 6c are simply the estimated signal trends in Figure 6a divided by the standard deviations of unforced trends in Figure 6b (computed with overlapping chunks). The $\overline{p_c}'$ ($\overline{p_f}'$) values in Figures 6d (6e) were computed with overlapping time series segments, and provide information on whether observed TLT trends are unusually large (small) with respect to multimodel sampling distributions of unforced (forced) trends. Values of $\overline{p_c}'$ and $\overline{p_f}' < 0.05$ are shaded in grey. For full details of statistical notation and all calculations, refer to Appendix A.

observational TLT data sets, S/N ratios increase from roughly 1.0 for 10-year trends to greater than 3.9 for 32-year trends. Because S/N ratios are very low on the 10-year timescale, one cannot use such short observational records to make meaningful inferences about the presence or absence of a slowly-evolving anthropogenic warming signal.

[30] On timescales longer than 17 years, the average trends in RSS and UAH near-global TLT data consistently exceed 95% of the unforced trends in the CMIP-3 control runs (Figure 6d), clearly indicating that the observed multi-decadal warming of the lower troposphere is too large to be explained by model estimates of natural internal variability.

This conclusion is dependent on the fidelity with which models simulate the amplitude of observed climate noise, particularly on multi-decadal timescales – an issue that we explore later in Section 6.

5. Consistency Between TLT Trends in Observations and 20CEN/A1B Runs

[31] We consider next the estimated $\overline{p_f}'$ values obtained for the comparison of modeled and observed TLT trends (Figure 6e). There is no timescale on which observed trends are statistically unusual (at the 5% level or better) relative to the multi-model sampling distribution of forced TLT trends. We conclude from this result that there is no inconsistency between observed near-global TLT trends (in the 10- to 32-year range examined here) and model estimates of the TLT response to anthropogenic forcing.

[32] The results in Figure 6e were obtained with a least-squares estimator for linear trends. It is interesting to consider whether the use of a trend estimator which is less sensitive to outliers [see, e.g., *Lanzante*, 1996; *Santer et al.*, 2000] leads to different conclusions regarding the consistency between TLT trends in observations and the model 20CEN/A1B runs.

[33] One such estimator minimizes the absolute deviations between the data and the linear fit [*Press et al.*, 1992]. We repeated all calculations of $\overline{p_f}'$ values using this 'least absolute deviations' (LAD) approach. On the longest and shortest timescales considered in our analysis, the least-squares and LAD trend estimators produce very similar results (see auxiliary material, Figure S1). For timescales ranging from roughly 19 to 30 years, the LAD estimator yields systematically higher values of $\overline{p_f}'$ – i.e., model forced trends are in closer agreement with observations. We conclude from this that our use of a least-squares estimator provides a conservative estimate of the consistency between TLT trends in observations and the 20CEN/A1B runs.

[34] Recently, it has been claimed that model-based estimates of global-scale TLT changes are a factor of three larger than the observed 'residual' TLT trend (J. R. Christy, Testimony in Hearing before the Subcommittee on Energy and Power, Committee on Energy and Commerce, House of Representatives, March 8, 2011, http://republicans.energy-commerce.house.gov/Media/file/Hearings/Energy/030811/Christy.pdf) (hereinafter Christy, online document, 2011). This residual trend was estimated after statistical removal of ENSO and volcanic signals from UAH TLT data, but not from model data. The net effect of removing ENSO and volcanic signals was to reduce the UAH TLT trend over 1979 to 2010 from 0.14 to 0.09°C/decade (Christy, online document, 2011).

[35] Our comparison of 'raw' modeled and observed trends (Figure 6f) does not involve removal of ENSO and volcanic effects from observations alone, and is not restricted to a single period of record. We find that for the range of TLT trends considered here, there is no trend length at which the multi-model average trend, $\overline{\overline{b_f}}$, is more than 1.73 times larger than $\overline{b_o}$, the average observed TLT trend (see Figure 6f). Across the 10- to 32-year range of trend lengths, the average ratio of $\overline{\overline{b_f}}/\overline{b_o}$ is 1.35 for RSS v3.2, 1.46 for RSS v3.3, and 1.55 for UAH.

[36] Possible reasons for the fact that $\overline{\overline{b_f}}$ is consistently larger than $\overline{b_o}$ will be discussed in detail in subsequent work, which will examine the latitude-altitude structure of differences between modeled and observed atmospheric temperature trends. Here, it is sufficient to note that many of the 20CEN/A1B simulations neglect negative forcings arising from stratospheric ozone depletion, volcanic dust, and indirect aerosol effects on clouds. Even CMIP-3 simulations which include these factors were performed roughly 7-10 years ago, and thus do not include solar irradiance changes over the last 11-year solar cycle [*Wigley*, 2010; *Kaufmann et al.*, 2011], decreases in stratospheric water vapor concentrations over 2000 to 2009 [*Solomon et al.*, 2010], and increases in volcanic aerosol loadings over the last decade [*Vernier et al.*, 2011; *Solomon et al.*, 2011]. It is likely that omission of these negative forcings contributes to the positive bias in the model average TLT trends in Figure 6f. Given the considerable technical challenges involved in adjusting satellite-based estimates of TLT changes for inhomogeneities [*Mears et al.*, 2006, 2011], a residual cool bias in the observations cannot be ruled out, and may also contribute to the offset between the model and observed average TLT trends.

[37] A notable feature of our results is the decrease in values of $\overline{p_f}'$ for trend lengths $L > 20$ years (Figure 6e). This is partly a consequence of the decrease in noise amplitude with increasing trend length, so that any errors in model signal trends are less obscured by noise on longer timescales. The timescale dependence of $\overline{p_f}'$ is also affected by forcing discontinuities at the splice point between the 20CEN and A1B runs [*Arblaster et al.*, 2011]. These discontinuities occur between 21 and 22 years after the 1979 start date used for comparisons with observational TLT data. They have the largest impact on 25- to 30-year trends spanning the splice point.

[38] Another factor contributing to the decrease in $\overline{p_f}'$ with increasing trend length is the occurrence of several La Niña events near the end of the satellite record (see Figure 7b). The cooling induced by this clustering of La Niñas appears to be statistically unusual with respect to average model estimates of interannual variability (Figure 7a), and has a larger influence on observed 30-year TLT trends than on 10-year trends. This is because each observational TLT data set has only 25 overlapping 30-year trends, all of which are affected by the La Niña-induced cooling at the end of the satellite record. For shorter 10-year trends, however, there are 265 overlapping samples, which diminishes the impact of the recent cluster of La Niñas. Note that the addition of observed TLT data for 2010 almost doubled the RSS and UAH $\overline{p_f}'$ values for 30-year trends – i.e., observed trends became less unusual relative to the multi-model sampling distribution of forced trends.

