Using climate model simulations to constrain observations

Benjamin D. Santer[a,*], Stephen Po-Chedley[a], Carl Mears[b], John C. Fyfe[c], Nathan Gillett[c], Qiang Fu[d], Jeffrey F. Painter[a], Susan Solomon[e], Andrea K. Steiner[f], Frank J. Wentz[b], Mark D. Zelinka[a], and Cheng-Zhi Zou[g]

[a]*Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA.*

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

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

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

[e]*Massachusetts Institute of Technology, Earth, Atmospheric, and Planetary Sciences, Cambridge, MA 02139, USA.*

[f]*Wegener Center for Climate and Global Change, University of Graz, A-8010 Graz, Austria.*

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

[*]*Corresponding author*: santer1@llnl.gov. *Date*: March 21, 2021

ABSTRACT

We compare atmospheric temperature changes in satellite data and in older and newer multi-model and single-model ensembles performed under phases 5 and 6 of the Coupled Model Intercomparison Project (CMIP5 and CMIP6). In the lower stratosphere, multi-decadal stratospheric cooling during the period of strong ozone depletion is smaller in newer CMIP6 simulations than in CMIP5 or satellite data. In the troposphere, however, despite differences in the forcings and climate sensitivity of the CMIP5 and CMIP6 ensembles, their ensemble-average global warming over the satellite era is remarkably similar. We also examine four well-understood properties of tropical behavior governed by basic physical processes. The first three properties are ratios between trends in water vapor (WV) and trends in sea surface temperature (SST), the temperature of the lower troposphere (TLT), and the temperature of the mid- to upper troposphere (TMT). The fourth property is the ratio between TMT and SST trends. All four trend ratios are tightly constrained in CMIP simulations. Observed ratios diverge markedly when calculated with SST, TLT, and TMT trends produced by different groups. Observed data sets with larger warming of the tropical ocean surface and tropical troposphere yield atmospheric moistening that is closer to model results. For the TMT/SST ratio, model-data consistency depends on the selected combination of observed data sets used to estimate TMT and SST trends. If model expectations of these four covariance relationships are realistic, one interpretation of our findings is that they reflect a systematic low bias in satellite tropospheric temperature trends. Alternately, the observed atmospheric moistening signal may be overestimated. Given the large structural uncertainties in observed tropical TMT and SST trends, and because satellite WV data are available from one group only, it is difficult to determine which interpretation is more credible. Nevertheless, our analysis illustrates the diagnostic power of simultaneously considering multiple complementary variables and points towards possible problems with certain observed data sets.

1. Introduction

Since publication of the first assessment report of the Intergovernmental Panel on Climate Change (IPCC) in 1990, there have been major improvements in our ability to model the climate system (Randall et al. 2007; Trenberth et al. 2007; Flato et al. 2013; Hartmann et al. 2013). Thirty years ago, the climate science community performed single simulations with a small number of pioneering atmosphere-ocean models. Today, more complex Earth System Models (ESMs) are used to generate large multi-model and single-model ensembles of simulations (Kay et al. 2015; Fyfe et al. 2017; Eyring et al. 2019; Deser et al. 2020). Standard benchmark simulations, performed repeatedly with improved versions of uncoupled and coupled models, have over the last several decades exposed and in some cases reduced systematic errors in model representation of many different aspects of Earth's climate (Gates et al. 1999; Randall et al. 2007; Flato et al. 2013; Sperber et al. 2013; Bellenger et al. 2014).

In tandem with advances in modeling, there have been improvements in the forcings used in model simulations of historical climate change (Solomon et al. 2011; Fyfe et al. 2013; Schmidt et al. 2014; Checa-Garcia et al. 2018). Observations have also improved with advances in the ability of scientists to identify and adjust for non-climatic effects (Wentz and Schabel 1998; Mears et al. 2003; Fu and Johanson 2005; Mears and Wentz 2005; Karl et al. 2006, 2015; Po-Chedley et al. 2015). This evolution of models, forcings, and observations is ongoing.

The last IPCC assessment report, published in 2013, relied on CMIP5 simulations performed with roughly four dozen models (Taylor et al. 2012). The 2021 IPCC assessment will evaluate output from a larger collection of CMIP6 models and an expanded set of experiments (Eyring et al. 2016, 2019). Our interest here is in comparing atmospheric temperature changes in CMIP5, CMIP6, the latest satellite data (Mears and Wentz 2017; Zou and Wang 2011; Spencer et al. 2017), and a state-of-the-art reanalysis of weather observations with a weather forecast model (Simmons et al. 2020). We seek to determine whether: 1) there are important differences between atmospheric temperature changes in CMIP5 and CMIP6; and 2) models and observations show consistency in theoretically and physically based constraints on tropical behavior – the amplification of tropical warming with increasing height, and the ratios between trends in tropical water vapor and trends in temperature at different levels. We show that the combination of these constraints provides new information on model/data consistency.

There are several reasons for our focus on atmospheric temperature. First, discrepancies between modeled and observed atmospheric temperature changes have received scientific and political attention for over 20 years (NRC 2000; Karl et al. 2006; Thorne et al. 2011; Fu et al. 2011; Po-Chedley and Fu 2012; US Senate 2015; Santer et al. 2017a,b; Po-Chedley et al. 2021). Determining the causes of these differences remains a priority. Second, estimates of atmospheric temperature from satellites have recently undergone important revision, primarily due to improved understanding of the effects of drifts in satellite orbits and instrument calibration (Po-Chedley et al. 2015; Mears and Wentz 2016, 2017; Zou and Qian 2016; Zou et al. 2018; Spencer et al. 2017). Reanalysis models and data assimilation systems have also evolved (Hersbach et al. 2020; Simmons et al. 2020). Our goal is to reassess model-data consistency in the light of these improvements to observations, models, and external forcings.

The structure of our paper is as follows. Sections 2 and 3 introduce the observational and model data analyzed in our study. Section 4 discusses basic features of atmospheric temperature time series and trends. Trend comparisons are over the full satellite era and over periods of stratospheric ozone depletion and recovery. Section 5 examines the relative sizes of forced and unforced temperature changes on different timescales, and considers whether observed changes are consistent with results from the forced simulations. The statistical methodology in Section 5 follows Santer et al. (2011)

and is provided in the Supplementary Materials (SM) with only minor modifications. Section 6 focuses on the covariability of different aspects of tropical climate change. We examine ratios between tropical trends in column-integrated water vapor (WV) and sea surface temperature (SST), WV and the temperature of the lower troposphere (TLT), WV and the temperature of the mid- to upper troposphere (TMT), and TMT and SST. These four ratios are compared in observations and multi-model and single-model ensembles. Prospects for using such covariability information to constrain divergent observations are considered in Section 7. Appendices A and B provide information regarding the calculation of synthetic satellite temperatures and the adjustment of tropospheric temperature for stratospheric cooling influence.

2. Observational data

a. Satellite temperature data

Since late 1978, NOAA polar-orbiting satellites have monitored the microwave emissions from oxygen molecules using the Microwave Sounding Unit (MSU) and the Advanced Microwave Sounding Unit (AMSU; Mears and Wentz 2017; Spencer et al. 2017; Zou et al. 2018). Microwave emissions are proportional to the temperature of broad atmospheric layers. By measuring at different microwave frequencies, MSU and AMSU provide estimates of TLT, TMT, and the temperature of the lower stratosphere (TLS).

We analyze TLS and TMT data sets produced by RSS (Mears and Wentz 2016), STAR (Zou and Qian 2016), and UAH (Spencer et al. 2017). Only RSS and UAH supply TLT measurements. We rely on the most recent data set versions: RSS 4.0, STAR 4.1, and UAH 6.0. The University of Washington (UW) also produces a TMT data set, but this is available for the tropics only (Po-Chedley et al. 2015). We did not use UW TMT data for the present study.

We consider three different versions of the RSS atmospheric temperature data. As noted in Mears and Wentz (2017), “a total of nine MSU instruments cover the period from 1978 to 2005, followed by a series of AMSU instruments that began in mid-1998 and continue to the present”. MSU and AMSU do not measure at the same microwave frequencies; different plausible choices can be made in merging their estimated brightness temperatures.

