Causes of differences in model and satellite tropospheric warming rates

Benjamin D. Santer[1], John C. Fyfe[2], Giuliana Pallotta[1], Gregory M. Flato[2], Gerald A. Meehl[3], Matthew H. England[4], Ed Hawkins[5], Michael E. Mann[6], Jeffrey F. Painter[1], Céline Bonfils[1], Ivana Cvijanovic[1], Carl Mears[7], Frank J. Wentz[7], Stephen Po-Chedley[8], Qiang Fu[8] & Cheng-Zhi Zou[9]

[1]Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA 94550, USA.

[2]Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, British Columbia, V8W 2Y2, Canada.

[3]National Center for Atmospheric Research, Boulder, Colorado 80307, USA.

[4]ARC Centre of Excellence for Climate System Science, University of New South Wales, New South Wales 2052, Australia.

[5]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading RG6 6BB, UK.

[6]Department of Meteorology and Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania, USA.

[7]Remote Sensing Systems, Santa Rosa, CA 95401, USA.

[8]Dept. of Atmospheric Sciences, University of Washington, Seattle, WA 98195, USA.

[9]Center for Satellite Applications and Research, NOAA/NESDIS, Camp Springs, Maryland 20746, USA.

Submitted to *Nature Geoscience*

February 10, 2017

In the early 21st century, observed surface and tropospheric warming trends were smaller than trends estimated from the average of a large multi-model ensemble [1, 2, 3]. Because observations and "free running" model simulations do not have the same phasing of natural internal variability, decadal differences in simulated and observed warming rates invariably occur [4, 5, 6, 7, 8]. It is unclear whether the magnitude of such differences can be explained by internal climate variability alone [9, 10]. We address this question using global changes in tropospheric temperature estimated from satellites and climate models. Here we show that in the last two decades of the 20th century, differences between modeled and observed tropospheric temperature trends are broadly consistent with internal variability. Over most of the early 21st century, however, model tropospheric warming is substantially larger than observed; warming rate differences are generally outside the range of trends arising from internal variability. There is a low probability (between $< 1\%$ and $\approx 7\%$) that multi-decadal internal variability fully explains this asymmetry in the statistical significance of the late 20th and early 21st century results. Our study provides new and independent support for findings that model overestimation of warming in the early 21st century is partly due to systematic model deficiencies in representing some of the post-2000 external cooling influences experienced by the real world [11, 12, 13, 14].

The Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) contained prominent discussion of differences between warming rates in observations and model simulations [15, 16]. The focus of this discussion was on two issues: the cause or causes of a putative "slowdown" in observed surface and tropospheric warming during the early 21st century, and the reasons for the inability of most climate model simulations to capture this "slowdown".

Since publication of the Fifth Assessment Report, progress has been made in improving scientific understanding of both issues [3]. With regard to the first issue, there is now widespread recognition that the observed "slowdown" was not due to a single cause, as several studies had claimed [9, 10]. Recent research points towards the combined effects of internal variability [5, 6, 7, 8, 9], the cooling caused by a succession of moderate early 21st century volcanic eruptions [1, 12, 17, 18, 19], an unusually long and low solar minimum during the last solar cycle [13], a decrease in stratospheric water vapor [11], and an increase in forcing by anthropogenic sulfate aerosols [14, 20]. The scientific challenge is to reliably estimate the individual contributions of these factors to the observed "slowdown", and to better quantify uncertainties in the observed temperature data used to study this phenomenon [21, 22].

The second issue has also been the subject of considerable attention. There is now widespread acceptance that larger simulated warming during the "slowdown" period arises from a combination of factors. These factors include systematic deficiencies in

how models represent early 21st century changes in key natural and anthropogenic external forcings [12, 13, 14, 17, 18, 23], model errors in response to forcing [24], errors in the observations themselves [21, 22], different sequences of internal variability in the real world and the model world [5, 6, 7, 8, 9, 25], and differences in the physical quantities being compared in observations and model simulations [26].

Other aspects of the "slowdown" – such as its statistical significance – remain the subject of vigorous scientific debate. One viewpoint is that the "slowdown" is a routine decadal fluctuation in temperature, and is not statistically different from previous manifestations of internal climate variability [27, 28]. Another perspective is that the "slowdown" was physically unusual, and resulted from the unusual temporal coincidence of a negative (cooling) phase of the Interdecadal Pacific Oscillation (IPO) and multiple external cooling influences [3]. From this second perspective, the question of whether the "slowdown" was routine cannot be viewed in purely statistical terms: it must also be evaluated in the context of the time histories of radiative forcing and internal variability. These statistical and physical perspectives are not mutually exclusive. Both views of the "slowdown" have scientific value.

Most statistical studies of the "slowdown" have framed the "slowdown" as a routine fluctuation [27], with a warming rate that is not significantly different from 20th century warming [28]. To the best of our knowledge, no previous study has explicitly considered whether differences between simulated and observed tropospheric warming

rates during the "slowdown" period are significantly larger than can be plausibly explained by internal variability alone. Failure to find such significant trend differences would lend support to the routine nature of the "slowdown".

We perform such significance tests here using satellite- and model-based estimates of tropospheric temperature. There are two reasons for this choice. First, satellite tropospheric temperature measurements have time-invariant, near-global coverage [22, 29, 30]. In contrast, there are large, non-random temporal changes in spatial coverage in the observed surface temperature datasets used in most "slowdown" studies [21, 31]. Second, satellite tropospheric temperature datasets have been a key component of recent claims that current climate models are too sensitive (by a factor of three or more) to human-caused changes in greenhouse gases [32, 33]. Errors of this magnitude would diminish confidence in model projections of future climate change. It is therefore critically important to evaluate the validity of such claims.

We focus on satellite-based measurements of global-scale changes in the temperature of the mid- to upper troposphere (TMT). TMT data with near-global coverage are available from three groups: Remote Sensing Systems (RSS) [22], the Center for Satellite Applications and Research (STAR) [30], and the University of Alabama at Huntsville (UAH) [34]. Older and more recent dataset versions are provided by each of these groups (see Supplementary Material). A fourth group (the University of Washington; UW) [29] produces TMT data for a tropical domain. We briefly discuss

both tropical-scale TMT changes and global-scale changes in the temperature of the lower troposphere (TLT); the latter are provided by RSS and UAH only.

Because TMT receives a contribution from the cooling of the stratosphere, a standard regression-based approach was employed to correct for this influence [35]. Correction yields a more representative measure of bulk changes in tropospheric temperature [36, 37, 38], and was performed for both satellite and model TMT data, as described in the Supplementary Material, Section 2.

Model synthetic TMT data are from simulations of historical climate change ("HIST") and from simulations of 21st century climate change under Representative Concentration Pathway 8.5 (RCP8.5). We also analyze control runs with no changes in external forcings. The historical and RCP8.5 integrations yield information on the tropospheric temperature response to combined anthropogenic and natural external forcing. The control runs provide estimates of the natural internal climate variability, which we use in assessing the significance of differences between modeled and observed TMT trends.

The HIST numerical experiments typically begin in the mid-19th century and end in 2005. To compare models and observations over the full satellite temperature record (January 1979 to June 2016), HIST temperatures were spliced with temperatures from the RCP8.5 runs ("HIST+8.5"; see Supplementary Material, Section 3.3). The HIST, RCP8.5, and control simulations were performed under phase 5 of the Coupled Model

Intercomparison Project (CMIP5) [39].