6. Comparison of Modeled and Observed TLT Variability

[39] If models systematically underestimate the amplitude of near-global TLT variability on multi-decadal timescales, the S/N ratios in Figure 6c are likely to be spuriously inflated. Whether such a systematic error exists is difficult to determine. Because satellite TLT records are relatively short, observations cannot provide a strong constraint on
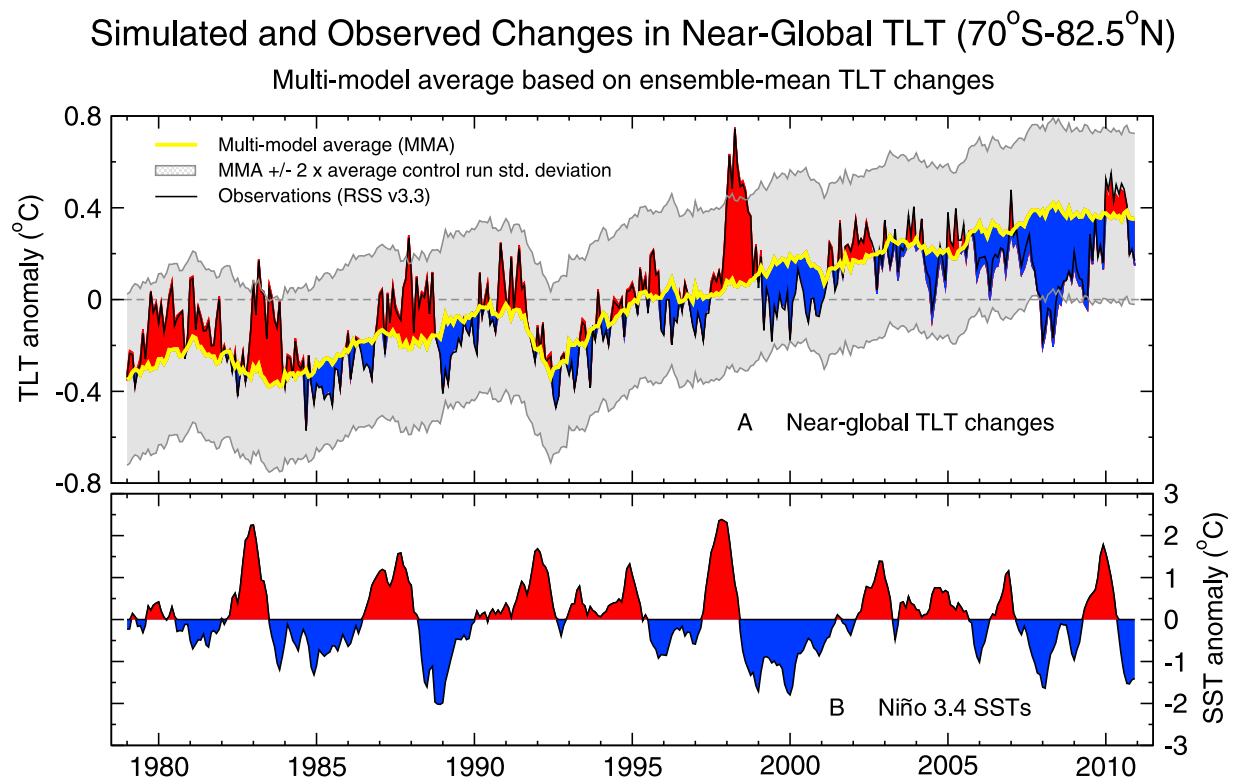
Simulated and Observed Changes in Near-Global TLT (70°S-82.5°N)
Multi-model average based on ensemble-mean TLT changes

Figure 7. Time series (a) of simulated and observed changes in monthly mean near-global TLT over 1979 to 2010 and (b) of observed changes in Niño 3.4 SSTs. Observational results in Figure 7a are from RSS version 3.3. Model results are multimodel average TLT changes computed from the spliced 20CEN/ A1B runs. All anomalies are defined as in Figure 3. RSS TLT anomalies are shaded red (blue) if they are larger (smaller) than the multimodel average. The grey shaded envelope is the multimodel average ± twice the standard deviation of the average control run interannual variability in 22 CMIP-3 control runs, cal-culated as described in Appendix A. The recent cooling near the end of observed TLT records is outside the 'model average' envelope of interannual variability. The observed Niño 3.4 SST data in Figure 7b are from NOAA ERSST version 3b [*Smith et al.*, 2008], and have a trend of −0.038°C over January 1979 to December 2010. Note the lag of roughly six months between the Niño 3.4 SST variability and global-scale TLT anomalies [*Wigley*, 2010].

model-based estimates of low-frequency TLT variability. Nor is it clear whether well-constrained observed estimates of high-frequency TLT variability provide useful insights into the direction and size of model variability errors on multi-decadal timescales.

[40] To investigate these issues, we applied high-pass and band-pass filters to the modeled and observed TLT data. The band-pass filtering focuses on TLT variability on timescales of 10 years, with half-power points at 5 and 20 years, while the high-pass filter has a half-power point at two years, and excludes all variability on timescales longer than 5 years (see auxiliary material, Figure S2). An example of the high-pass and band-pass filtered components of the RSS v3.3 TLT data is shown in Figure 8a. All filtering operations were performed after removal of least-squares linear trends, to avoid inflation of the variability estimate by the trend.

[41] Results shown in Figure 9 are the temporal standard deviations of the high- and band-pass filtered TLT data, calculated over the 384-month period from January 1979 to December 2010. Model information in Figure 9 is from the spliced 20CEN/A1B runs rather than the control simulations.

The former are the simulations most relevant for direct comparison with observations. This is because the TLT variability in observations and in many of the 20CEN/A1B runs (see auxiliary material) has both an internally generated component and a component related to volcanic and solar forcing. In contrast, TLT variability in the control runs is solely generated by processes internal to the climate system, and has no contribution from natural external forcing.

[42] There is no evidence from Figure 9 that models systematically underestimate the amplitude of observed TLT variability on timescales of 5–20 years. For the band-pass filtered data, the ratio R_{band} between the multi-model aver-age temporal standard deviation and the observed temporal standard deviation is roughly 1.2 for RSS v3.2 and RSS v3.3 and 1.1 for UAH – i.e., the model variability on 5–20-year timescales is actually 10–20% larger than observed. If the observational estimates are reliable, the 'observed' S/N ratios in Figure 6c are likely underestimates of the true S/N ratios on multi-decadal timescales.

[43] One interesting aspect of Figure 9 is that the CMIP-3 models have a relatively weak functional relationship between the amplitude of high- and low-frequency TLT