Mears and Wentz (2016) employed three approaches to merge MSU and AMSU data:

1. MSU and AMSU measurements were used during the merge period from mid-1998 to 2003.

2. Only AMSU data were used after 1999. MSU data were excluded after 1999.

3. MSU data were used after 1999. AMSU data were excluded before 2003.

These approaches are referred to subsequently as “baseline”, “AMSU merge”, and “MSU merge” (respectively), and are described in more detail in the SM. In Sections 5 and 6, we address the question of whether these three RSS data sets yield different statistical inferences regarding the correspondence between simulated and observed measures of climate change.

All satellite temperature data sets analyzed here are in the form of monthly means on the same 2.5°× 2.5° latitude/longitude grid. Near-global averages of TLS, TMT and TLT were calculated over areas of common coverage in the RSS, UAH, and STAR datasets (82.5°N to 82.5°S for TLS and TMT, and 82.5°N to 70°S for TLT). At the time this analysis was performed, satellite temperature data for full 12-month years were available for the 492-month period from January 1979 to December 2019.

b. SST data

Section 6 considers two ratio statistics involving SST. The first is $R_{\{\mathrm{WV/SST}\}}$, the ratio between tropical trends in WV and SST (Wentz and Schabel 2000; Held and Soden 2006; Mears et al.

2007; Mears and Wentz 2016). The second is $R_{\{\mathrm{TMT/SST}\}}$, the ratio of tropical TMT and SST trends (Wentz and Schabel 2000; Santer et al. 2005; Po-Chedley et al. 2015). We seek to determine whether simulated and observed values of these ratio statistics are consistent, and how model-data agreement is affected by structural uncertainty in observed SST data. This uncertainty arises from differences in raw data, the methods used to adjust raw data for known inhomogeneities, treatment of sea ice, and the decisions made in merging information from ship-based measurements, buoys, floats, and satellites (Karl et al. 2006, 2015; Morice et al. 2012; Hausfather et al. 2017). We quantify structural uncertainty in SST data by calculating $R_{\{\mathrm{WV/SST}\}}$ and $R_{\{\mathrm{TMT/SST}\}}$ with four commonly used observational records (Po-Chedley et al. 2021):

1. Version 2 of the Centennial In Situ Observation-Based Estimates of the Variability of SST and Marine Meteorological Variables (COBE; Hirahara et al. 2014).
2. Version 5 of the NOAA Extended Reconstructed SST data set (ERSST; Huang et al. 2017).
3. Version 1 of the Hadley Center Sea Ice and SST data set (HadISST; Rayner et al. 2003).
4. Version 3 of the Hadley Center SST data set (HadSST; Kennedy et al. 2011).

All data sets except HadSST are spatially complete over the ocean domain of interest (20°N-20°S).

c. Satellite water vapor data

The satellite WV data used here were produced by RSS and are from 11 different satellite-based microwave radiometers (Wentz 2013). The procedures for intercalibrating and merging information from these instruments and for estimating uncertainties in satellite WV trends are described in detail elsewhere (Mears et al. 2018). The WV retrievals are based on measurements of microwave emissions from the 22-GHz water vapor absorption line. The distinctive shape of this line provides robust retrievals that are less problematic than other types of satellite measurement.

The signal-to-noise ratio (S/N) for detecting moistening in the lower troposphere by a measurement of water vapor is several times larger than for MSU-based measurements of air temperature (Wentz and Schabel 2000). Relative to WV information from radiosondes and early reanalysis products, the RSS WV data set was judged by Trenberth et al. (2005) to provide the most credible estimate of means, variability, and trends over oceans.

RSS WV data were available for the 384 months from January 1988 to December 2019 on a 1°× 1° latitude/longitude grid. Due to the high emissivity of the land surface, WV retrievals are provided over oceans only. Our focus here is on WV trends spatially averaged over tropical oceans (20°N-20°S), where there is well-understood covariability between temperature and saturation vapor pressure (Iribarne and Godson 1981).

Because of changes in satellite capabilities, footprint size, and rain and land masking, the spatial coverage of the RSS WV data changes over time. This results in the systematic addition of grid cells with WV data in the western Pacific and near the maritime continent. To avoid the introduction of trend biases arising from coverage changes, we imposed a "fixed coverage" mask – i.e., our analysis of the satellite WV data was restricted to the subset of grid-points with continuous coverage over the 384-month analysis period. After regridding model WV data to the observational grid, the same "fixed coverage" mask was applied to all model simulations of historical climate change.

d. Reanalysis data

Reanalyses employ an atmospheric numerical weather forecast model with no changes over time in the model itself (Bengtsson and Shukla 1988; Kalnay et al. 1996). They provide a well-tested framework for blending and constraining assimilated weather information from different sources; each source is typically characterized by different accuracy and different temporal and spatial coverage.

The ERA5 reanalysis product of the European Centre for Medium-Range Weather Forecasts (ECMWF) recently superseded the ERA-Interim reanalysis. ERA5 was generated with a high-resolution version ($\approx$ 31 km horizontal resolution, 137 vertical levels) of the ECMWF operational forecast model and a 4-D variational data assimilation system (Hersbach et al. 2020). According to Simmons et al. (2020), ERA5 exhibited "a pronounced cold bias for the years 2000 to 2006".

ERA5.1, which spans the affected 2000 to 2006 period, corrects this error and yields "analyses with better global-mean temperatures in the stratosphere and uppermost troposphere than provided by ERA5" (Simmons et al. 2020). Inclusion of ERA5.1 results allows us to test whether blending model and observational information in a state-of-the-art reanalysis framework provides layer-average atmospheric temperature information similar to those available from actual RSS, STAR, and UAH satellite data.

3. Model output

a. CMIP5 simulations

We used model output from phase 5 of the Coupled Model Intercomparison Project (CMIP5) (Taylor et al. 2012). The description of the CMIP5 data sets provided in the next two paragraphs follows Santer et al. (2017a).

Our focus here is on three different types of CMIP5 numerical experiment: 1) simulations with estimated historical changes in human and natural external forcings; 2) simulations with 21st century changes in greenhouse gases and anthropogenic aerosols prescribed according to the Representative Concentration Pathway 8.5 (RCP8.5; Meinshausen et al. 2011);[i] and 3) pre-industrial control runs with no changes in external influences on climate.

[i]RCP8.5 has radiative forcing of approximately 8.5 W/m^2 in 2100, eventually stabilizing at roughly 12 W/m^2.

Most CMIP5 historical simulations end in December 2005. RCP8.5 simulations were initiated from conditions of the climate system at the end of the historical run. To avoid truncating comparisons between modeled and observed climate change trends in December 2005, we spliced together output from the historical simulations and the RCP8.5 runs. We refer to these spliced simulations subsequently as “extended HIST” runs.

In total, we analyzed 123 individual extended HIST realizations performed with 28 different CMIP5 models. We excluded models that did not consider the scattering and absorption of radiation by stratospheric volcanic aerosols (Santer et al. 2013), and therefore lack short-term lower stratospheric warming signals after the eruptions of El Chichón in 1982 and Pinatubo in 1991. Including these models in the calculation of multi-model average (MMA) temperature changes would bias the MMA estimate of volcanic TLS signals.

Details of the start dates, end dates, and lengths of the historical integrations and RCP8.5 runs are given in Supplemental Table S1. Supplemental Table S2 provides information on the 36 CMIP5 pre-industrial control runs used to calculate climate noise estimates. The control integrations allow us to determine S/N characteristics of atmospheric temperature changes (see Section 5).

b. CMIP6 simulations

We also analyze TLS, TMT, TLT, WV, and SST from model simulations performed under phase 6 of CMIP. These simulations rely on newer versions of CMIP5 models, often with more comprehensive representation of earth system processes (Eyring et al. 2016), and with contributions from modeling groups that did not participate in CMIP5. Efforts were made in CMIP6 to improve the representation of external forcings with known systematic errors in CMIP5, such as volcanic and solar forcing in the early 21st century (Solomon et al. 2011; Kopp and Lean 2011; Ridley et al. 2014; Schmidt et al. 2014; Gillett et al. 2016).