The multi-model average (MMA) of TMT changes in the HIST+8.5 simulations is noticeably smoother than any individual observational TMT time series (see, *e.g.*, the RSS results in Fig. 1A). This difference in the amplitude of variability is expected [1, 6, 40]. In "free running" simulations with coupled models of the climate system, the phasing of internally generated climate variability is random. By averaging over 49 different realizations of HIST+8.5 (performed with 37 different climate models), the amplitude of random variability is reduced, more clearly revealing the underlying temperature response to external forcings. In the real world, however, there is only one sequence of internal climate variability.

In the MMA, tropospheric warming over the satellite era is larger than in observations [41]. Differences between simulated and observed warming rates are noticeable in the early 21st century (Figs. 1A, B). Another prominent feature of the observational results in Fig. 1 is the pronounced tropospheric warming associated with major El Niño events. In observations, the large 1982/83 El Niño partly obscured cooling caused by the 1982 eruption of El Chichón. Because of the above-described noise reduction arising from averaging over realizations and models, the cooling signatures of El Chichón and Pinatubo are clearer in the MMA [1, 42]. Removal of temperature variability induced by the internally generated El Niño/Southern Oscillation (ENSO) improves the agreement between volcanic cooling signals in the MMA and in satel-

lite tropospheric temperature data, but does not fully explain mismatches between simulated and observed tropospheric warming during the early 21st century [1].

Next, we assess whether there are statistically significant differences between estimated tropospheric temperature changes in models and individual satellite temperature datasets. We operate on the difference series $\Delta T_{f-o}(k,t) = \overline{\overline{T}}_f(t) - T_o(k,t)$, where k is an index over the number of observational datasets, t is an index over time (in months), $\overline{\overline{T}}_f(t)$ is the MMA, and $T_o(k,t)$ is an individual observational temperature time series (see Supplementary Material, Section 4.1).

Our significance testing procedure rests on two assumptions. First, we assume that the MMA provides a credible, "noise free" estimate of the true (but unknown) externally forced tropospheric temperature signal in the real world. If this assumption were valid, the difference series $\Delta T_{f-o}(k,t)$ should largely reflect the departures of the observed realization of internal variability from the externally forced signal.[1] A second necessary assumption is that the CMIP5 control runs provide unbiased estimates of the amplitude of natural internal variability, particularly on interannual to multi-decadal timescales (see Supplementary Material, Section 4.6).[2]

[1]It is well-known, however, that systematic errors in certain external forcings in the HIST+8.5 simulations introduce biases in the MMA [11, 12, 13, 14, 20]. The behavior of $\Delta T_{f-o}(k,t)$ must also receive a contribution from these forcing errors. We seek to determine whether this contribution is large enough to be discriminated from the effects of internal variability.

[2]Another (less critical) assumption implicit in our analysis is that the statistical properties of "real-world" internal variability have not been strongly modulated by external forcing during the

Under these two assumptions, we formulate the null hypothesis that departures between the expected and observed tropospheric temperature trends are consistent with internal climate noise. Rejection of the null hypothesis can have multiple explanations: systematic deficiencies in the model external forcings applied in the HIST+8.5 simulations,[3] errors in the climate response to these forcings, errors in the simulated spectrum of internal variability, and residual inhomogeneities in the satellite temperature measurements. These explanations are not mutually exclusive.

In previous studies of early 21st century differences between simulated and observed warming rates, most attention focused on temperature changes over specific 10- to 15-year periods [2, 3, 21, 43]. The appropriateness of specific analysis period choices has been the subject of debate [3, 21]. To avoid such debate, we focus instead on L-year analysis timescales. We consider six timescales here: $L = 10$, 11, 12, 13, 14 and 15 years. For each timescale, an L-year "window" is advanced by one month at a time through the $\Delta T_{f-o}(k,t)$ difference series. A least-squares linear trend is calculated for each window. We refer to these subsequently as "maximally overlapping trends". A key element of our analysis is that the significance of each overlapping L-year trend in $\Delta T_{f-o}(k,t)$ is evaluated relative to control run distributions of L-year temperature trends – *i.e.*, the difference series trend is compared with the behavior of noise trends on the same timescale.

satellite era.

[3]Such as the neglect of moderate volcanic eruptions in the early 21st century [12, 17, 18, 19].

Maximally overlapping trends are plotted in the left column of Fig. 2. As expected, shorter L-year trends are noisier. For example, for 10-year windows ending close to 1999, difference series trends show large negative values associated with the observed warming caused by the 1997/98 El Niño. The use of longer trend-fitting periods damps such end-point effects.

The overlapping trends in Fig. 2 have large, positive values for L-year windows sampling a substantial portion of the early 21st century. During this period, the average simulated warming is larger than the observed tropospheric warming in each satellite dataset. We use CMIP5 control runs[4] to estimate the probability that trends in $\Delta T_{f-o}(k,t)$ are either unusually large or unusually small relative to unforced temperature trends (see Supplementary Material, Section 4.3). The resulting empirical p-values are plotted in Fig. 2 (right-hand column).

For most L-year trends ending after 2005, model-versus-observed differences in tropospheric warming are significantly larger (at the 10% level or better) and more persistent than can be explained by natural internal variability alone. This result holds for all six satellite TMT datasets examined here. In contrast, L-year difference series trends ending before 2005 are generally not significantly larger than unforced TMT trends in the CMIP5 control runs. Qualitatively similar results are obtained

[4]Other means of estimating unforced variability temperature variability are also available, using either statistical models of observational temperature data [44, 45] or observed residuals after removal of an estimated externally forced temperature component [46, 47].

for TMT averaged over the tropics, as well as for near-global changes in TLT (see Supplementary Figures S1 and S2, respectively).

In each panel in the right-hand column of Fig. 2, there are both upper and lower rejection regions for the null hypothesis stipulated above ("departures between the expected and observed tropospheric temperature trends are consistent with internal climate noise"). The upper rejection regions ($p \geq 0.9$ and ≤ 1) are for significant negative trends in $\Delta T_{f-o}(k,t)$; the lower rejection regions ($p \geq 0$ and ≤ 0.1) are for significant positive trends in the difference series. Under the null hypothesis, significant negative and positive trends in $\Delta T_{f-o}(k,t)$ should be equally likely.[5] This is clearly not the case. Significant positive trends dominate, particularly for trends with start dates after 2000. There is only one small group of significant negative trends in $\Delta T_{f-o}(k,t)$ – the group with end points close to the anomalous warmth of the 1997/98 El Niño.

Other features of Fig. 2 are also of interest. Consider, for example, the group of positive 10-year trends ending between approximately 1990 and 1993 (Fig. 2B). As noted above, El Chichón's cooling signal in satellite TMT data was partly masked by the 1982/83 El Niño; this warm anomaly decreases the size of observed TMT trends starting close to the time of the Chichón eruption. Because the phasing of

[5]Distributions of significant trends generated under the null hypothesis have average values that are in accord with this expectation – *i.e.*, they have the same numbers of significant positive and significant negative trends (see discussion in Sections 4.5 and 4.6 of the Supplementary Material).

ENSO is random in the HIST+8.5 simulations, the Chichón-induced cooling of TMT is larger and clearer in the MMA [1, 42], so simulated TMT trends commencing close to the Chichón eruption tend to be larger than in observations (Figs. 1A and B). The influence of the 1982/83 El Niño diminishes as the trend fitting period is increased.

The large tropospheric warming caused by the 2015/16 El Niño event also has a pronounced effect. As shorter (10- to 13-year) sliding windows sample this observed warming spike, the size of trends in the $\Delta T_{f-o}(k,t)$ difference series decreases, and p-values increase (Figs. 2B, D, F, H). However, as the longer 14- to 15-year analysis sliding windows approach the end of the TMT records, even the anomalous observed warmth of late 2015 and early 2016 does not negate the larger simulated warming during most of the "slowdown" period – *i.e.*, trends in $\Delta T_{f-o}(k,t)$ remain significantly larger than unforced trends (Figs. 2J, L).