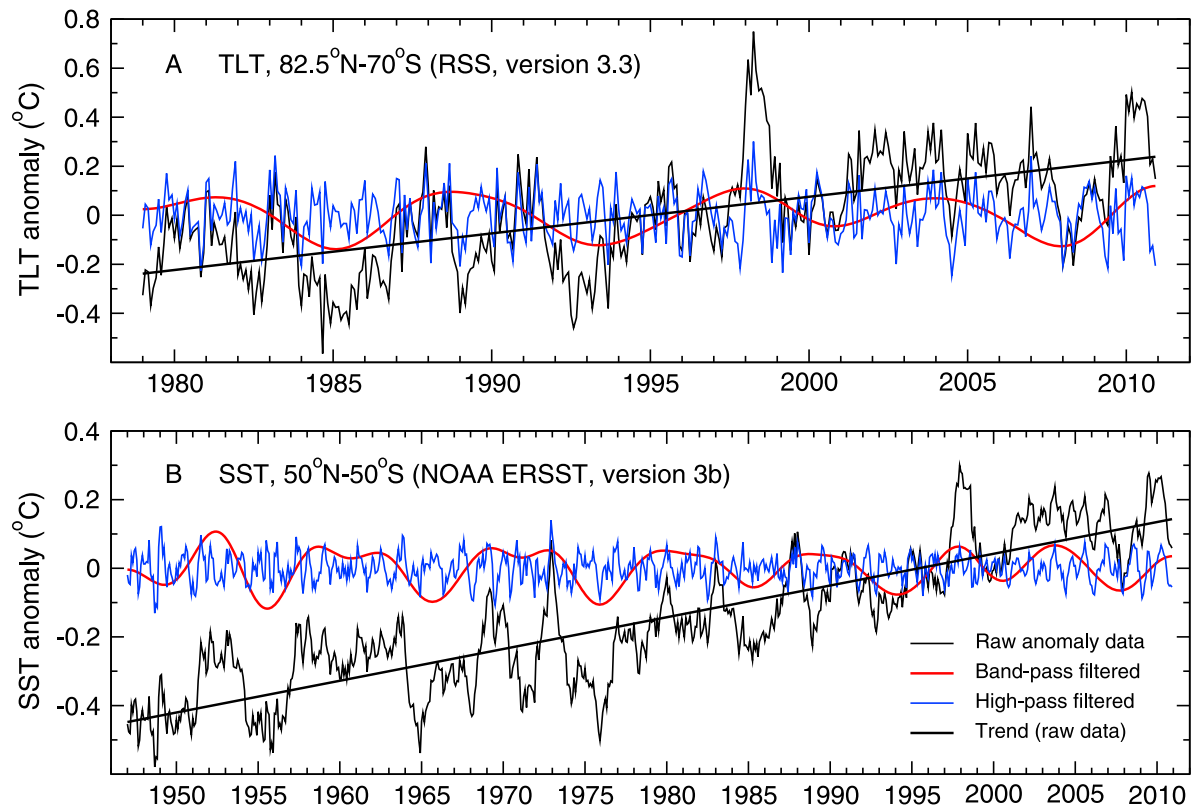## Band- and High-Pass Filtering of Global-Scale TLT and SST Anomalies

Figure 8. Observed time series of monthly mean anomalies of raw and filtered (a) TLT and (b) SST data. The analysis periods are 1979–2010 (TLT) and 1947–2010 (SST). TLT results are from version 3.3 of the RSS data set [*Mears et al.*, 2011], and were spatially averaged over 82.5°N–70°S. SST data are from version 3b of the NOAA ERSST data set [*Smith et al.*, 2008], and were averaged over a slightly smaller domain (50°N–50°S). After removal of a least squares linear trend from the raw anomalies, a Butterworth filter was used to perform band-pass and high-pass filtering. There is virtually no overlap between the frequencies isolated by the high- and band-pass filters. The least squares linear trends in TLT (0.149°C/decade) and SST (0.092°C/decade) are also shown. The correlation between the band-pass filtered TLT and SST time series in Figures 8a and 8b is 0.92 for the period of overlap.

variability. The correlation between the high- and band-pass filtered results is only 0.48, and the direction of the model-average variability bias differs in the high- and low-frequency variability cases. While R_{band} exceeds 1.0, the ratio R_{high} is 0.96 for all three observational data sets, indicating that the observed high-frequency variability is slightly larger than the multi-model average.

[44] These results suggest that model errors in well-observed interannual variability may not provide reliable information on the size and direction of model errors in low-frequency variability. This reflects the fact that different modes of variability have different characteristic timescales. Model performance in simulating ENSO physics, and in capturing the interannual variability induced by ENSO, is not necessarily an accurate predictor of model skill in representing longer-timescale modes of climate variability (like the Pacific Decadal Oscillation and the Atlantic Multidecadal Oscillation).

[45] Because of the relatively short length of observational TLT records, there is considerable uncertainty in the obser-

vational estimate of low-frequency TLT variability. Longer observed SST records provide a stronger constraint on model estimates of decadal variability. We therefore estimated R_{band} and R_{high} with band- and high-pass filtered SST data from the CMIP-3 20CEN/A1B runs and version 3b of the NOAA ERSST observational data set [*Smith et al.*, 2008]. Filtering of SST data was performed as described above for TLT, but using the longer, 64-year period from January 1947 to December 2010 (see Figure 8b). Choice of this period avoids problems with observational SST data in the early 1940s [*Thompson et al.*, 2008]. All SSTs were spatially-averaged over 50°N–50°S, which minimizes model-versus-observed SST differences associated with inaccurate simulation of the latitudinal extent of ice margins.

[46] As in the case of TLT, the analysis of filtered SST data yields $R_{band} > 1.0$ and $R_{high} < 1.0$ ($R_{band} = 1.13$ and $R_{high} = 0.91$, respectively; see Figure 10). In contrast to *Swanson et al.* [2009], we find no evidence that CMIP-3 models systematically underestimate the amplitude of observed decadal SST variability. Since changes in SST and

Simulated and Observed Temporal Variability of Near-Global TLT Changes

Detrended TLT anomalies, Butterworth band- and high-pass filtering
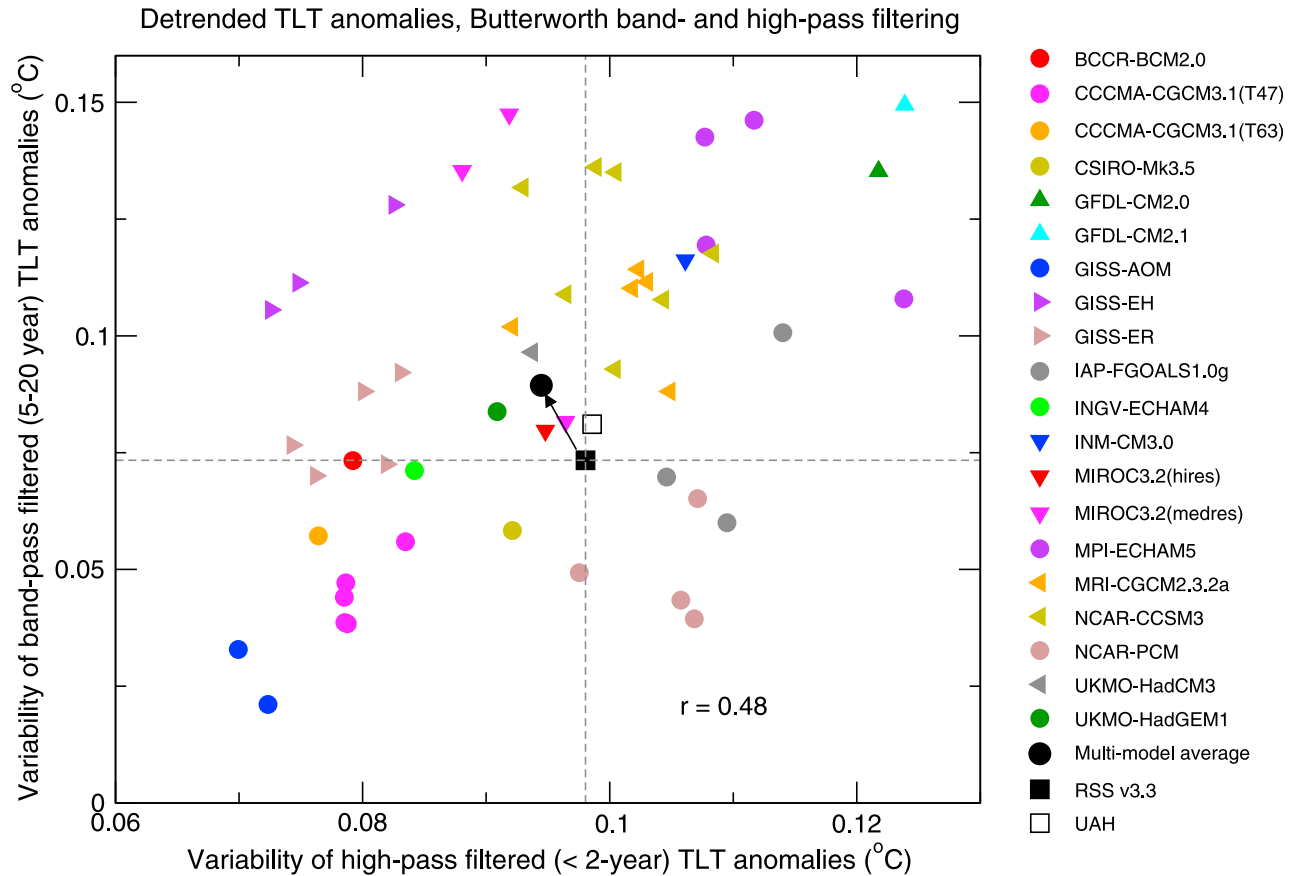