At the time this research was performed, the CMIP6 archive was still being populated with model simulation output. For pre-industrial control runs, output was available from 30 different models. For the analysis of forced simulations, the CMIP6 historical runs[ii] from 22 different models were spliced with results from scenario integrations.

Multiple Shared Socioeconomic Pathway (SSP) scenarios were available for splicing (Riahi et al. 2017). We chose the SSP5 scenario here. SSP5 most closely approximates the radiative forcing in the CMIP5 RCP8.5 simulation. The differences in radiative forcing between the five SSPs are very small over the satellite era (Riahi et al. 2017), so the choice of scenario is unlikely to affect our model-versus-data comparisons.

In the case of TMT, TLT, SST, and WV, we analyzed 166 realizations. For reasons discussed in Section 3c, the sample size was smaller for TLS (116 extended HIST realizations performed with 21 models). Further details of the CMIP6 extended HIST and control simulations are provided in Supplemental Tables S3 and S4, respectively.

c. Large initial condition ensembles

Large initial condition ensembles (LEs) are routinely performed by climate modeling groups (Deser et al. 2012; Fyfe et al. 2017; Deser et al. 2020). Typical LE sizes range from 30 to 100. Individual LE members are generated with the same model and external forcings, but are initialized from different conditions of the climate system. Each LE member provides a unique realization of the "noise" of natural internal variability superimposed on the underlying climate "signal" (the response to the changes in forcing).

We used four different LEs to quantify uncertainties in temperature and WV trends arising from multi-decadal internal variability. Two LEs applied CMIP5 historical forcing until 2005 and CMIP

[ii] The CMIP6 historical runs typically end in December 2014.

RPC8.5 forcing thereafter. The other two LEs relied on CMIP6 forcing until 2014 and SSP5 forcing from 2015 to 2100. The CMIP5 LEs were performed with version 1 of the Community Earth System Model (CESM1; Deser et al. 2012) and with version 2 of the Canadian Earth System Model (CanESM2; Fyfe et al. 2017; Swart et al. 2018). The CESM1 and CanESM2 LEs consist of 40 and 50 members, respectively. The two 50-member CMIP6 LEs relied on version 5 of CanESM (CanESM5; Fyfe et al. 2021) and on version 6 of the Model for Interdisciplinary Research on Climate (MIROC6; Tatebe et al. 2019). All four LEs used different strategies for initialization of the individual ensemble members.[iii]

The CanESM5 LE exhibits anomalous aperiodic 1-2 month lower stratospheric warming events in certain ensemble members. These warming events are sufficiently large to influence decadal-timescale TLS trends but have minimal impact on decadal variability in tropospheric temperature. We therefore excluded the CanESM5 LE from the multi-model analysis of CMIP6 TLS trends, but included CanESM5 LE results in the multi-model analysis of TMT, TLT, WV, and SST.

4. Temperature time series and trends

a. Lower stratosphere

Figure 1A shows time series of near-global averages of TLS. The lower stratosphere cools over the full satellite era in all observational data sets and model extended HIST simulations. The main cause of this cooling is human-induced depletion of stratospheric ozone, with a smaller contribution from anthropogenic increases in atmospheric CO_2 (Solomon 1999; Ramaswamy et al. 2006; Thompson et al. 2012; Aquila et al. 2016; Maycock et al. 2018; Bandoro et al. 2018). Satellite-era decreases in TLS are punctuated by large episodic warming signals after the major eruptions of El Chichón

[iii]Differences include the selected starting year for the simulation, the strategy for perturbing initial conditions, and whether perturbations were applied to the atmosphere only or to the atmosphere and the ocean.

in 1982 and Pinatubo in 1991. Warming arises from absorption of incoming solar radiation and outgoing long-wave radiation by stratospheric volcanic aerosols (Robock 2000; Shine et al. 2003).

The CMIP6 multi-model average has an unrealistically small TLS signal after El Chichón (Fig. 2). Based on the MMA root-mean-square (RMS) errors between observed and simulated volcanic TLS signals, the TLS response to El Chichón is better captured in CMIP5 (Figs. 3A,C). For Pinatubo, the MMA RMS error is smaller in CMIP6 (Figs. 3B,D). These CMIP5-versus-CMIP6 differences are significant at the 5% level for the El Chichón signal, but not for the Pinatubo signal (see SM).

Volcanic signal differences in CMIP5 and CMIP6 arise from multiple factors. These include differences in the type and time history of information used for prescribing historical changes in volcanic aerosol loadings, the aerosol optical properties, and the implementation of these properties in calculating volcanic radiative forcing (Thomason et al. 2018). Rather than prescribing volcanic aerosol, at least one CMIP6 modeling group calculated volcanic aerosol loadings based on observed estimates of volcanically produced SO_2 (Mills et al. 2016; Danabasoglu et al. 2020). Separating and quantifying the impact of these different factors on volcanic temperature signals requires systematic numerical experimentation (Rieger et al. 2020; Fyfe et al. 2021).

Recent studies suggest that the Montreal Protocol led to a partial recovery of lower stratospheric ozone and TLS in the early 21st century (Solomon et al. 2016, 2017; Philipona et al. 2018; Petropavlovskikh et al. 2019; Banerjee et al. 2020). All model and observational TLS data sets analyzed here exhibit behavior consistent with ozone recovery: pronounced global-mean cooling of the lower stratosphere over the ozone depletion portion of the satellite record, followed by weaker cooling or near-zero trends over the recovery period (Solomon et al. 2017; Philipona et al. 2018; Steiner et al. 2020; Mitchell et al. 2020; see Fig. 4). The multi-model average TLS trends for these two periods are -0.36 and -0.07°C/decade in CMIP5 and -0.26 and -0.06°C/decade in CMIP6. During the ozone depletion period, the larger multi-model average lower stratospheric cooling in the older CMIP5 simulations is in better accord with satellite TLS trends, which range from −0.42 to −0.49°C/decade. This is partly due to the larger (negative) ozone-induced stratospheric radiative forcing in CMIP5 (Checa-Garcia et al. 2018).

Other factors may also contribute to reduced lower stratospheric cooling in CMIP6 over 1979 to 2000. These factors include CMIP5-versus-CMIP6 differences in forcing from tropospheric ozone (Checa-Garcia et al. 2018), volcanoes (see above) and stratospheric water vapor (Keeble et al. 2020), along with differences in the behavior of tropical upwelling. The zonal-mean structure of trends in TLS and TMT (Fig. 5) reveals prominent differences between CMIP5 and CMIP6 in the tropics, where any differences in the behavior of tropical upwelling should manifest (Ball et al. 2020). More detailed analyses and more systematic numerical experimentation will be required to quantify the relative contributions of forcing, response, chemistry, and dynamics to differences between CMIP5 and CMIP6 TLS trends (Checa-Garcia et al. 2018; Fyfe et al. 2021).

b. Troposphere

Multi-decadal warming of the global troposphere is a ubiquitous feature of the observations and all CMIP5 and CMIP6 forced simulations (Figs. 1B, C). Over the full satellite era, the MMA tropospheric warming rate is very similar in CMIP5 and CMIP6 (0.28 and 0.29°C/decade, respectively). This holds both for TMT and TLT (Fig. 6A). The similarity of the CMIP5 and CMIP6 results is noteworthy given that CMIP6 has a larger number of models with higher Transient Climate Response (TCR) and higher Effective Climate Sensitivity (ECS) (Zelinka et al. 2020; Flynn and Mauritsen 2020; Meehl et al. 2020). An independent analysis of surface temperature supports our finding: despite higher average TCR and ECS in CMIP6, the MMA historical surface warming rate is comparable in older and newer generations of CMIP models, possibly due to a larger response to anthropogenic aerosol forcing in CMIP6 (Flynn and Mauritsen 2020; Fyfe et al. 2021).