Figure 2 reveals large structural uncertainties in satellite TMT datasets. These uncertainties reflect different choices in dataset construction, primarily related to the treatment of orbital drift, the impact of orbital drift on sampling the diurnal cycle of atmospheric temperature [22, 29, 30, 34, 48], and the influence of instrument body temperature [49, 50]. For example, versions 5.6 and 6.0 of the UAH TMT dataset have pronounced differences in tropospheric warming in the first third of the satellite record. These differences (which are probably due to an update in how the UAH group deals with instrument bias correction) are large enough to lead to different decisions

regarding the statistical significance of initial trends in $\Delta T_{f-o}(k,t)$ (compare the UAH 5.6 and 6.0 results in Figs. 2B, D, and F).

Our use of older and newer versions of satellite TMT records highlights the evolutionary nature of these datasets. This evolutionary understanding is not always well understood outside of the scientific community [32], which is why we choose to illustrate it in Fig. 2. In the following analysis, however, we focus on newer dataset versions, which incorporate adjustments for recently identified inhomogeneities, and are likely to be improved relative to earlier dataset versions [22, 29].

The analysis in Fig. 2 focuses on the significance of individual trends in $\Delta T_{f-o}(k,t)$. It does not consider whether overall asymmetries in p-values (such as the preponderance of significant positive trends in the difference series) could be due to internal variability alone. To address this question, we define three asymmetry metrics. The first is the $\gamma_1(k,l)$ statistic, which measures asymmetry in the distribution of significant positive and significant negative trends in $\Delta T_{f-o}(k,t)$. The second and third are the $\gamma_2(k,l)$ and $\gamma_3(k,l)$ statistics, which yield information on asymmetries in the temporal distribution of the individual p-values in Fig. 2.

To assess these asymmetries in temporal distribution, the $N_{f-o}(l)$ maximally overlapping difference series trends in Fig. 2 are divided into two sets of approximately equal size (SET 1 and SET 2). This is done for each value of the trend length L. The difference in the total number of significant positive trends in SET 1 and SET 2 is

$\gamma_2(k,l)$. The difference in "set-average" p-values is $\gamma_3(k,l)$. Further information on all three asymmetry statistics is given in Section 4.5 of the Supplementary Material.

Figure 3 shows asymmetry statistics for the specific case of maximally overlapping 10-year trends in $\Delta T_{f-o}(k,t)$. The actual values of the asymmetry statistics $\gamma_1(k,l)$, $\gamma_2(k,l)$, and $\gamma_3(k,l)$ (shown in Figs. 3A, C, and E, respectively) reflect the above-described features of Fig. 2: a preponderance of significant positive trends in $\Delta T_{f-o}(k,t)$, a larger number of significant positive trends in SET 2 than in SET 1, and a sharp decrease in average p-values between SET 1 and SET 2. We seek to estimate the likelihood that these actual values could be due to internal variability alone.[6] We refer to these probabilities subsequently as $p_{\gamma_1}(k,l)$, $p_{\gamma_2}(k,l)$, and $p_{\gamma_3}(k,l)$.

Our procedure for estimating $p_{\gamma_1}(k,l)$, $p_{\gamma_2}(k,l)$, and $p_{\gamma_3}(k,l)$ involves randomly selecting 5,000 surrogate "observed" TMT time series from the CMIP5 control runs (see Supplementary Material Figs. S3 and S4). Maximally overlapping L-year trends in the surrogate observations are then compared with control run distributions of unforced L-year trends. This yields 5,000-member null distributions of $\gamma_1(l,m)^*$, $\gamma_2(l,m)^*$, and $\gamma_3(l,m)^*$, where we know *a priori* that the statistical properties of the null distributions are solely influenced by internal variability.[7] The actual values

[6]See Section 4.6 of Supplementary Material regarding the credibility of model-based internal variability estimates on multi-decadal timescales.

[7]Where l and m are (respectively) indices over the number of values of the trend length L and the number of surrogate observational time series, and * denotes a statistic calculated with surrogate

of the asymmetry statistics are compared with the Monte Carlo-generated distributions to obtain estimates of $p_{\gamma_1}(k,l)$, $p_{\gamma_2}(k,l)$, and $p_{\gamma_3}(k,l)$ (see Figs. 3B, D, and F). Full details of this significance testing procedure are given in Section 4.5 of the Supplementary Material.

Figure 4 summarizes these probability estimates. By averaging over satellite datasets and analysis timescales, we obtain the overall probabilities $\overline{\overline{p_{\gamma_1}}}$, $\overline{\overline{p_{\gamma_2}}}$, and $\overline{\overline{p_{\gamma_3}}}$ (the magenta horizontal lines in each panel of Fig. 4). For the statistic gauging the asymmetry in the numbers of positive and negative difference series trends, $\overline{\overline{p_{\gamma_1}}} \approx 0.004$. On average, therefore, there is only a 1 in 250 chance that the actual preponderance of significant positive trends in $\Delta T_{f-o}(k,t)$ could be due to internal variability alone (Fig. 4A).

Consider next the temporal asymmetries between SET 1 and SET 2 (Figs. 4B and C). The likelihood is very small ($\overline{\overline{p_{\gamma_2}}} \approx 0.003$) that random internal fluctuations in climate could fully explain why SET 2 has a larger number of significant positive difference series trends in $\Delta T_{f-o}(k,t)$. Finally, the overall likelihood that the actual decline in average p-values between SET 1 and SET 2 is due to random chance alone is $\overline{\overline{p_{\gamma_3}}} \approx 0.07$. Other aspects of these overall significance results, such as the noticeable timescale-dependence of $\overline{\overline{p_{\gamma_3}}}$ in Fig. 4C, are described in detail in Section 4.5 of the Supplementary Material.

---

observations.

The credibility of the estimated p-values in Figs. 2 and 4 is dependent on the reliability of model-based estimates of natural variability. If CMIP5 models systematically underestimated the amplitude of tropospheric temperature variability on 10- to 15-year timescales, it would systematically decrease p-values in Fig. 2, and spuriously inflate the significance of individual difference series trends. In previous work, we found no evidence of such a systematic low bias. On average, CMIP5 models slightly overestimated the amplitude of decadal variability in TMT [51], suggesting that the p-values in Fig. 2 may be conservative.

It is more difficult to assess the credibility of our estimated probabilities for the overall asymmetry statistics shown in Fig. 4. Such an evaluation requires information on model performance in capturing the "real-world" variability of tropospheric temperature on longer 30- to 40-year timescales.[8] This information is not directly available from relatively short satellite TMT records, and must instead be inferred from other sources (see Supplementary Material, Section 4.6). Such indirect sources do not support a systematic model underestimate of tropospheric temperature variability on 30- to 40-year timescales [52].

In summary, we attempted to assess the significance of differences between expected and observed tropospheric temperature trends. The first part of this assessment was performed by fitting L-year trends to $\Delta T_{f-o}(k,t)$, the time series of differ-

[8]These are the timescales on which we are trying to determine the likelihood of obtaining unforced multi-decadal changes in the three asymmetry statistics analyzed here.

ences between tropospheric temperature changes in the CMIP5 multi-model average (MMA) and in individual satellite datasets. We systematically varied the length of L and the start dates of the L-year trends. Each individual trend in $\Delta T_{f-o}(k,t)$ was compared with distributions of unforced L-year temperature trends from model control runs. This yielded an estimate of the probability that an individual trend in $\Delta T_{f-o}(k,t)$ could be due to internal variability alone.