Figure 9. Comparison of simulated and observed temporal variability of near-global TLT anomalies. Variability on monthly to interannual timescales (*x*-axis) is plotted against variability on timescales of 5–20 years (*y*-axis). Filtering of model and observational TLT data was performed as described in Figure 8 and in the main text. All results are for the 384-month period from January 1979 to December 2010. Model results were calculated using synthetic TLT data from 51 realizations of the spliced 20CEN/A1B runs. The triangles denote 20CEN simulations which include volcanic forcing (see auxiliary material). With the exception of MPI-ECHAM5 and IAP-FGOALS1.0g, model simulations without volcanic forcing tend to have smaller values of low-frequency TLT variability. The dashed lines are centered on the RSS v3.3 result. Temporal standard deviations for the band- and high-pass filtered RSS v3.2 data (not shown) are virtually identical to the RSS v3.3 results. The multimodel average is the 'weighted' form (see Appendix A), calculated using the ensemble-mean temporal standard deviations.

TLT are closely coupled when averaged over sufficiently large spatial domains and timescales [*Wentz and Schabel*, 2000], our analysis of SST variability supports the direct analysis of TLT variability in Figure 9. Our SST results suggest that model-based estimates of TLT variability are adequate for the purpose of estimating the noise component of S/N ratios – at least on the 5- to 20-year timescales we focused on in the band-pass filtered data.

[47] In view of the difficulty of reliably constraining model variability estimates on multi-decadal timescales, it is useful to consider the implications of a large model error in the amplitude of low-frequency TLT variability. Even if we assumed that model-based estimates of internal variability on the 32-year timescale were biased low by 50%, S/N ratios for near-global TLT trends over the full satellite era would still be highly significant, and would exceed 2.6 in all

three observational TLT data sets (2.91 for RSS v3.2, 2.67 for RSS v3.3, and 2.61 for UAH). As noted above, our analysis of band-pass filtered TLT and SST data suggests that on average, the CMIP-3 models actually overestimate observed temperature variability on the 10-year timescale. Thus a 50% variability underestimate on the 32-year timescale appears unlikely, but cannot be definitively ruled out given the relatively short observational records available.

7. Discussion and Conclusions

[48] Efforts to apply rigorous statistical methods to the problem of identifying human effects on climate commenced over 30 years ago [*Hasselmann*, 1979]. At the inception of this endeavor, it was recognized that any human-caused climate change signal is embedded in the noise of natural
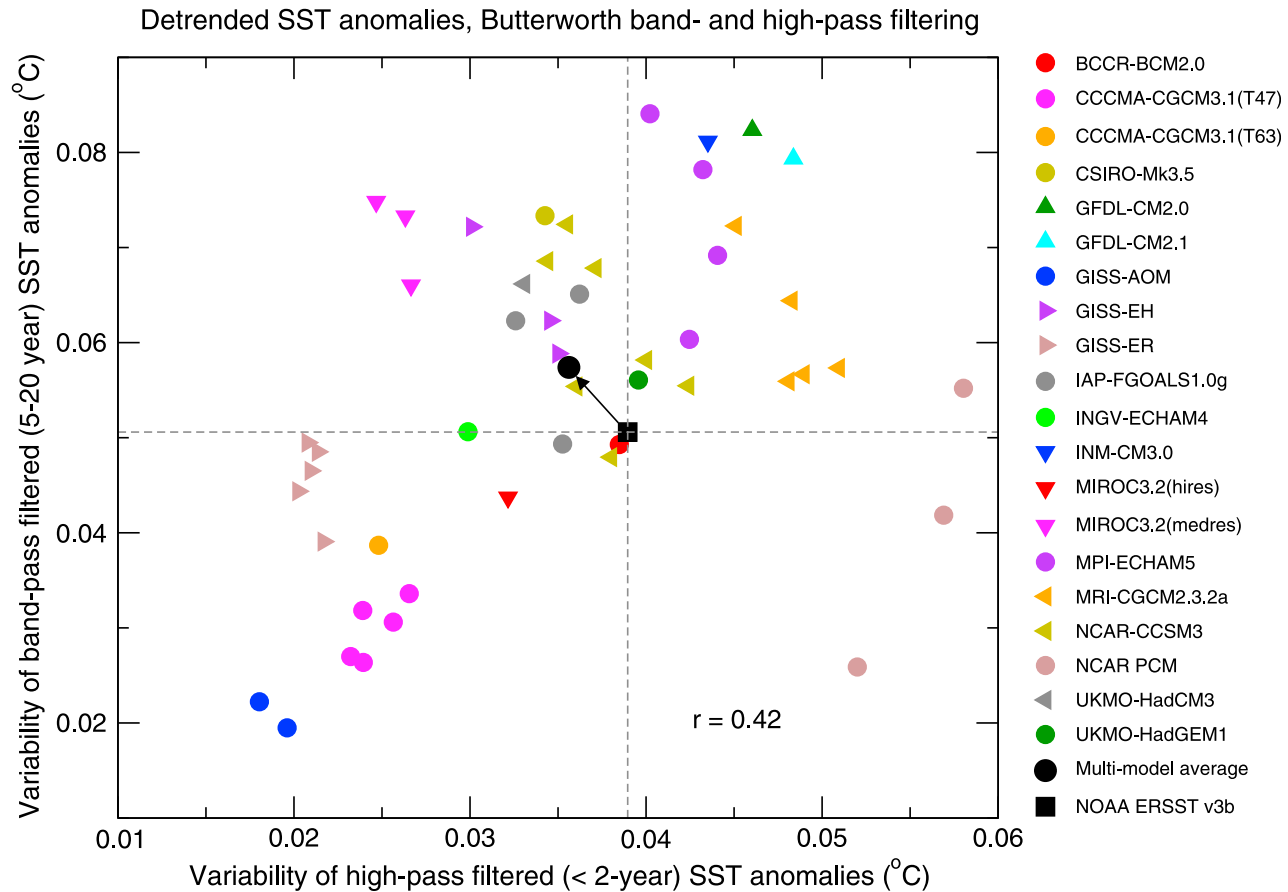
Figure 10. Same as Figure 9 but for the comparison of the simulated and observed temporal variability of near-global SST anomalies. Filtering of model and observational SST data was performed as described in Figure 8 and in the main text. All results are for the 768-month period from January 1947 to December 2010. There is one fewer 20CEN/A1B realization than in Figure 9 because realization 2 of the NCAR PCM 20CEN run commences in 1961, and does not cover the entire analysis period (see auxiliary material, Table S2). SST data were spatially averaged over 50°N–50°S. Observational SSTs are from version 3b of the NOAA ERSST data set [*Smith et al.*, 2008].

climate variability, and that separation of human and natural influences requires information on signal and noise properties over a range of timescales.