In the four single-model large ensembles, the spread of TMT and TLT trends arising from internal variability is substantial, spanning 31 to 47% of the trend spread in the CMIP5 and CMIP6 multi-model ensembles (Fig. 6A).[iv] These results are consistent with other recent comparisons of LE spread to multi-model ensemble spread (Mitchell et al. 2020; Po-Chedley et al. 2021).

Observed trends in global-mean tropospheric temperature range from 0.13 to 0.19°C/decade for TMT and from 0.13 to 0.21°C/decade for TLT (Fig. 6A). For TLT, over 84% of the total number of CMIP5 and CMIP6 extended HIST realizations analyzed here have trends exceeding the largest observational result; the corresponding figure is 91% for corrected TMT trends. Related work suggests that the smaller observed warming is partly due to an unusual manifestation of natural internal variability. Model realizations with phasing of internal variability similar to the observations yield global-mean and tropical tropospheric temperature trends that are within the range of satellite results (Po-Chedley et al. 2021).

In all individual extended HIST realizations, the ratio $R_{\{TMT/TLT\}}$ between global-mean trends in TMT and TLT is close to unity (Fig. 6B). This narrow range occurs despite differences in external forcings, ECS, and internal variability in the multi-model and single-model ensembles, and despite differences in the patterns of warming in TMT and TLT (Santer et al. 2019). The CMIP5 and CMIP6 sampling distributions of $R_{\{TMT/TLT\}}$ encompass the UAH, ERA5.1, and RSS "MSU merge" results. The latter data set relies solely on information from earlier MSU instruments during the 1999 to 2003 overlap period between MSU and AMSU measurements (see Section 2a). The other two RSS data sets, "AMSU merge" and baseline, depart noticeably from the model-based expectations, yielding $R_{\{TMT/TLT\}}$ values significantly less than one.

[iv]This percentage represents $(s_{LE}/s_{CMIP}) * 100$, where s_{LE} is the standard deviation of the sampling distribution of trends in an individual CMIP5 LE or CMIP6 LE and s_{CMIP} is the standard deviation of the sampling distribution of ensemble-mean trends in the corresponding CMIP5 or CMIP6 multi-model ensemble containing the LE.

It is now recognized that there were systematic deficiencies in the early 21st century solar and volcanic forcing used in CMIP5 (Kopp and Lean 2011; Solomon et al. 2012; Flato et al. 2013; Schmidt et al. 2014). Efforts were made to improve representation of both forcings in CMIP6 (Eyring et al. 2016; Gillett et al. 2016; Thomason et al. 2018; Rieger et al. 2020). We find, however, that CMIP5 and CMIP6 multi-model average trends in TMT are virtually identical over 2001 to 2019 (Fig. 7). Since other external forcings also changed between these two generations of models (Checa-Garcia et al. 2018), isolating the climate impact of improvements in volcanic or solar forcing is challenging. Such diagnosis will benefit from simulations in which the same physical climate model is run with different versions of individual forcings (Fyfe et al. 2021).

Tropospheric trends in ERA5.1 exhibit several notable differences relative to the satellite data sets (Hersbach et al. 2020). Reanalysis TMT trends are smaller than in all satellite data sets over 1979 to 2000 and larger than in all satellite data sets over 2001 to 2019 (Fig. 7). Over the 2002 to 2018 period covered by Global Positioning Satellite (GPS) radio occultation measurements, both GPS data and radiosondes yield trends in the middle troposphere that are in reasonable accord with the ERA5.1 results (Steiner et al. 2020).

While the satellite data analyzed here are derived from measurements of microwave emissions alone, ERA5.1 uses a state-of-the-art 4D assimilation system to constrain a weather forecast model with a wide range of multi-variable measurements from satellites, radiosondes, and surface stations (Hersbach et al. 2020; Simmons et al. 2020). Detailed observing system experiments can help to understand the impact of different features of the assimilation system and assimilated data (Bormann et al. 2019). Such studies will be useful in reconciling the trend differences found here and elsewhere (Steiner et al. 2020) between microwave sounders and ERA5.1.

5. Signal-to-noise properties and model-data signal differences

In previous statistical comparisons of modeled and observed temperature changes, discussion often focused on the appropriateness of different comparison periods (Santer et al. 2011). This can be uninformative if attention is restricted to a short segment of the overall temperature record. Here we analyze atmospheric temperature changes over all N_L maximally overlapping L-year periods (see SM). We consider four different values of L: 10, 20, 30, and 40 years. For each value of L, sampling variability is reduced by averaging over all N_L individual measures of temperature change. As we show below, examining timescale-average behavior can have diagnostic value.

Figure 8 shows two different types of statistic: trends and regression coefficients. Results are from individual observational data sets and from distributions of statistics in forced and unforced simulations.

Consider the trend results first. Rows 1-3 of Fig. 8 display trends in TLS, TMT, and TLT (respectively) for our four selected values of the timescale L. With increasing L, the amplitude of internally generated trends decreases. As a result, the standard deviations of the forced and unforced trend distributions decrease. For all three atmospheric layers, forced and unforced trend distributions are completely separated at $L = 40$ years (Figs. 8D, H, and L). This is a simple visual illustration of the timescale-dependence of signal and noise, and of the difficulty in their separation on shorter, noisier timescales of 1-2 decades (Santer et al. 2011).

Despite the evolution in model complexity and resolution between CMIP5 and CMIP6, the sampling distributions of unforced atmospheric temperature trends are remarkably similar in the two generations of coupled models. The same is true for the sampling distributions of forced trends on 10- and 20-year timescales. On longer 30- and 40-year timescales, however, small differences are apparent in the distributions of forced tropospheric temperature trends in CMIP5 and CMIP6.

These may arise because CMIP5 and CMIP6 do not have identical multi-decadal evolution of certain external forcings (Checa-Garcia et al. 2018; Fyfe et al. 2021).

Figure 8 also provides information on the consistency between global-mean temperature trends in observations and the extended HIST simulations. On shorter 10- and 20-year timescales, all observed TLS, TMT, and TLT trends are contained within the respective CMIP5 and CMIP6 distributions of forced trends. The same is true for observed TLS trends on longer 30- and 40-year timescales (Figs. 8C, D). For TMT and TLT, however, only observed data sets with larger tropospheric warming rates are within the model 30- and 40-year distributions of forced trends. The UAH-inferred warming on these timescales is invariably smaller than model expectations (Figs. 8G, H, K, and L).

Amplification of warming with increasing height is a well-known and well-understood property of the tropical atmosphere (Stone and Carlson 1979; Santer et al. 2005; Held and Soden 2006). Figures 8M-P display one measure of tropical amplification behavior – the regression coefficient $b_{\{\mathrm{TMT:TLT}\}}$ between time series of tropical ocean averages of TMT and TLT. All model and observational values of $b_{\{\mathrm{TMT:TLT}\}}$ are greater than 1, indicating that temperature changes in the mid- to upper troposphere exceed those in the lower troposphere. The means and widths of the CMIP5 and CMIP6 sampling distributions of $b_{\{\mathrm{TMT:TLT}\}}$ are relatively insensitive to increases in L, and show substantial overlap for the forced and unforced runs. The model results imply that $b_{\{\mathrm{TMT:TLT}\}}$ is both timescale-invariant and insensitive to forcing, and that its values may impose a robust, physically based constraint on observations (Santer et al. 2005; Held and Soden 2006).

Observational values of $b_{\{\mathrm{TMT:TLT}\}}$ show a number of interesting features. First, the ERA5.1 and RSS "MSU merge" results are well within the range of model expectations on all four timescales considered here. In terms of this tropical amplification metric, therefore, there is no fundamental discrepancy between simulations and all observations.

Second, as in the model simulations, $b_{\{TMT:TLT\}}$ is timescale-invariant for UAH, ERA5.1, and the RSS "MSU merge" case. While the three RSS sensitivity tests have almost identical $b_{\{TMT:TLT\}}$ values for L = 10 years (Fig. 8M), the RSS baseline and "AMSU correct" data sets yield regression coefficients that decrease in size as L increases, and are generally outside the range of model results for 30- and 40-year timescales (Figs. 8O-P). On these longer timescales, the maximally overlapping L-year windows always sample the 1998 to 2003 transition between earlier and more advanced microwave sounders, and thus are more likely to reflect the impact of different merging choices on amplification behavior (see Section 2a).