The twin assumptions in our significance assessment are that the MMA is a reasonable estimate of the response to external forcing, and that internal variability estimates from the CMIP5 control runs are reliable. If these assumptions are valid, obtaining significant positive trends in $\Delta T_{f-o}(k,t)$ should be as likely as obtaining significant negative $\Delta T_{f-o}(k,t)$ trends. This is not the case (see Fig. 2). Significant positive trends – indicating model tropospheric warming that is substantially larger than observed – are far more frequent occurrences, and are not randomly distributed as a function of time. Almost all of the significant positive trends in $\Delta T_{f-o}(k,t)$ occur preferentially for L-year trends that sample a substantial portion of the "slowdown" in observed warming in the early 21st century.

In the second part of our analysis, we defined three statistical measures of asymmetry in the p-value results shown in Fig. 2. These measures relate to imbalances in the total numbers of significant positive and negative trends in $\Delta T_{f-o}(k,t)$, to differences in the temporal distribution of significant positive trends, and to decadal changes in

average p-values (see Fig. 3). We addressed the question of whether the actual values of these asymmetry statistics could themselves be due to multi-decadal internal variability. A Monte Carlo analysis suggests that on average,[9] the likelihood is small that the actual multi-decadal differences between expected and observed warming rates are purely random. This likelihood ranges from 1 in 300 to 1 in 15, depending on the choice of asymmetry statistic (see Fig. 4). It is still conceivable, however, that a very unusual observed manifestation of internal variability is the primary driver of our significance results [7].

The most plausible interpretation of these results is that the observed "slowdown" was not a routine manifestation of internal variability. Instead, it reflects the unusual temporal coincidence of multiple cooling influences, arising from the transition to a negative phase of the IPO in roughly 1999 [3, 5, 7, 25], the phasing of other internal variability modes [8, 10], a succession of moderate early 21st century volcanic eruptions [12, 17, 18, 53, 54], an anomalous solar cycle [13], a decrease in stratospheric water vapor [11], and an increase in anthropogenic sulfate forcing [14, 20].

This interpretation is not solely based on our statistical analysis of the time series of differences between simulated and observed tropospheric warming rates. It also relies on physical understanding. This has two components. The first is an understanding of observational changes in the key modes of internal variability, volcanic

[9] Averaging is over three different satellite datasets and six analysis timescales.

aerosols, solar irradiance, water vapor, and anthropogenic aerosols. Changes in these internal and external factors, while subject to some irreducible uncertainty, are real, and have been documented in many of the above-cited studies.

The second component is based on models, and relates to our understanding of the forcings used in the HIST+8.5 simulations, and the phasing of internal variability in individual HIST+8.5 realizations. It is known that these simulations had systematic errors in external forcing during the "slowdown" period (such as the neglect of a sequence of moderate volcanic eruptions after 2000), and therefore underestimated the tropospheric and surface cooling experienced by the real world [1, 11, 12, 14, 18, 53, 54, 23, 55]. These contributions have been evaluated in many model simulations. Likewise, a number of model and observational studies have attempted to quantify the temperature impact of model-versus-observed differences in the phasing of key modes of variability [3, 5, 6, 7, 8, 9, 10].

It has been posited that the differences between modeled and observed tropospheric warming rates are solely attributable to a fundamental error in model sensitivity to anthropogenic greenhouse gas increases [33]. If this explanation were correct, it would undermine confidence in model projections of future climate change. Several aspects of the structure of $\Delta T_{f-o}(k,t)$ cast doubt on the "sensitivity error" explanation. First, it is difficult to understand why significant differences between expected and observed warming rates should be preferentially concentrated in the second part

of the satellite TMT record (see Fig. 2). A fundamental model sensitivity error should be manifest more uniformly in time. Second, a large sensitivity error should appear not only in trend behavior, but also in the response to major volcanic eruptions [42]. After removal of ENSO variability, however, there are no large systematic model errors in tropospheric cooling following the eruptions of El Chichón in 1982 and Pinatubo in 1991 [1]. In summary, the evidence presented here and elsewhere [56, 57, 41] does not support the claim that current climate models are (on average) a factor of 3 to 4 too sensitive to anthropogenic greenhouse gas increases [33].

The temporary "slowdown" in warming has provided the scientific community with a valuable opportunity to advance understanding of internal variability and external forcing, and to develop improved climate observations, forcing estimates, and model simulations. Further work is necessary, particularly to reliably quantify the relative magnitudes of the internally generated and externally forced components of surface and atmospheric temperature change.[10] The science of what has erroneously been called a warming "hiatus" has not yet reached a quietus, and is unlikely to do so in the next several years.

[10]It would be useful to extend the statistical analysis presented here to the lower stratosphere. This is of interest for multiple reasons: 1) the substantial contribution of the cooling stratosphere to TMT [35, 36]; 2) the existence of systematic errors in stratospheric ozone forcing in the CMIP5 HIST+8.5 simulations [23, 58]; and 3) the fact that low-frequency changes in observed lower stratospheric temperature data [59] appear to be in phase with both stratospheric ozone recovery [60] and with the decadal changes in TMT.

Acknowledgments

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison (PCMDI) provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. The views, opinions, and findings contained in this report are those of the authors and should not be construed as a position, policy, or decision of the U.S. Government, the U.S. Department of Energy, or the National Oceanic and Atmospheric Administration.

References

[1] B. D. Santer, C. Bonfils, J. Painter, M. Zelinka, C. Mears, S. Solomon, G. A. Schmidt, J. C. Fyfe, J. N. S. Cole, L. Nazarenko, K. E. Taylor, and F. J. Wentz. Volcanic contribution to decadal changes in tropospheric temperature. *Nat. Geosci.*, 7:185–189, 2014.

[2] J. C. Fyfe, N. P. Gillett, and F. W. Zwiers. Overestimated global warming over the past 20 years. *Nat. Clim. Change*, 3:767–769, 2013.

[3] J. C. Fyfe, G. A. Meehl, M. H. England, M. E. Mann, B. D. Santer, G. M. Flato, E. Hawkins, N. P. Gillett, S.-P. Xie, Y. Kosaka, and N. C. Swart. Making sense of the early-2000s warming slowdown. *Nat. Clim. Change*, 6:224–228, 2016.

[4] D. R. Easterling and M. F. Wehner. Is the climate warming or cooling? *Geophys. Res. Lett.*, 36(L08706), 2009.

[5] G. A. Meehl, H. Teng, and J. M. Arblaster. Climate model simulations of the observed early-2000s hiatus of global warming. *Nat. Clim. Change*, 4:898–902, 2014.

[6] J. S. Risbey, S. Lewandowsky, C. Langlais, D. P. Monselesan, T. J. O'Kane, and N. Oreskes. Well-estimated global surface warming in climate projections selected for ENSO phase. *Nat. Clim. Change*, 4:835–840, 2014.

[7] M. H. England, S. McGregor, P. Spence, G. A. Meehl, A. Timmermamm, W. Cai, A. Sen Gupta, M. J. McPhaden, A. Purich, and A. Santoso. Recent intensification of wind–driven circulation in the Pacific and the ongoing warming hiatus. *Nat. Clim. Change*, 4:222–227, 2014.

[8] B. A. Steinman, M. E. Mann, and S. K. Miller. Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures. *Science*, 347:988–991, 2015.

[9] Y. Kosaka and S.-P. Xie. Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature*, 501:403–407, 2013.

[10] X. Chen and K. K. Tung. Varying planetary heat sink led to global-warming slowdown and acceleration. *Science*, 345:897–903, 2014.