[49] We have provided such information here for the specific example of near-global changes in lower tropospheric temperature (TLT). We relied on control runs from the CMIP-3 multi-model archive for our estimates of climate noise. Estimates of externally forced climate-change signals were obtained from three different sets of satellite-based observations and from CMIP-3 simulations of 20th and 21st-century climate change. In contrast to almost all previous work, we compared modeled and observed TLT changes on multiple timescales (using maximally overlapping trends) rather than over a single period of record. For timescales less than the record length, this strategy reduces the impact of climate noise on estimates of the signal component of observed (and simulated) temperature trends. The fact that our $\overline{p_f}'$ values (even for 30-year TLT trends) are sensitive to the addition of a single year of observational data (see Figure 6e) indicates the dangers of ignoring the effects of interannual variability on signal

estimates, as was done, for example, by *Douglass et al.* [2008].

[50] Because of the large effect of year-to-year variability on decadal trends, roughly 10% of the 10-year TLT trends in the 20CEN/A1B runs are less than zero (Figure 4a). This result shows that anthropogenically forced models can replicate the recent muted warming of the surface [*Easterling and Wehner*, 2009; *Knight et al.*, 2009] and the lower troposphere. Claims that minimal warming over a single decade undermine findings of a slowly-evolving externally-forced warming signal (e.g., as in *Investor's Business Daily* [2008] and W. Happer, Testimony on climate science in the political arena, Hearing before the Select Committee on Energy Independence and Global Warming, House of Representatives, 111th Congress, May 20, 2010, http://globalwarming.house.gov/files/HRG/052010SciencePolicy/happer.pdf) are simply incorrect.

[51] Our estimated signal-to-noise (S/N) ratios for global-scale TLT changes were less than 1.0 on the 10-year timescale (Figure 6c). On the 32-year timescale, however, S/N exceeded 3.9 in all three observational TLT data sets.

The latter result shows that natural internal variability, as simulated by current climate models, is a highly unlikely explanation for the observed lower tropospheric warming over the satellite era (Figure 6d). Comparisons between simulated and observed low-frequency TLT variability suggest that our estimates of S/N ratios on 5–20 year timescales are conservative (Figures 9 and 10). The strong timescale dependence of S/N ratios arises primarily because of the large decrease in noise amplitude as the period used for trend fitting increases (Figure 6b).

[52] On all timescales examined here, the TLT trends in the observational satellite data sets are not statistically unusual relative to model-based distributions of externally forced TLT trends (Figure 6e). While this consistency is encouraging, it should be qualified by noting that: (1) the multi-model average TLT trend is always larger than the average observed TLT trend (Figures 6a and 6f); and (2) as the trend fitting period increases, values of $\overline{p_f}'$ decline, indicating that average observed trends are increasingly more unusual with respect to the multi-model distribution of forced trends (Figure 6e). Possible explanations for these results include the neglect of negative forcings in many of the CMIP-3 simulations of forced climate change (see auxiliary material, Table S1), omission of recent temporal changes in solar and volcanic forcing [*Wigley*, 2010; *Kaufmann et al.*, 2011; *Vernier et al.*, 2011; *Solomon et al.*, 2011], forcing discontinuities at the 'splice points' between CMIP-3 simulations of 20th and 21st century climate change [*Arblaster et al.*, 2011], model response errors, residual observational errors [*Mears et al.*, 2011], and an unusual manifestation of natural internal variability in the observations (see Figure 7a).

[53] Although we considered three different observational estimates of TLT changes (and one observational estimate of SST changes), our analysis does not comprehensively explore the impact of data uncertainties on model evaluation. In the future, studies of the consistency between simulated and observed temperature trends will be able to employ both multi-model ensembles (like CMIP-3) and new ensembles of observational upper-air [*Mears et al.*, 2011; *Thorne et al.*, 2011b] and SST data [*Kennedy et al.*, 2011a, 2011b]. These observational ensembles account for multiple sources of uncertainty in the construction of 'climate-quality' temperature records, and will allow us to explicitly include observational uncertainty in estimates of $\overline{p_f}'$ values.

[54] In summary, because of the effects of natural internal climate variability, we do not expect each year to be inexorably warmer than the preceding year, or each decade to be warmer than the last decade, even in the presence of strong anthropogenic forcing of the climate system. The clear message from our signal-to-noise analysis is that multi-decadal records are required for identifying human effects on tropospheric temperature. Minimal warming over a single decade does not disprove the existence of a slowly-evolving anthropogenic warming signal.

Appendix A

A1. Statistical Notation

[55] In the following, we provide a brief introduction to the statistical notation (Table A1) used in our discussion of

Table A1. Notation

Notation	Definition
	Abbreviations
MMA	Multi-model average
MMSD	Multi-model sampling distribution
	Subscripts
o	Subscript denoting observational data
c	Subscript denoting output from model control runs
f	Subscript denoting output from model forced experiments
	Indices
i	Index over number of maximally overlapping trends in observations
j	Index over number of models (and control run/forced run realizations)
	Sample Sizes
L	Length of trend-fitting period (months)
N_o	No. of overlapping L-month trends in observed data set
N_c	No. of overlapping L-month trends in control run MMSD
N_f	No. of overlapping L-month trends in 20CEN/A1B MMSD
$N_c(j)$	No. of overlapping L-month trends in jth model control run
$N_f(j)$	No. of overlapping L-month trends in jth model 20CEN/A1B run
N_{model}	No. of models (22 for control runs, 20 for 20CEN/A1B runs)
	Summation Variables
$K_c(i)$	No. of overlapping L-month trends in control run MMSD $> b_o(i)$
$K_f(i)$	No. of overlapping L-month trends in 20CEN/A1B MMSD $< b_o(i)$
$K_c(i, j)$	No. of overlapping L-month trends in jth model control run $> b_o(i)$
$K_f(i, j)$	No. of overlapping L-month trends in jth model 20CEN/A1B run $< b_o(i)$
	Linear Trends
$b_o(i)$	Linear trend for ith L-month segment of observed time series
$b_f(i, j)$	Linear trend for ith L-month segment of jth model's 20CEN/A1B time series
$\overline{b_o}$	Average (over index i) of $b_o(i)$
$\overline{b_f}(j)$	Average (over index i) of $b_f(i, j)$
$\overline{\overline{b_f}}$	MMA of overlapping L-month trends from 20CEN/A1B runs
	Statistics for Model-Versus-Observed Trend Comparisons
$p_c(i)$	Unweighted p-value for comparison of $b_o(i)$ and control run MMSD
$p_f(i)$	Unweighted p-value for comparison of $b_o(i)$ and 20CEN/A1B MMSD
$p_c(i, j)$	p-value for comparison of $b_o(i)$ and jth model control run
$p_f(i, j)$	p-value for comparison of $b_o(i)$ and jth model 20CEN/A1B run
$p_c(i)'$	Weighted p-value, model average of $p_c(i, j)$
$p_f(i)'$	Weighted p-value, model average of $p_f(i, j)$
$\overline{p_c}'$	Weighted p-value, average over index i of $p_c(i)'$
$\overline{p_f}'$	Weighted p-value, average over index i of $p_f(i)'$

model-versus-observed trend comparisons. We perform such comparisons on timescales ranging from 120 months to 384 months (in increments of 12 months). For the sake of simplicity, we do not explicitly include the selected analysis timescale in our notation.