Third, the UAH $b_{\{TMT:TLT\}}$ value is $\approx$ 1.1 on all four timescales and is smaller than almost all model results. The anomalous UAH value is due to a change in the method used by the UAH group to estimate TLT (Spencer et al. 2017). The impact of this change was to increase the height of the effective weighting function for TLT, thus decreasing the vertical separation between the TLT and corrected TMT weighting functions. This leads to a smaller amplification ratio. To maintain continuity with previous tropical amplification studies (Santer et al. 2017b) and to increase the amplification signal, the model, RSS, and ERA5.1 results shown here do not use the new UAH approach for calculating TLT.

6. Covariability of different aspects of tropical climate change

Properties of the climate system that are controlled by well-understood physical mechanisms and are tightly constrained in model simulations may be useful for reducing large uncertainties in observed temperature trends (Santer et al. 2005). We consider four such properties here. The first three properties are ratios between tropical WV trends and trends in tropical SST, TLT, and corrected TMT.[v] We refer to these ratios as $R_{\{WV/SST\}}$, $R_{\{WV/TLT\}}$, and $R_{\{WV/TMT\}}$ (respectively).

---

[v]Because satellite WV data are available over ocean only, we computed $R_{\{WV/TLT\}}$ and $R_{\{WV/TMT\}}$ using "ocean only" TLT and TMT trends. Temperature gradients are weak in the tropical free troposphere, so whether we use TLT and TMT trends calculated over ocean only or over land

The relationship between temperature and saturation vapor pressure changes is governed by the Clausius-Clapeyron (C-C) equation (Iribarne and Godson 1981). If relative humidity remains approximately constant as temperature increases, C-C predicts the increase in columnar content of WV (Wentz and Schabel 2000; Held and Soden 2006; Mears et al. 2007; O'Gorman and Muller 2010).

The fourth property we examine, the trend ratio $R_{\{\mathrm{TMT/SST}\}}$, is a measure of the amplification of tropical SST changes in the tropical troposphere. Its behavior is governed by moist thermodynamics (Stone and Carlson 1979; Held and Soden 2006). $R_{\{\mathrm{TMT/SST}\}}$ provides information that differs from that of $b_{\{\mathrm{TMT:TLT}\}}$, the regression-based amplification metric considered in Section 5.[vi]

In a climate model, these four ratios are internally and physically consistent. The observed covariability of tropical WV, tropospheric temperature, and SST should also exhibit internal and physical consistency. As we show below, however, observed values of $R_{\{\mathrm{WV/SST}\}}$, $R_{\{\mathrm{WV/TLT}\}}$, $R_{\{\mathrm{WV/TMT}\}}$, and $R_{\{\mathrm{TMT/SST}\}}$ can be inconsistent for certain combinations of observed data sets, and may depart noticeably from model expectations.

Such departures can have at least three explanations. First, WV, tropospheric temperature, and SST are measured independently by different instruments on different satellites and/or measurement platforms. Each variable has different measurement accuracy and errors. These measurement differences can affect the estimated covariability between multidecadal trends in WV, tropospheric temperature, and SST.

---

and ocean has minimal impact on our results. To be consistent in terms of the domain analyzed, the TMT trends in $R_{\{\mathrm{TMT/SST}\}}$ also rely on data averaged over tropical oceans only.

[vi] $b_{\{\mathrm{TMT:TLT}\}}$ was useful for examining whether the TMT and TLT time series produced by an individual research group yielded internally consistent amplification behavior. $b_{\{\mathrm{TMT:TLT}\}}$ used TMT and TLT information from the same microwave sensors flown on the same satellites. In contrast, observed values of $R_{\{\mathrm{TMT/SST}\}}$ provide information on the physical consistency between multidecadal trends in SST and TMT measurements that are processed by different research groups, and that are obtained using different types of measurement platforms.

Second, the tropospheric temperature and SST data sets analyzed here were generated by multiple research groups. In the case of TMT and TLT, each research group uses different procedures to adjust for drifts in satellite orbits and instrument calibration, to merge measurements from multiple satellites, and to merge brightness temperatures estimated from earlier and more recent microwave sounders. For SST, groups use different methods to blend information from ships, buoys, drifting floats, and satellites, to adjust for changes over time in how SST measurements were made, and to infill SSTs in data-sparse regions. The decisions made in adjusting tropospheric temperature and SST for these known non-climatic influences can affect trends (Karl et al. 2006, 2015; Hausfather et al. 2017; Mears et al. 2011; Mears and Wentz 2016, 2017; Zou and Qian 2016; Zou et al. 2018; Spencer et al. 2017; Po-Chedley et al. 2015), and can therefore influence the estimated covariability between real-world tropical temperature and WV changes (or between observed trends in SST and TMT). Trends in satellite WV data are also sensitive to data set construction choices (Mears et al. 2018), but we currently have uncertainty estimates from the RSS group only.[vii]

Third, models may have incomplete or inaccurate representation of the basic physics driving observed tropical covariability relationships on multidecadal timescales. This seems unlikely (Held and Soden 2006), particularly given the fact that on interannual timescales, observed tropical covariability relationships between surface and tropospheric temperature (Santer et al. 2005) and between temperature and WV (Mears et al. 2007) are well captured by models (see Section 7).

Figure 9 shows scatter plots of the individual trend components of the four ratio statistics. For each statistic, model results are tightly constrained in the CMIP5 and CMIP6 multi-model ensembles. At least 96% of the variance in simulated WV trends (plotted on the y-axis in panels A-C) and in simulated TMT trends (plotted on the y-axis of panel D) is explained by simulated trends in the

[vii] We do not use the reanalysis-derived WV trend in estimating structural uncertainties in observed WV trends. Other research has found possible problems with WV trends inferred from reanalysis products (Bengtsson et al. 2004; Wang et al. 2020).

independent (x-axis) variable. This indicates that the four covariance relationships of interest here are relatively insensitive to model differences in the applied historical forcings, the temperature and WV responses to these forcings, and the properties of simulated multi-decadal internal variability. A related inference is that even though most of the mass of atmospheric water vapor resides in the lower troposphere, simulated tropical SST, TLT, and TMT trends impose similar constraints on simulated tropical WV trends – i.e., there is no evidence that on multidecadal timescales, SST or TLT explain noticeably more of the WV variance than TMT.

The regression fits to the CMIP5 and CMIP6 trends are 8.5 and 8.7%/decade for WV and SST, 6.3 and 6.4%/decade for WV and TLT, and 5.3 and 5.5%/decade for WV and TMT (Figs. 9A-C, respectively). The decrease in regression slope in the progression from panels A to C in Fig. 9 reflects the fact that tropical temperature changes closely follow a moist adiabatic lapse rate (Stone and Carlson 1979). As the magnitude of warming amplifies with increasing height, the slope of the regression between temperature trends and moisture trends decreases. The regression slope for simulated tropical SST and TMT trends (1.6 for both CMIP5 and CMIP6; see Fig. 9D) is also consistent with MALR expectations.

Unlike the model covariance relationships in Fig. 9, all four sets of observed covariance relationships show substantial spread. The tight clustering of model expectations and the large observational uncertainty are clearer if we directly compare trend ratios (Fig. 10).[viii] This comparison reveals that observed SST and tropospheric temperature data sets with the largest tropical warming over 1988 to 2019 have $R_{\{\mathrm{WV/SST}\}}$, $R_{\{\mathrm{WV/TLT}\}}$, and $R_{\{\mathrm{WV/TMT}\}}$ ratios closest to the model results (Figs. 10A-C).

[viii]The lowest and highest observational values for $R_{\{\mathrm{WV/SST}\}}$, $R_{\{\mathrm{WV/TLT}\}}$, $R_{\{\mathrm{WV/TMT}\}}$, and $R_{\{\mathrm{TMT/SST}\}}$ vary by factors of 1.6, 1.7, 1.8, and 2.9, respectively. The larger range for $R_{\{\mathrm{TMT/SST}\}}$ arises because there is appreciable observational uncertainty in both the numerator and denominator of the ratio. In the three ratios involving WV, the structural uncertainty of observed trends can be estimated in the denominator only.