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

[12] S. Solomon, J. S. Daniel, R. R. Neely, J.-P. Vernier, E. G. Dutton, and L. W. Thomason. The persistently variable "background" stratospheric aerosol layer and global climate change. *Science*, 333:866–870, 2011.

[13] G. Kopp and J. L. Lean. A new, lower value of total solar irradiance: Evidence and climate significance. *Geophys. Res. Lett.*, 38, 2011.

[14] G. A. Schmidt, D. T. Shindell, and K. Tsigaridis. Reconciling warming trends. *Nat. Geosci.*, 7:1–3, 2014.

[15] IPCC. Summary for Policymakers. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, page 29. Cambridge University Press, 2013.

[16] G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen. Evaluation of climate models. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 741–866. Cambridge University Press, 2013.

[17] R. R. Neely, O. B. Toon, S. Solomon, J.-P. Vernier, C. Alvarez, J. M. English, K. H. Rosenlof, M. J. Mills, C. G. Bardeen, J. S. Daniel, and J. P. Thayer. Recent anthropogenic increases in SO_2 from Asia have minimal impact on stratospheric aerosol. *Geophys. Res. Lett.*, 40:1–6, 2013.

[18] D. A. Ridley, S. Solomon, J. E. Barnes, V. D. Burlakov, T. Deshler, S. I. Dolgii, A. B. Herber, T. Nagai, R. R. Neely III, A. V. Nevzorov, C. Ritter, T. Sakai, B. D. Santer, M. Sato, A. Schmidt, O. Uchino, and J.-P. Vernier. Total volcanic stratospheric aerosol optical depths and implications for global climate change. *Geophys. Res. Lett.*, 41:7763–7769, 2014.

[19] J.-P. Vernier, L. W. Thomason, J.-P. Pommereau, A. Bourassa, J. Pelon, A. Garnier, A. Hauchecorne, L. Blanot, C. Trepte, D. Degenstein, and F. Vargas. Major influence of tropical volcanic eruptions on the stratospheric aerosol layer during the last decade. *Geophys. Res. Lett.*, 38, 2011.

[20] D. M. Smith, B. B. B. Booth, N. J. Dunstone, R. Eadie, L. Hermanson, G. S. Jones, A. A. Scaife, K. L. Sheen, and V. Thompson. Role of volcanic and anthropogenic aerosols in the recent global surface warming slowdown. *Nat. Clim. Change*, 6:936–940, 2016.

[21] T. R. Karl, A. Arguez, B. Huang, J. H. Lawrimore, J. R. McMahon, M. J. Menne, T. C. Peterson, R. S. Vose, and H.-M. Zhang. Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, 348:1469–1472, 2015.

[22] C. Mears and F. J. Wentz. Sensitivity of satellite-derived tropospheric temperature trends to the diurnal cycle adjustment. *J. Clim.*, 29:3629–3646, 2016.

[23] S. Solomon, P. J. Young, and B. Hassler. Uncertainties in the evolution of stratospheric ozone and implications for recent temperature changes in the tropical

lower stratosphere. *Geophys. Res. Lett.*, 39, 2012.

[24] K. E. Trenberth and J. T. Fasullo. Simulation of present-day and twenty-first-century energy budgets of the Southern Oceans. *J. Clim.*, 23:440–454, 2010.

[25] K. E. Trenberth. Has there been a hiatus? *Science*, 349:791–792, 2015.

[26] K. Cowtan, Z. Hausfather, E. Hawkins, P. Jacobs, M. E. Mann, S. K. Miller, B. A. Steinman, M. B. Stolpe, and R. G. Way. Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophys. Res. Lett.*, 42(15):6526–6534, 2015.

[27] S. Lewandowsky, J. S. Risbey, and N. Oreskes. The "pause" in global warming: Turning a routine fluctuation into a problem for science. *Bull. Amer. Meteor. Soc.*, 97(5):723–733, 2016.

[28] N. Cahill, S. Rahmstorf, and A. C. Parnell. Change points of global temperature. *Environ. Res. Lett.*, 10, 2015.

[29] S. Po-Chedley, T. J. Thorsen, and Q. Fu. Removing diurnal cycle contamination in satellite-derived tropospheric temperatures: Understanding tropical tropospheric trend discrepancies. *J. Clim.*, 28:2274–2290, 2015.

[30] C.-Z. Zou and W. Wang. Inter-satellite calibration of AMSU-A observations for weather and climate applications. *J. Geophys. Res.*, 116, 2011.

[31] K. Cowtan and R. G. Way. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quart. J. Roy. Met. Soc.*, 140:1935–1944, 2014.

[32] US Senate. Data or Dogma? Promoting open inquiry in the debate over the magnitude of human impact on Earth's climate, 2015. [Available online at http://www.commerce.senate.gov/public/index.cfm/2015/12/data-or-dogma-promoting-open-inquiry-in-the-debate-over-the-magnitude -of-human-impact-on-earth-s-climate].

[33] J. R. Christy. Testimony in Hearing before the U.S. Senate Committee on Commerce, Science, and Transportation, Subcommittee on Space, Science, and Competitiveness, December 8, 2015, 2015. [Available online at http://www.commerce.senate.gov/public/index.cfm/2015/12/data-or-dogma-promoting-open-inquiry-in-the-debate-over-the-magnitude-of-human-impact-on-earth-s-climate].

[34] J. R. Christy, W. B. Norris, R. W. Spencer, and J. J. Hnilo. Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements. *J. Geophys. Res.*, 112, 2007.

[35] Q. Fu, C. M. Johanson, S. G. Warren, and D. J. Seidel. Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature*, 429:55–58, 2004.

[36] Q. Fu and C. M. Johanson. Stratospheric influences on MSU-derived tropospheric temperature trends: A direct error analysis. *J. Clim.*, 17:4636–4640, 2004.

[37] Q. Fu, S. Manabe, and C. M. Johanson. On the warming in the tropical upper troposphere: Models versus observations. *Geophys. Res. Lett.*, 38, 2011.

[38] S. Po-Chedley and Q. Fu. Discrepancies in tropical upper tropospheric warming between atmospheric circulation models and satellites. *Environ. Res. Lett.*, 7, 2012.

[39] K. E. Taylor, R. J. Stouffer, and G. A. Meehl. An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, 93:485–498, 2012.

[40] B. D. Santer, C. Mears, C. Doutriaux, P. Caldwell, P. J. Gleckler, T. M. L. Wigley, S. Solomon, N. P. Gillett, D. Ivanova, T. R. Karl, J. R. Lanzante, G. A. Meehl, P. A. Stott, K.E. Taylor, P. W. Thorne, M. F. Wehner, and F. J. Wentz. Separating signal and noise in atmospheric temperature changes: The importance of timescale. *J. Geophys. Res.*, 116, 2011.

[41] B. D. Santer, S. Solomon, G. Pallotta, C. Mears, S. Po-Chedley, Q. Fu, F. Wentz, C.-Z. Zou, J. Painter, I. Cvijanovic, and C. Bonfils. Comparing tropospheric warming in climate models and satellite data. *J. Clim.*, 30:373–392, 2017.

[42] T. M. L. Wigley, C. M. Ammann, B. D. Santer, and S. C. B. Raper. The effect of climate sensitivity on the response to volcanic forcing. *J. Geophys. Res.*, 110, 2005.

[43] D. J. A. Johansson, B. C. O'Neill, C. Tebaldi, and O. Häggström. Equilibrium climate sensitivity in light of observations over the warming hiatus. *Nat. Clim. Change*, 5:449–453, 2015.

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

[45] P. T. Brown, W. Li, E. C. Cordero, and S. A. Mauget. Comparing the model-simulated global warming signal to observations using empirical estimates of unforced noise. *Nature Sci. Rep.*, 5(9957), 2016.