[56] Note that the index j is a combined index over models and over realizations of either the control run or 20CEN/A1B run. For example, the PCM model has four different realizations of the spliced 20CEN/A1B run (see Table S2 in the auxiliary material). In the $L = 120$-month case, and for

maximally overlapping trends calculated over the 384-month period January 1979 to December 2010, PCM provides 265×4 samples of forced TLT trends, and $N_f(j) = 1060$. All 1060 of these 120-month TLT trends were used in computing $p_f(i, j)$ values for PCM. Alternately, some models have multiple realizations of the pre-industrial control run, or (because of missing data), multiple segments of a single pre-industrial control run (see Table S3 in the auxiliary material). For example, two realizations of the GISS-AOM control run were performed, each of length 3012 months. So for the $L = 120$-month case, the GISS-AOM provides 2893×2 samples of overlapping 120-month trends, and the GISS-AOM value of $N_c(j) = 5786$.

A2. Statistical Analysis

A2.1. Calculation of p-Values

[57] We calculate two different types of p-value. The first type, $p_c(i)$ (where i is an index over the number of maximally overlapping observed trends), is for comparisons of observed trends and trends estimated from CMIP-3 model control runs with no changes in natural or anthropogenic forcings. The second type of p-value, $p_f(i)$, is based on comparisons of observed trends against the externally-forced trends in the spliced 20CEN/A1B experiments.

[58] As used here and subsequently, 'overlapping' signifies trend overlap by all but one month. For $L = 120$ months, the first trend is over January 1979 to December 1988, the second trend is over February 1979 to January 1989, etc. Note that all least-squares linear trends were computed from time series of monthly-mean anomalies of spatially-averaged (82.5°N–70°S) observed and simulated TLT data. Anomalies in the 20CEN/A1B runs were defined relative to climatological monthly means over the 384-month period January 1979 to December 2010. Control run anomalies were defined relative to climatological monthly means over the full length of each model's control integration.

[59] We compute both 'unweighted' and 'weighted' forms of $p_c(i)$ and $p_f(i)$. The weighted forms, $p_c(i)'$ and $p_f(i)'$, are distinguished by the use of prime notation ('), and account for inter-model differences in either the length/number of realizations of the control run or in the number of realizations of the spliced 20CEN/A1B run (respectively).

[60] Consider first the 'unweighted' form of $p_c(i)$. For a stipulated trend length L (in months), the $p_c(i)$ value is defined as:

$$p_c(i) = K_c(i)/N_c$$
$$i = 1, \ldots, N_o \tag{A1}$$

where $K_c(i)$ is the number of L-month trends in the MMSD of control run trends that are larger than $b_o(i)$ (the current L-month observed trend), N_c is the total number of overlapping L-month trends in the MMSD of control run trends, and N_o is the total number of overlapping L-month observed trends in the 384-month analysis period. For $L = 120$ months, $N_c = 120965$ and $N_o = 265$.

[61] The time series of spatially-averaged TLT anomalies from individual models are not concatenated prior to trend calculation (which could spuriously inflate trends spanning the 'splice point' between two different model control runs).

Instead, overlapping trends are calculated separately from each realization of each individual model's TLT time series, and each model's TLT trends are then accumulated in a multi-model trend distribution.

[62] In the 'weighted' form, $p_c(i)'$, individual $p_c(i, j)$ values are first calculated separately for each model, and the accumulated $p_c(i, j)$ values are then averaged:

$$p_c(i)' = \sum_{j=1}^{N_{model}} p_c(i, j)/N_{model}$$
$$i = 1, \ldots, N_o \tag{A2}$$

where j is an combined index over the number of models and the number of control run realizations per model, and N_{model} (the number of CMIP-3 models with pre-industrial control runs from which synthetic MSU temperatures could be calculated) = 22. The individual $p_c(i, j)$ values for each model are calculated as follows:

$$p_c(i, j) = K_c(i, j)/N_c(j)$$
$$i = 1, \ldots, N_o; \; j = 1, \ldots, N_{model} \tag{A3}$$

where $K_c(i, j)$ is, for the ith observed trend and the jth model, the number of L-month trends in the pre-industrial control run larger than $b_o(i)$.

[63] Values of $p_c(i)$ and $p_c(i)'$ are very similar, indicating that inter-model differences in control run length do not distort our estimates of whether observed TLT trends are unusually large relative to trends arising from internally-generated variability. We show only 'weighted' $p_c(i)'$ values in the main text.

[64] In comparisons involving forced trends from the CMIP-3 20CEN/A1B runs, we seek to determine whether the model TLT trends are unusually large relative to observed trends, as some analysts have claimed [*Douglass et al.*, 2008]. Values of $p_f(i)$ are defined in an analogous way to $p_c(i)$ values:

$$p_f(i) = K_f(i)/N_f$$
$$i = 1, \ldots, N_o \tag{A4}$$

where $K_f(i)$ is the number of L-month trends in the 20CEN/A1B MMSD that are smaller than the current observed trend, N_f is the total number of overlapping L-month trends in the 20CEN/A1B MMSD, and N_o is the total number of overlapping L-month observed trends in the 384-month analysis period.

[65] Unlike $p_c(i)$ calculations with the CMIP-3 pre-industrial control runs (where synthetic TLT data for the full length of each control run were used in the calculations, but only 384 months of observational TLT data were analyzed), all $p_f(i)$ and $p_f(i)'$ values were computed using the same 384-month period (January 1979 to December 2010) in the spliced 20CEN/A1B runs and the observations. The spliced 20CEN/A1B runs provide a total of 51 realizations of forced TLT changes over January 1979 to December 2010. For the case of overlapping 120-month trends, $N_f = 13515$ (265×15).

[66] As in the comparisons with control run trends, a 'weighted' form of $p_f(i)$ can be calculated:

$$p_f(i)' = \sum_{j=1}^{N_{model}} p_f(i,\,j)/N_{model} \tag{A5}$$
$$i = 1, \ldots, N_o$$

where N_{model} (the number of CMIP-3 models with 20CEN and A1B runs from which synthetic MSU temperatures could be calculated) = 20, and $p_f(i, j)$ is defined in an analogous way to $p_c(i, j)$ in equation (A3). Averaging the N_o individual values of $p_f(i)'$ yields $\overline{p_f}'$:

$$\overline{p_f}' = \sum_{i=1}^{N_o} p_f(i)'/N_o \tag{A6}$$
$$i = 1, \ldots, N_o$$

with $\overline{p_c}'$ defined similarly.

[67] Our use of maximally overlapping trends has the advantage of reducing the impact of seasonal and interannual noise on estimates of the signal components of TLT trends, both in the observations and in the spliced 20CEN/A1B runs. However, it has the disadvantage of decreasing the statistical independence of trend samples.