For all three ratios involving WV trends, there is minimal overlap between simulations and observations – observed ratios generally exceed model expectations. For $R_{\{WV/SST\}}$, only the COBE SST trend leads to a result consistent with model expectations (Fig. 10A). For both $R_{\{WV/TLT\}}$ and $R_{\{WV/TMT\}}$, observed trend ratios are larger than almost all of the 289 model results (Figs. 10B,C).[ix] The agreement between model and observed $R_{\{TMT/SST\}}$ values is closer, but depends on the selected combination of observed TMT and SST data sets (Fig. 10D).

We calculated Z-scores to summarize and synthesize the information in Fig. 10. For each observed ratio in Fig. 10, the Z-score is simply the difference between the observed result and the mean of the CMIP5 or CMIP6 multi-model average ratio, normalized by the CMIP5 or CMIP6 standard deviation of the sampling distribution of the ratio in question. The Z-scores in Fig. 11A are averages over the individual scores arising from structural uncertainty in observed SST trends; they are measures of the consistency between the simulated values of $R_{\{WV/TLT\}}$, $R_{\{WV/TMT\}}$, and $R_{\{TMT/SST\}}$ and the observed values of these ratios estimated with a specific TLT or TMT data set. The Z-scores in Fig. 11B are defined analogously, and are averages over the individual Z-scores arising from structural uncertainty in observed tropospheric temperature trends.

Under the assumption that the model-generated distributions of the four ratios are realistic representations of the true (but uncertain) real-world covariance relationships, the Z-scores allow us to make certain inferences about the likelihood that individual observed SST and tropospheric temperature data sets are consistent with model expectations and with other other observations. In Fig. 11A, for example, STAR and RSS "MSU merge" – the data sets with the largest observed tropospheric warming trends – are closest to the model expectations of WV/tropospheric temperature trend ratios, and therefore have the smallest Z-scores for $R_{\{WV/TLT\}}$ and $R_{\{WV/TMT\}}$. In contrast, the

[ix] For each ratio, there are 123 values for CMIP5 and 166 for CMIP6. For $R_{\{WV/TLT\}}$ and $R_{\{WV/TMT\}}$, only 4 and 3 of the 289 extended HIST realizations (respectively) have scaling ratios exceeding the smallest observed value.

muted tropospheric warming in UAH leads to $R_{\{\mathrm{WV/TLT}\}}$ and $R_{\{\mathrm{WV/TMT}\}}$ values that are significantly larger than model expectations, thus leading to large UAH Z-scores for these two ratios. Based on $R_{\{\mathrm{WV/TLT}\}}$ and $R_{\{\mathrm{WV/TMT}\}}$ alone, therefore, we might infer that the smaller tropical tropospheric warming trend is UAH is less credible.

This inference assumes that the observed trend in tropical WV is accurate. A substantially smaller observed WV trend would decrease the UAH-derived $R_{\{\mathrm{WV/TLT}\}}$ and $R_{\{\mathrm{WV/TMT}\}}$ ratios, bringing them in closer agreement with model expectations. Since we do not have estimates of the observed WV trend from multiple research groups, it is difficult to assess the likelihood that the true (but uncertain) real-world WV trend is markedly smaller than the RSS WV trend estimate.

By considering the $R_{\{\mathrm{TMT/SST}\}}$ ratio, however, we can bring in independently monitored observed SST data. This allows us to explore the constraint that observed SST trends impose on the size of observed TMT trends. The COBE, ERSST, and HadSST data sets, when considered in combination with the UAH TMT trend, lead to UAH-based $R_{\{\mathrm{TMT/SST}\}}$ ratios that are significantly smaller than climate model and MALR expectations (Fig. 10D). Only the muted tropical surface warming in the HadISST data set yields a UAH-based $R_{\{\mathrm{TMT/SST}\}}$ ratio that is marginally consistent with model expectations. The weaker surface warming in HadISST is inconsistent with independently monitored WV data (see Figs. 10A and 11B).

To summarize, the reduced tropical tropospheric warming in UAH is not supported by: 1) an independent estimate of atmospheric moistening from satellite data; 2) all independent estimates of observed sea surface warming except HadISST; and 3) all model and theoretical expectations of $R_{\{\mathrm{WV/TLT}\}}$, $R_{\{\mathrm{WV/TMT}\}}$, and $R_{\{\mathrm{TMT/SST}\}}$. In turn, the HadISST tropical SST trend that is marginally consistent with the muted UAH tropospheric warming is not supported by independently monitored satellite WV or TMT data, or by model and theoretical expectations of $R_{\{\mathrm{WV/SST}\}}$ and $R_{\{\mathrm{TMT/SST}\}}$.

The above analysis focused on comparing simulated and observed measures of tropical covariability. It is also of interest to compare modeled and observed values of the individual components of these covariability metrics. In the case of WV, 21% of the model WV trends are smaller than the satellite-estimated WV trend in Fig. 9A. For SST, TLT, and TMT, only 17%, 12%, and 12% of the model trends are within the range of observed results (Figs. 9A-C, respectively).

There are multiple interpretations of this finding. One interpretation is that the higher level of consistency between simulated and observed tropical WV trends reflects a systematic low bias in observed tropical TLT and TMT trends over 1988 to 2019. An alternative explanation is that the satellite WV trend is overestimated. It is difficult to discriminate between these two possibilities without additional information, such as well-quantified estimates of uncertainties in observed WV trends from different research groups.

7. Conclusions

Relative to CMIP5, the more recent CMIP6 models have higher resolution (on average), more complete numerical portrayal of Earth's climate system, and nominally improved representation of external forcings (Eyring et al. 2016). These advances do not guarantee improved agreement between simulations and observations. This is apparent in at least two aspects of model performance analyzed here: lower stratospheric cooling over the ozone depletion period and the stratospheric temperature response to the El Chichón eruption. Understanding why these features are more accurately represented in CMIP5 will require more systematic diagnostic efforts to disentangle evolutionary changes in models from evolutionary changes in model forcings (Fyfe et al. 2021).

The development of satellite temperature data sets remains a work in progress. Adjustments for known non-climatic factors can have significant impact on observed trends in tropospheric temperature, as well as on basic physical properties related to tropospheric warming (Karl et al.

2006; Mears et al. 2011; Mears and Wentz 2016, 2017; Zou and Qian 2016; Zou et al. 2018; Spencer et al. 2017; Po-Chedley et al. 2015). Multi-model and single-model large ensembles tightly constrain four such physical properties – the ratio between tropical trends in WV and SST, WV and TLT, WV and TMT, and TMT and SST. These are denoted here by $R_{\{\mathrm{WV/SST}\}}$, $R_{\{\mathrm{WV/TLT}\}}$, $R_{\{\mathrm{WV/TMT}\}}$, and $R_{\{\mathrm{TMT/SST}\}}$, respectively. Comparing modeled and observed values of such basic covariance relationships has the advantage (relative to single-variable comparisons) that results are less sensitive to model-versus-observed differences in the phasing of internal variability (Santer et al. 2005; Po-Chedley et al. 2021).

We find significant differences between simulated and observed values of $R_{\{\mathrm{WV/SST}\}}$, $R_{\{\mathrm{WV/TLT}\}}$, $R_{\{\mathrm{WV/TMT}\}}$, with observations exceeding model expectations in most cases (Figs. 10A-C). Observed data sets with larger warming of the tropical ocean surface and tropical troposphere yield trend ratios that are closer to model results. For $R_{\{\mathrm{TMT/SST}\}}$, model-data consistency depends on the selected combination of observed data sets used to estimate TMT and SST trends (Fig. 10D).

One interpretation of our findings is that they are due to a systematic low bias in satellite tropospheric temperature trends – i.e., that the size of the observed tropical moistening signal is greater than can be explained by the independently observed warming of the tropical troposphere. Alternately, the observed atmospheric moistening signal may be overestimated. Given the large structural uncertainties in observed tropical TMT and SST trends, and because satellite WV data are available from one group only, it is difficult to determine which interpretation is more credible.