[46] M. R. Allen and S. F. B. Tett. Checking for model consistency in optimal fingerprinting. *Cli. Dyn.*, 15:419–434, 1999.

[47] M. E. Mann, S. Rahmstorf, B. A. Steinman, M. Tingley, and S. K. Miller. The likelihood of recent warmth. *Nat. Sci. Rep.*, 6:19831, 2016.

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

[49] C. A. Mears, M. C. Schabel, and F. J. Wentz. A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Clim.*, 16:3650–3664, 2003.

[50] S. Po-Chedley and Q. Fu. A bias in the mid-tropospheric channel warm target factor on the NOAA-9 Microwave Sounding Unit. *J. Atmos. Oceanic Technol.*, 29:646–652, 2012.

[51] B. D. Santer, J. F. Painter, C. A. Mears, C. Doutriaux, P. Caldwell, J. M. Arblaster, P. J. Cameron-Smith, N. P. Gillett, P. J. Gleckler, J. Lanzante, J. Perlwitz, S. Solomon, P. A. Stott, K. E. Taylor, L. Terray, P. W. Thorne, M. F. Wehner, F. J. Wentz, T. M. L. Wigley, L. J. Wilcox, and C.-Z. Zou. Identifying human influences on atmospheric temperature. *Proc. Nat. Acad. Sci.*, 110:26–33, 2013.

[52] J. Imbers, A. Lopez, C. Huntingford, and M. R. Allen. Testing the robustness of anthropogenic climate change detection statements using different empirical models. *J. Geophys. Res.*, 118:3192–3199, 2013.

[53] J. C. Fyfe, K. von Salzen, J. N. S. Cole, N. P. Gillett, and J.-P. Vernier. Surface response to stratospheric aerosol changes in a coupled atmosphere-ocean model. *Geophys. Res. Lett.*, 40:584–588, 2013.

[54] J. M. Haywood, A. Jones, and G. S. Jones. The impact of volcanic eruptions in the period 2000-2013 on global mean temperature trends evaluated in the HadGEM2-ES climate model. *Atmos. Sci. Lett.*, 15:92–96, 2013.

[55] B. Hassler, P. J. Young, R. W. Portmann, G. E. Bodeker, J. S. Daniel, K. H. Rosenlof, and S. Solomon. Comparison of three vertically resolved ozone data sets: climatology, trends and radiative forcings. *Atmos. Chem. Phys.*, 13:5533–5550, 2013.

[56] M. Huber and R. Knutti. Natural variability, radiative forcing and climate response in the recent hiatus reconciled. *Nat. Geosci.*, 7:651–656, 2014.

[57] J. Marotzke and P. M. Forster. Forcing, feedback and internal variability in global temperature trends. *Nature*, 517:565–570, 2015.

[58] V. Eyring, J. M. Arblaster, I. Cionni, J. Sedláček, J. Perlwitz, P. J. Young, S. Bekki, D. Bergmann, P. Cameron-Smith, W. J. Collins, G. Faluvegi, K.-D. Gottschaldt, L. W. Horowitz, D. E. Kinnison, J.-F. Lamarque, D. R. Marsh, D. Saint-Martin, D. T. Shindell, K. Sudo, S. Szopa, and S. Watanabe. Long-term ozone changes ozone and associated climate impacts in CMIP5 simulations. *J. Geophys. Res.*, 118:5029–5060, 2013.

[59] C.-Z. Zou and H. Qian. Stratospheric temperature climate record from merged SSU and AMSU-A observations. *J. Atmos. Ocean. Tech.*, 33:1967–1984, 2016.

[60] S. Solomon, D. J. Ivy, D. Kinnison, M. J. Mills, R. R. Neely III, and A. Schmidt. Emergence of healing in the Antarctic ozone layer. *Science*, 353:269–274, 2016.

Figure 1: Time series (panel A) and difference series (panel B) of monthly-mean near-global anomalies in simulated and observed tropospheric temperature. Results are for the temperature of the mid- to upper troposphere (TMT), corrected for lower stratospheric cooling [35] and spatially averaged over 82.5°N-82.5°S. Multi-model average (MMA) temperature data are from simulations of externally forced climate change performed with 37 different CMIP5 models; satellite TMT data are for RSS version 4.0 only [22]. Model TMT data were computed using vertical weighting functions that approximate the satellite-based vertical sampling of the atmosphere [51]. The time series of differences between the MMA and the RSS data is shown in both raw form and smoothed with a 12-month running mean (panel B). All anomalies are relative to climatological monthly means calculated over January 1979 to December 2015. Observational (model) records extend to June 2016 (December 2016). The vertical purple line is plotted at the time of the maximum global-mean tropospheric warming during the 1997/98 El Niño. The vertical green lines denote the eruption dates of El Chichón and Pinatubo. See the Supplementary Material for further details of the model and observational datasets and the regression-based method used to correct TMT.

Figure 2: Trends (left column) in near-global averages of TMT difference series. The six difference series were computed by subtracting each of the six individual satellite TMT records from the HIST+8.5 multi-model average TMT time series (see Fig. 1). Maximally overlapping trends were fit to each difference series. Results are for trend

lengths of $L = 10, 11, \ldots, 15$ years; the overlap between successive L-year trends is by all but one month. The p-values associated with each L-year difference series trend (right column) were obtained by testing against multi-model distributions of unforced L-year TMT trends from 36 different CMIP5 control runs. The final month of each satellite TMT time series is June 2016. Results are plotted on the last month of the trend-fitting period. The grey shading denotes the rejection region (at a stipulated 10% significance level) for the null hypothesis that the difference between expected and observed TMT trends could be due to internal variability alone. Each panel in the right-hand column has two rejection regions: the lower (upper) region is for large positive (large negative) trends in the model-minus-observed difference series. The lower (upper) rejection region spans the p-value range 0 to 0.1 (0.9 to 1.0). The y-axis range was extended to -0.06 to facilitate visual display of p-values at or close to zero. To calculate the actual values of the $\gamma_2(k, l)$ and $\gamma_3(k, l)$ statistics in Figs. 3D and F, the maximally overlapping L-year trends were divided into two sets of approximately equal size ("SET 1" and "SET 2"; see Supplementary Material, Section 4.4). The dashed vertical lines in the right-hand column panels denote the final month of the last L-year trend in SET 1.

Figure 3: Asymmetries in the statistical significance of differences between individual expected and observed trends in near-global averages of corrected TMT. All results are for maximally overlapping 10-year trends that are calculated from the $\Delta T_{f-o}(k, t)$ difference time series. The total number of significant trends in $\Delta T_{f-o}(k, t)$ is strongly

asymmetric, with more trends that are significantly positive than trends that are significantly negative (panel A). Trend significance is assessed at the 10% level relative to model-based estimates of natural internal variability on the 10-year timescale. Actual values of the γ_1 statistic (the vertical lines in panel B) are the differences between the numbers of significant positive and negative trends in panel A. Significant trends in $\Delta T_{f-o}(k,t)$ also show a pronounced asymmetry in the temporal distribution of positive trends (panel C). To quantify this asymmetry and evaluate whether it is unusual, we split the total number of maximally overlapping 10-year trends in $\Delta T_{f-o}(k,t)$ into two equally sized sets (SET 1 and SET 2). The number of positive trends achieving significance at the 10% level or better is consistently larger in SET 2. The vertical lines in panel D are the actual values of the γ_2 statistic, which is the difference between the SET 1 and SET 2 bars in panel C. The average p-values in SET 1 and SET 2 are shown in panel E; the difference between these set-average values is the γ_3 statistic (see vertical lines in panel F). The grey histograms in panels B, D, and F were generated using 5,000 realizations of surrogate observations, randomly selected from the 36 CMIP5 control runs (see Supplementary Material, Section 4.5). The distributions provide information on whether the actual values of the statistics (the vertical lines) could have been obtained by natural variability alone.