[68] While non-independence of samples is an important issue in formal statistical significance testing, it is not a serious concern here. This is because our $p_c(i)'$ and $p_f(i)'$ values are not used as a basis for formal statistical tests. Instead, they simply provide useful information on whether observed TLT trends are unusually large relative to model-based estimates of unforced trends, or unusually small relative to model estimates of externally-forced trends. Note also that we process observed TLT data and model output in identical ways, with the same overlap between successive L-month trends – i.e., we are not generating fundamentally different temporal autocorrelation structure in the model and observational trend samples.

[69] The key point is that whether we employ overlapping or non-overlapping model trends has very small impact on estimates of $p_c(i)'$ or $p_f(i)'$. This suggests that the sample sizes of non-overlapping trends (in both the CMIP-3 control runs and the 20CEN/A1B runs) may be adequate for obtaining reasonable estimates of $p_c(i)'$ and $p_f(i)'$.

[70] However, because of the relatively short length of satellite temperature records, the use of non-overlapping observed TLT trends can have a large impact on both $p_c(i)'$ and $p_f(i)'$. For each observational TLT data set, the 1979 to 2010 analysis period contains three non-overlapping 10-year trends, two non-overlapping trends >10 years and ≤16 years, and only one non-overlapping trend >16 years and ≤32 years. As shown in Figure 3a of the main text, the use of non-overlapping time series segments does not adequately sample the impact of interannual variability on trends. This is why we focus primarily on $p_c(i)'$ and $p_f(i)'$ values calculated with overlapping L-month observed trends.

[71] The implicit assumption in all of our p-value calculations is that results from individual models are independent. This assumption is almost certainly unjustified [*Masson and Knutti*, 2011]. While it would be interesting to explore the sensitivity of trend consistency results to the selection of different subsets of "independent" CMIP-3 models, we do not perform such an analysis here. We suspect that the identification of "independent" model subsets may be sensitive to the variables, statistical procedures, and metrics used to assess inter-model dependencies.

A2.2. Calculation of Signal-to-Noise Ratios

[72] Two types of signal-to-noise ratio are shown in Figure 6. The first is the 'observed' signal-to-noise ratio, R_o, in Figure 6c:

$$R_o = \overline{b_o}/s\{b_c\} \tag{A7}$$

where $\overline{b_o}$ is the average of all overlapping L-month observed TLT trends, and $s\{b_c\}$ is the standard deviation of the MMSD of overlapping L-month control run TLT trends. The model signal-to-noise ratio in Figure 6c, R_f, is defined similarly:

$$R_f = \overline{\overline{b_f}}/s\{b_c\} \tag{A8}$$

where $\overline{\overline{b_f}}$ is the MMA of the overlapping L-month TLT trends obtained from the 20CEN/A1B runs. Figure 6c shows R_o and R_f for 23 different values of L (120, 132, 144, …, 372, and 384 months).

[73] Note that, in an analogous way to the calculation of unweighted and weighted p-values, unweighted and weighted forms of the MMA can be computed. The unweighted MMA is simply the arithmetic average of all available overlapping, L-month trends in the 20CEN/A1B runs. The weighted MMA (which is what we use here, and what we show in Figure 6a of the main text) is calculated by first computing (for the jth model) the average of all available overlapping, L-month trends in all realizations of the jth models's 20CEN/A1B runs, and then averaging these ensemble-mean 'average' trends. For the sample sizes of forced TLT trends available here, weighted and unweighted forms of the MMA yield very similar results.

A2.3. Calculation of Model Average TLT Variability

[74] The grey shaded envelope in Figure 7a provides information on the model average TLT variability. This was estimated in the following way. For the first realization of each of the CMIP-3 pre-industrial control runs for which synthetic MSU temperatures could be calculated (see Table S3 in the auxiliary material), we computed spatial averages of TLT over 82.5°N–70°S (the latitudinal extent of RSS TLT data). From each of these 22 time series, we extracted the first 1200 months of near-global TLT data, defined anomalies relative to climatological monthly means, and then detrended the anomaly data. These detrended anomalies were used to calculate $s_c^2(j)$, the temporal TLT variance of the jth model. The model average variability, $\overline{s_c}$, is given by:

$$\overline{s_c} = \sqrt{\overline{s_c^2}} \tag{A9}$$

where $\overline{s_c^2}$ is the model-average temporal variance. Use of the first 1200 months of each model's pre-industrial control run (and use of the first realization only) ensures that our estimate of $\overline{s_c}$ is not biased by inter-model differences in the length and number of realizations of the control run.

References

Allen, M. R., and S. F. B. Tett (1999), Checking for model consistency in optimal fingerprinting, *Clim. Dyn.*, *15*, 419–434.

Allen, M. R., C. T. Mutlow, G. M. C. Blumberg, J. R. Christy, R. T. McNider, and D. T. Llewellyn-Jones (1994), Global change detection, *Nature*, *370*, 24–25.

Arblaster, J. M., G. A. Meehl, and D. J. Karoly (2011), Future climate change in the Southern Hemisphere: Competing effects of ozone and greenhouse gases, *Geophys. Res. Lett.*, *38*, L02701, doi:10.1029/2010GL045384.

Barnett, T. P., D. W. Pierce, and R. Schnur (2001), Detection of anthropogenic climate change in the world's oceans, *Science*, *292*, 270–274.

Barnett, T. P., D. W. Pierce, K. M. AchutaRao, P. J. Gleckler, B. D. Santer, J. M. Gregory, and W. M. Washington (2005), Penetration of human-induced warming into the world's oceans, *Science*, *309*, 284–287.

Christy, J. R., W. B. Norris, R. W. Spencer, and J. J. Hnilo (2007), Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements, *J. Geophys. Res.*, *112*, D06102, doi:10.1029/2005JD006881.

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

Easterling, D. R., and M. F. Wehner (2009), Is the climate warming or cooling?, *Geophys. Res. Lett.*, *36*, L08706, doi:10.1029/2009GL037810.

Forster, P., et al. (2007), Changes in atmospheric constituents and in radiative forcing, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 130–234, Cambridge Univ. Press, Cambridge, U. K.

Fu, Q., S. Manabe, and C. M. Johanson (2011), On the warming in the tropical upper troposphere: Models versus observations, *Geophys. Res. Lett.*, *38*, L15704, doi:10.1029/2011GL048101.

Gaffen, D., et al. (2000), Multi-decadal changes in the vertical temperature structure of the tropical troposphere, *Science*, *287*, 1239–1241.

Gillett, N. P., F. W. Zwiers, A. J. Weaver, G. C. Hegerl, M. R. Allen, and P. A. Stott (2002), Detecting anthropogenic influence with a multi-model ensemble, *Geophys. Res. Lett.*, *29*(20), 1970, doi:10.1029/2002GL015836.

Hasselmann, K. (1979), On the signal-to-noise problem in atmospheric response studies, in *Meteorology of Tropical Oceans*, edited by D. B. Shaw, pp. 251–259, R. Meteorol. Soc., London.

Hasselmann, K. (1993), Optimal fingerprints for the detection of time dependent climate change, *J. Clim.*, *6*, 1957–1971.

Hegerl, G. C., and F. W. Zwiers (2011), Use of models in detection and attribution of climate change, *Wiley Interdisc. Rev. Clim. Change*, *2*, 570–591, doi:10.1002/wcc.121.

Hegerl, G. C., H. von Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones (1996), Detecting greenhouse-gas-induced climate change with an optimal fingerprint method, *J. Clim.*, *9*, 2281–2306.

Hegerl, G. C., et al. (2007), Understanding and attributing climate change, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 664–745, Cambridge Univ. Press, Cambridge, U. K.