What we can say with confidence, however, is that decisions regarding how to merge MSU and AMSU TMT data have substantial impact on observed tropical TMT trends. This is evident from the three RSS sensitivity tests examined here. These sensitivity tests point towards merging decisions as a significant contributory factor to uncertainties in observed $R_{\{\mathrm{WV/TMT}\}}$ and $R_{\{\mathrm{TMT/SST}\}}$ trend ratios (Figs. 10C,D).

Two further points are relevant to the question of whether the model-observed differences in Figs. 10A-C are mainly due to underestimated observed tropospheric temperature trends or to an overestimated satellite WV trend. First, there is some evidence that observational uncertainties may be smaller in satellite WV data than in satellite tropospheric temperature data (Wentz 2013; see Section 2c). Second, when the individual trend components of our four trend ratios are examined, the agreement between models and observations is better for WV and SST trends than for TMT or TLT trends. This difference in model-data consistency, taken together with higher measurement accuracy of WV and the results of the RSS sensitivity tests, suggests that underestimated observed tropospheric warming is plausible. This inference is predicated on the assumption that the model-based covariance constraints are realistic.

While our analysis does not definitively resolve the cause or causes of significant differences between modeled and observed tropospheric warming trends, it does illustrate the diagnostic power of simultaneously considering multiple complementary variables (Wentz and Schabel 2000). Our study also highlights the strong internal and physical consistency between the model constraints derived from multidecadal tropical trends in WV, TMT, and SST. Examining additional independently monitored constraints may be helpful in reducing the currently large uncertainties in observations of tropical climate change.

Acknowledgments. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling 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. This work was performed under the auspices of the U.S. Department of Energy (DOE) by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. At LLNL, B.D.S., S.P.-C., M.D.Z., and J.P. were supported by the Regional and Global Model Analysis Program of the Office of Science at the DOE. S.P.-C. was also supported under LDRD 18-ERD-054. All primary satellite, reanalysis, and model temperature data sets used here are publicly available. Synthetic satellite temperatures calculated from model simulations and the ERA 5.1 reanalysis are provided at: https://pcmdi.llnl.gov/research/DandA/. We thank Adrian Simmons at ECMWF for assistance with ERA 5.1 data and internal reviewers at NOAA and CCCma for helpful comments.

APPENDIX A

Calculation of synthetic satellite temperatures from model data

a. Calculation of synthetic satellite temperatures

We use a local weighting function method developed at RSS to calculate synthetic satellite temperatures from CMIP5 and CMIP6 output and from the ERA5.1 reanalysis (Santer et al. 2017b). At each grid-point, simulated temperature profiles were convolved with local weighting functions. The weights depend on the grid-point surface pressure, the surface type (land, ocean, or sea ice), and the selected layer-average temperature (TLS, TMT, or TLT). The local weighting function method provides more accurate estimates of synthetic satellite temperatures than use of a global-mean weighting function, particularly over high elevation regions.

APPENDIX B

Method used for correcting TMT data

Trends in TMT estimated from microwave sounders receive a substantial contribution from the cooling of the lower stratosphere (Fu et al. 2004; Fu and Johanson 2004, 2005; Johanson and Fu 2006). In Fu et al. (2004), a regression-based method was developed for removing the bulk of this stratospheric cooling component of TMT. This method has been validated with both observed and model atmospheric temperature data (Fu and Johanson 2004; Gillett et al. 2004; Kiehl et al. 2005). We calculated two different versions of corrected TMT, the first with latitudinally fixed and the second with latitudinally varying regression coefficients. We refer to these subsequently as TMT_1 and TMT_2, respectively. The main text discusses corrected TMT_1 only, and does not use the subscript 1 to identify corrected TMT.

The regression equation applied in Fu and Johanson (2005) for calculating corrected TMT is:

$$\mathrm{TMT} = a_{24}\mathrm{TMT} + (1 - a_{24})\mathrm{TLS} \tag{B1}$$

For TMT_1, we use a_{24} =1.1 at each latitude. For TMT_2, a_{24} =1.1 between 30°N and 30°S, and a_{24} =1.2 poleward of 30°. This is consistent with how we have calculated TMT_1 and TMT_2 in previous work (Santer et al. 2017b).

The advantage of TMT_2 is that lower stratospheric cooling makes a larger contribution to TMT trends at mid- to high latitudes. The latitudinally varying regression coefficients in TMT_2 remove more of this extratropical cooling. We prefer to use the more conservative TMT_1 here. In practice, the choice of TMT_1 or TMT_2 has minimal influence on the statistical significance of differences between the modeled and observed statistics of interest here (temperature trends and a regression-based measure of the amplification of warming with increasing height in the tropical atmosphere).

References

Aquila, V., W. H. Swartz, D. W. Waugh, P. R. Colarco, S. Pawson, L. M. Polvani, and R. S. Stolarski, 2016: Isolating the roles of different forcing agents in global stratospheric temperature changes using model integrations with incrementally added single forcings. *J. Geophys. Res.*, **121**, 8067–8082, doi:10.1002/2015JD023841.

Ball, W. T., G. Chiodo, M. Abalas, and J. Alsing, 2020: Inconsistencies between chemistry climate model and observed lower stratospheric trends since 1998. *Atmos. Chem. Phys.*, doi: doi.org/10.5194/acp-2019-734, (in review).

Bandoro, J., S. Solomon, B. D. Santer, D. Kinnison, and M. Mills, 2018: Detectability of the impacts of ozone-depleting substances and greenhouse gases upon global stratospheric ozone accounting for nonlinearities in historical forcings. *Atmos. Chem. Phys.*, **18**, 143–166, doi: doi.org/10.5194/acp-18-143-2018.

Banerjee, A., J. C. Fyfe, L. M. Polvani, D. Waugh, and K.-L. Chang, 2020: A pause in Southern Hemisphere circulation trends due to the Montreal Protocol. *Nature*, **579**, 544–548.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard, 2014: ENSO representation in climate models: from CMIP3 to CMIP5. *Cli. Dyn.*, **42**, 1999–2018, doi: 10.1007/s00382-013-1783-z.

Bengtsson, L., S. Hagemann, and K. I. Hodges, 2004: Can climate trends be calculated from reanalysis? *J. Geophys. Res.*, **109**, D11111, doi:10.1029/2004JD004536.

Bengtsson, L., and J. Shukla, 1988: Integration of space and in situ observations to study global climate change. *Bull. Am. Meteorol. Soc.*, **69**, 1130–1143.

Bormann, N., H. Lawrence, and J. Farnan, 2019: Global observing system experiments in the ECMWF assimilation system. Technical Memo 839, European Centre for Medium-Range Weather Forecasts, 24 pp. doi:10.21957/sr184iyz.

Checa-Garcia, R., M. I. Hegglin, D. Kinnison, D. A. Plummer, and K. P. Shine, 2018: Historical tropospheric and stratospheric ozone radiative forcing using the CMIP6 database. *Geophys. Res. Lett.*, **45**, 3264–3273, doi:10.1002/2017GL076770.

Danabasoglu, G., and Coauthors, 2020: The Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, **12**, e2019MS001916, doi: 10.1029/2019MS001916.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate projections: The role of internal variability. *Cli. Dyn.*, **38**, 527–546.

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Clim. Change*, **10**, 277–286.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Mod. Dev.*, **9(5)**, 1937–1958, doi:10.5194/gmd-9-1937-2016.

Eyring, V., and Coauthors, 2019: Taking model evaluation to the next level. *Nat. Clim. Change*, **9**, 102–110.

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment i Report of the*

Intergovernmental Panel on Climate Change, 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, Eds., Cambridge University Press, 741–866.

Flynn, C. M., and T. Mauritsen, 2020: On the climate sensitivity and historical warming evolution in recent coupled model ensembles. *Atmos. Chem. Phys.*, **20**, 7829–7842, doi:doi.org/10.5194/acp-2019-1175.