Figure 4: Overall statistical significance of the asymmetry statistics, $\gamma_1(k,l)$, $\gamma_2(k,l)$, and $\gamma_3(k,l)$, as a function of the selected observational dataset used to compute the $\Delta T_{f-o}(k,t)$ difference time series and the selected analysis timescale. For definition

of the asymmetry statistics, refer to the caption of Fig. 3 and the Supplementary Material, Section 4.4. Results in Fig. 4 are estimates of $p_{\gamma_1}(k,l)$, $p_{\gamma_2}(k,l)$, and $p_{\gamma_3}(k,l)$, the probabilities that each actual value of the asymmetry statistic could have been obtained by natural internal variability alone (panels A, B, and C, respectively). The magenta lines in panels 4A, B, and C are values of $\overline{\overline{p_{\gamma_1}}}$, $\overline{\overline{p_{\gamma_2}}}$, and $\overline{\overline{p_{\gamma_3}}}$. These are the averages (over the three recent observational datasets and the six analysis timescales) of $p_{\gamma_1}(k,l)$, $p_{\gamma_2}(k,l)$, and $p_{\gamma_3}(k,l)$. Zero values of the probabilities are indicated by colored arrows. Note that the y-axis range in Figs. 4A and B is an order of magnitude smaller than in Fig. 4C.