Henson, S. A., et al. (2010), Detection of anthropogenic climate change in satellite records of ocean chlorophyll and productivity, *Biogeosciences*, *7*, 621–640.

Intergovernmental Panel on Climate Change (IPCC) (2007), Summary for policymakers, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., pp. 1–7, Cambridge Univ. Press, Cambridge, U. K.

Investor's Business Daily (2008), Alarmists still heated even as world cools, 4 November.

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

Karl, T. R., S. J. Hassol, C. D. Miller, and W. L. Murray (Eds.) (2006), Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences – A report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research, report, Natl. Clim. Data Cent., Asheville, N. C.

Kaufmann, R. K., H. Kauppi, M. L. Mann, and J. H. Stock (2011), Reconciling anthropogenic climate change with observed temperature 1998–2008, *Proc. Natl. Acad. Sci. U. S. A.*, *108*, 11,790–11,793, doi:10.1073/pnas.1102467108.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby (2011a), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties, *J. Geophys. Res.*, *116*, D14103, doi:10.1029/2010JD015218.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby (2011b), Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization, *J. Geophys. Res.*, *116*, D14104, doi:10.1029/2010JD015220.

Knight, J., et al. (2009), Do global temperature trends over the last decade falsify climate predictions?, *Bull. Am. Meteorol. Soc.*, *90*, S22–S23.

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.

Liebmann, B., R. M. Dole, C. Jones, I. Bladé, and D. Allured (2010), Influence of choice of time period on global surface temperature trend estimates, *Bull. Am. Meteorol. Soc.*, *91*, 1485–1491, doi:10.1175/2010BAMS3030.1.

Madden, R. A., and V. Ramanathan (1980), Detecting climate change due to increasing carbon dioxide, *Science*, *209*, 763–768.

Masson, D., and R. Knutti (2011), Climate model genealogy, *Geophys. Res. Lett.*, *38*, L08703, doi:10.1029/2011GL046864.

McKitrick, R., S. McIntyre, and C. Herman (2010), Panel and multivariate methods for tests of trend equivalence in climate data series, *Atmos. Sci. Lett.*, *11*(4), 270–277.

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

Mears, C. A., C. E. Forest, R. W. Spencer, R. S. Vose, and R. W. Reynolds (2006), What is our understanding of the contribution made by observational or methodological uncertainties to the previously reported vertical differences in temperature trends?, in *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences, A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research*, edited by T. R. Karl et al., chap. 4, Natl. Clim. Data Cent., Asheville, N. C.

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

Meehl, G. A., et al. (2007), The WCRP CMIP3 multi-model dataset: A new era in climate change research, *Bull. Am. Meteorol. Soc.*, *88*, 1383–1394.

Nakićenović, N., and R. Swart (Eds.) (2000), *Special Report on Emissions Scenarios. A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*, Cambridge Univ. Press, Cambridge, U. K.

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

North, G. R., K. Y. Kim, S. S. P. Shen, and J. W. Hardin (1995), Detection of forced signals. Part 1: Filter theory, *J. Clim.*, *8*, 401–408.

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

Santer, B. D., W. Brüggemann, U. Cubasch, K. Hasselmann, H. Höck, E. Maier-Reimer, and U. Mikolajewicz (1994), Signal-to-noise analysis of time-dependent greenhouse warming experiments. Part 1: Pattern analysis, *Clim. Dyn.*, *9*, 267–285.

Santer, B. D., et al. (1995), Ocean variability and its influence on the detectability of greenhouse warming signals, *J. Geophys. Res.*, *100*, 10,693–10,725.

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

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

Santer, B. D., et al. (2006), How well can the observed vertical temperature changes be reconciled with our understanding of the causes of these

changes?, in *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*, edited by T. R. Karl et al., Natl. Clim. Data Cent., Asheville, N. C.

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

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore (2008), Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006), *J. Clim.*, *21*, 2283–2296.

Solomon, S., K. H. Rosenlof, R. W. Portman, J. S. Daniel, S. M. Davis, T. J. Sanford, and G.-K. Plattner (2010), Contributions of stratospheric water vapor to decadal changes in the rate of global warming, *Science*, *327*, 1219–1223.

Solomon, S., J. S. Daniel, R. R. Neely III, J. P. Vernier, E. G. Dutton, and L. W. Thomason (2011), The persistently variable "background" stratospheric aerosol layer and global climate change, *Science*, *333*, 866–870, doi:10.1126/science1206027.

Stott, P. A., and S. F. B. Tett (1998), Scale-dependent detection of climate change, *J. Clim.*, *11*, 3282–3294.

Stott, P. A., S. F. B. Tett, G. S. Jones, M. R. Allen, J. F. B. Mitchell, and G. J. Jenkins (2000), External control of 20th century temperature by natural and anthropogenic forcings, *Science*, *290*, 2133–2137.

Swanson, K. L., G. Sugihara, and A. A. Tsonis (2009), Long-term natural variability and 20th century climate change, *Proc. Natl. Acad. Sci. U. S. A.*, *106*, 16,120–16,123.

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

Thompson, D. W. J., J. J. Kennedy, J. M. Wallace, and P. D. Jones (2008), A large discontinuity in the mid-twentieth century in observed global-mean surface temperature, *Nature*, *453*, 646–649.

Thorne, P. W., et al. (2007), Tropical vertical temperature trends: A real discrepancy?, *Geophys. Res. Lett.*, *34*, L16702, doi:10.1029/2007GL029875.

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

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

Vernier, J.-P, et al. (2011), Major influence of tropical volcanic eruptions on the stratospheric aerosol layer during the last decade, *Geophys. Res. Lett.*, *38*, L12807, doi:10.1029/2011GL047563.

Weatherhead, E. C., et al. (1998), Factors affecting the detection of trends: Statistical considerations and applications to environmental data, *J. Geophys. Res.*, *103*, 17,149–17,161.

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

Wentz, F. J., and M. Schabel (2000), Precise climate monitoring using complementary satellite data sets, *Nature*, *403*, 414–416.

Wigley, T. M. L. (2010), A review and assessment of past and future changes in global-mean temperature, report, 74 pp., Electr. Power Res. Inst., Palo Alto, Calif.

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

Wigley, T. M. L., and S. C. B. Raper (1990), Natural variability of the climate system and detection of the greenhouse effect, *Nature*, *344*, 324–327.

P. Caldwell, C. Doutriaux, P. J. Gleckler, D. Ivanova, B. D. Santer, and K. E. Taylor, Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. (santer1@llnl.gov)

N. P. Gillett, Canadian Centre for Climate Modelling and Analysis, Environment Canada, Victoria, BC V8W 3V6, Canada.

T. R. Karl and P. W. Thorne, National Climatic Data Center, National Oceanic and Atmospheric Administration, Asheville, NC 28801, USA.

J. R. Lanzante, Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration, Princeton, NJ 08542, USA.

C. Mears and F. J. Wentz, Remote Sensing Systems, Santa Rosa, CA 95401, USA.

G. A. Meehl and T. M. L. Wigley, National Center for Atmospheric Research, Boulder, CO 80307, USA.

S. Solomon, Department of Atmospheric and Oceanic Sciences, University of Colorado at Boulder, Boulder, CO 80305, USA.

P. A. Stott, Met Office Hadley Centre, Exeter EX1 3PB, UK.

M. F. Wehner, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.