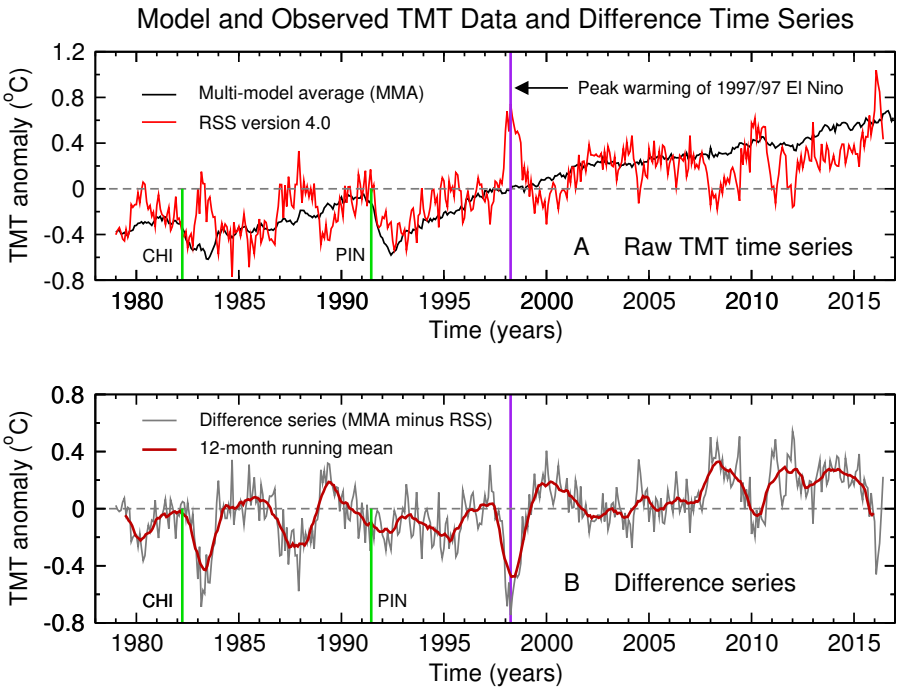

Figure 1: Santer *et al.*

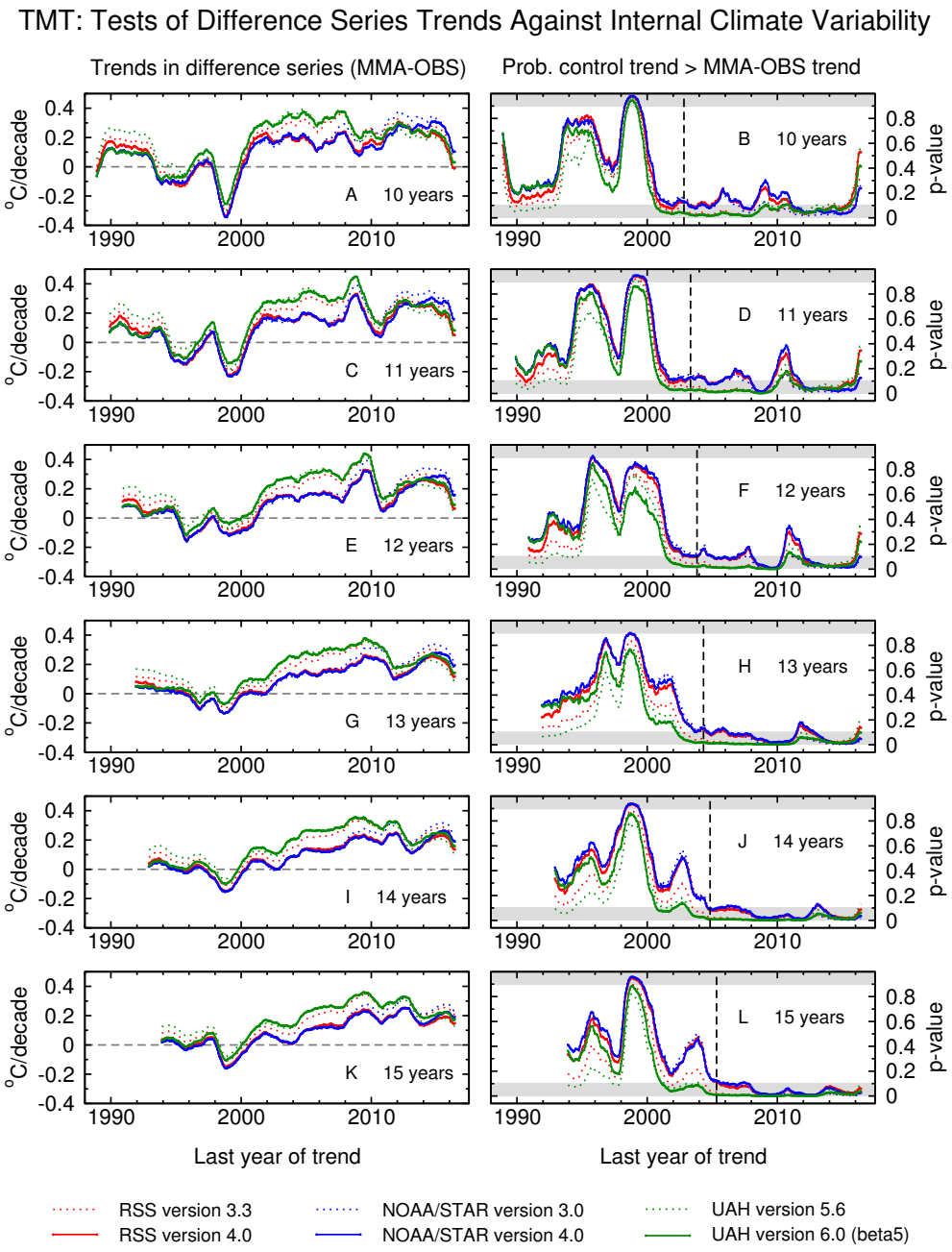

Figure 2: Santer *et al.*

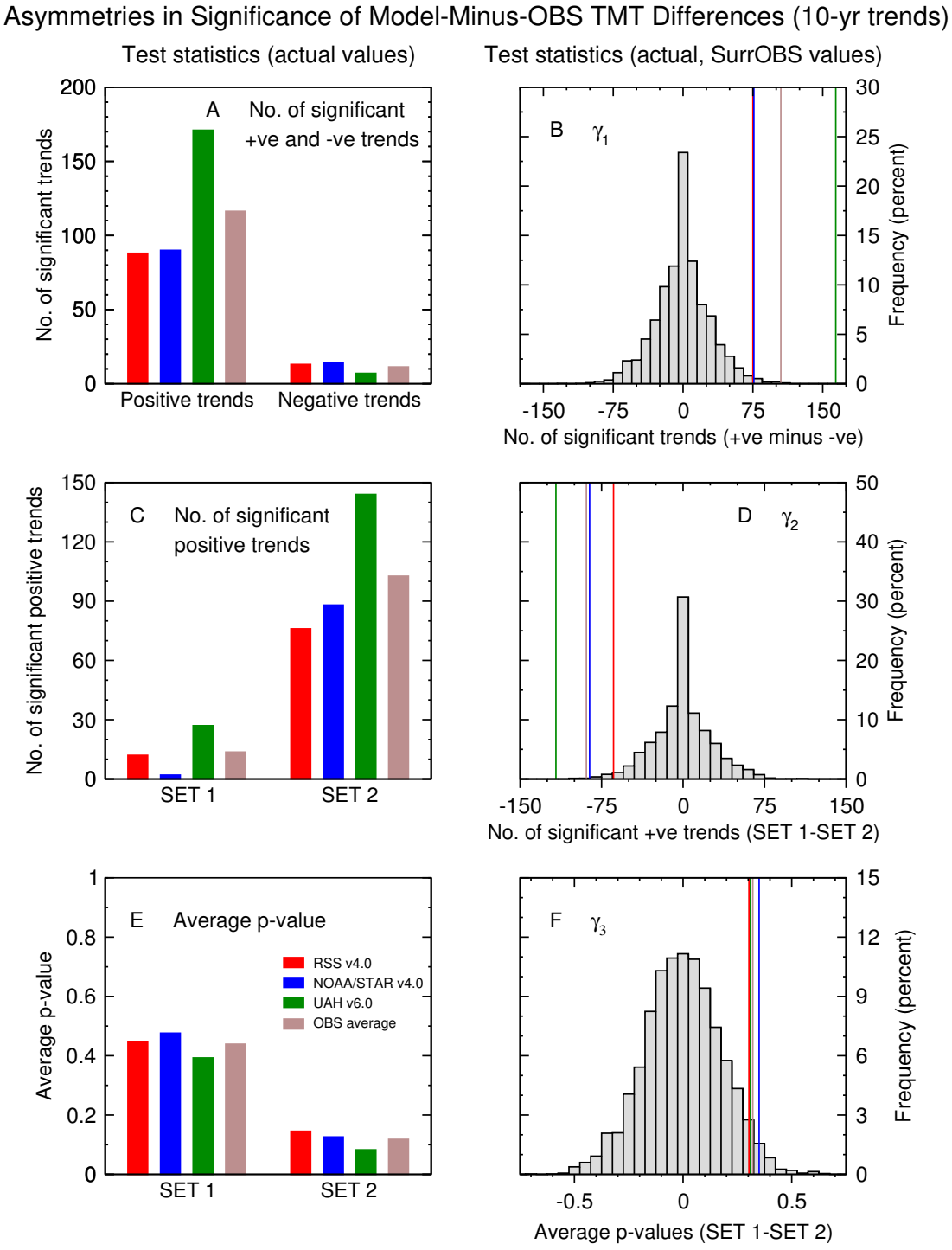

Figure 3: Santer *et al.*

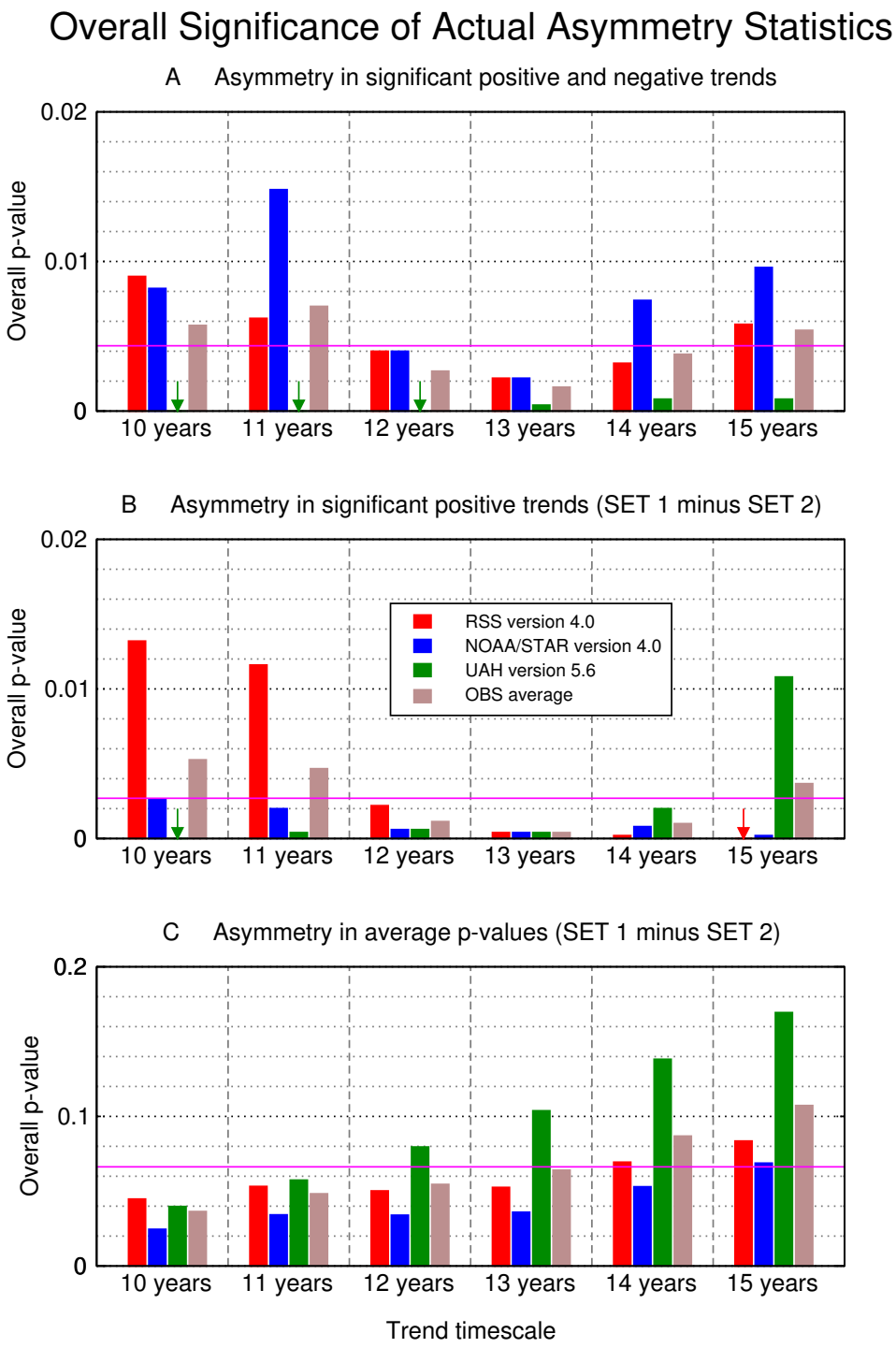

Figure 4: Santer *et al.*