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

Fu, Q., and C. M. Johanson, 2005: Satellite-derived vertical dependence of tropical tropospheric temperature trends. *Geophys. Res. Lett.*, **32**, L10703, doi:10.1029/2004GL022266.

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

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.

Fyfe, J. C., V. Kharin, B. D. Santer, R. N. S. Cole, and N. P. Gillett, 2021: Significant impact of forcing uncertainty in a large ensemble of climate model simulations. *Proc. Nat. Acad. Sci.*, (in review).

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

Fyfe, J. C., and Coauthors, 2017: Large near-term projected snowpack loss over the western United States. *Nature Communications*, **8**, doi:10.1038/ncomms14996.

Gates, W. L., and Coauthors, 1999: An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bull. Am. Meteor. Soc.*, **29**, 29–55.

Gillett, N. P., B. D. Santer, and A. J. Weaver, 2004: Quantifying the influence of stratospheric cooling on satellite-derived tropospheric temperature trends. *Nature*, **432**, doi:10.1038/nature03209.

Gillett, N. P., and Coauthors, 2016: The detection and attribution model intercomparison project (DAMIP v1.0) contribution to CMIP6. *Geosci. Mod. Dev.*, **9**, 3685–3697.

Hartmann, D. L., and Coauthors, 2013: Observations: Atmosphere and Surface. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 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, Eds., Cambridge University Press, 159–254.

Hausfather, Z., K. Cowtan, D. C. Clarke, P. Jacobs, M. Richardson, and R. Rohde, 2017: Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Sci. Adv.*, **3, e1601207**, doi:10.1126/sciadv.1601207.

Held, I. M., and B. J. Soden, 2006: Robust responses of the hydrological cycle to global warming. *J. Clim.*, **19**, 5686–5699.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Q. J. Roy. Met. Soc.*, **146**, 1999–2049.

Hirahara, S., M. Ishii, and Y. Fukuda, 2014: Centennial-scale sea surface temperature analysis and its uncertainty. *J. Clim.*, **27**, 57–75.

Huang, B., and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Clim.*, **30**, 8179–8205.

Iribarne, J. V., and W. L. Godson, 1981: *Atmospheric Thermodynamics*. D. Reidel, 276 pp.

Johanson, C. M., and Q. Fu, 2006: Robustness of tropospheric temperature trends from MSU Channels 2 and 4. *J. Clim.*, **19**, 4234–4242.

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

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*. National Oceanic and Atmospheric Administration, 164 pp.

Karl, T. R., and Coauthors, 2015: Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, **348**, 1469–1472.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model: Large ensemble project. *Bull. Amer. Met. Soc.*, **96**, 1333–1349.

Keeble, J., and Coauthors, 2020: Evaluating stratospheric ozone and water vapor changes in cmip6 models from 1850-2100. *Atmos. Chem. Phys.*, doi:10.5194/acp-2019-1202.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011: 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.

Kiehl, J. T., J. Caron, and J. J. Hack, 2005: On using global climate model simulations to assess the accuracy of MSU retrieval methods for tropospheric warming trends. *J. Clim.*, **18**, 2533–2539.

Kopp, G., and J. L. Lean, 2011: A new, lower value of total solar irradiance: Evidence and climate significance. *Geophys. Res. Lett.*, **38**, L01706, doi:10.1029/2010GL045777.

Maycock, A. C., and Coauthors, 2018: Revisiting the mystery of recent stratospheric temperature trends. *Geophys. Res. Lett.*, **45**, 9919–9933, doi:10.1029/2018GL078035.

Mears, C., D. K. Smith, L. Ricciardulli, J. Wang, H. Huelsing, and F. J. Wentz, 2018: Construction and uncertainty estimation of a satellite-derived total precipitable water data record over the world's oceans. *Earth and Space Sci.*, **5**, 197–210.

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

Mears, C., and F. J. Wentz, 2017: A satellite-derived lower-tropospheric atmospheric temperature dataset using an optimized adjustment for diurnal effects. *J. Clim.*, **30**, 7695–7718.

Mears, C., 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 technique. *J. Geophys. Res.*, **116**, D08112, doi:10.1029/2010JD014954.

Mears, C. A., B. D. Santer, F. J. Wentz, K. E. Taylor, and M. Wehner, 2007: The relationship between temperature and precipitable water changes over tropical oceans. *Geophys. Res. Lett.*, **34**, L2470, doi:10.1029/2007GL031936.

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

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

Meehl, G. A., C. A. Senior, V. Eyring, G. Flato, J.-F. Lamarque, R. J. Stouffer, K. E. Taylor, and M. Schlund, 2020: Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Sci. Advances*, **6**, doi:10.1126/sciadv.aba1981.

Meinshausen, M., and Coauthors, 2011: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, **109**, 213–241.

Mills, M. J., and Coauthors, 2016: Global volcanic aerosol properties derived from emissions, 1990–2014, using CESM1 (WACCM). *J. Geophys. Res. Atmos.*, **121**, 2332–2348.

Mitchell, D. M., Y. T. E. Lo, W. J. M. Seviour, L. Haimberger, and L. M. Polvani, 2020: The vertical profile of recent tropical temperature trends: Persistent model biases in the context of internal variability. *Env. Res. Lett.*, (in press), [https://iopscience.iop.org/article/10.1088/1748-9326/ab9af7].

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, doi:10.1029/2011JD017187.

NRC, 2000: *Reconciling observations of global temperature change*. National Academy Press, Washington D.C., 169 pp.

O'Gorman, P. A., and C. J. Muller, 2010: How closely do changes in surface and column water vapor follow Clausis-Clapeyron scaling in climate change simulations? *Env. Res. Lett.*, **5**, doi:10.1088/1748-9326/5/2/025207.

Petropavlovskikh, I., S. Godin-Beekmann, D. Hubert, R. Damadeo, B. Hassler, and V. S. (Eds.), 2019: SPARC/IO3C/GAW Report on Long-term Ozone Trends and Uncertainties in the Stratosphere. Tech. Rep. SPARC Report No. 9, GAW Report No. 241, WCRP-17/2018. [http://www.sparc-climate.org/publications/sparc-reports/sparc-report-no9].

Philipona, R., and Coauthors, 2018: Radiosondes show that after decades of cooling, the lower stratosphere is now warming. *J. Geophys. Res.*, **123**, 12 509–12 522, doi:10.1029/2018JDR028901.

Po-Chedley, S., and Q. Fu, 2012: Discrepancies in tropical upper tropospheric warming between atmospheric circulation models and satellites. *Environ. Res. Lett.*, **7**, doi:10.1088/1748-9326/7/4/044018.

Po-Chedley, S., B. D. Santer, S. Fueglistaler, M. D. Zelinka, P. Cameron-Smith, J. F. Painter, and Q. Fu, 2021: Natural variability drives model-observational differences in tropical tropospheric warming. *Proc. Nat. Acad. Sci.*, (in press).

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

Ramaswamy, V., M. D. Schwarzkopf, W. J. Randel, B. D. Santer, B. J. Soden, and G. L. Stenchikov, 2006: Anthropogenic and natural influences in the evolution of lower stratospheric cooling. *Science*, **311**, 1138–1141.

Randall, D. A., and Coauthors, 2007: Climate Models and Their Evaluation. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, 589–662.

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

Riahi, K., and Coauthors, 2017: The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Glob. Env. Change*, **42**, 153–168, doi:10.1016/j.gloenvcha.2016.05.009.

Ridley, D. A., and Coauthors, 2014: Total volcanic stratospheric aerosol optical depths and implications for global climate change. *Geophys. Res. Lett.*, **41**, 7763–7769, doi:10.1002/2014GL061541.

Rieger, L. A., J. N. S. Cole, J. C. Fyfe, S. Po-Chedley, P. Cameron-Smith, P. J. Durack, N. P. Gillett, and Q. Tang, 2020: Quantifying CanESM5 and EAMv1 sensitivities to volcanic forcing for the CMIP6 historical experiment. *Geosci. Mod. Dev.*, **13**, 4831–4843, doi:https://doi.org/10.5194/gmd-13-4831-2020.

Robock, A., 2000: Volcanic eruptions and climate. *Rev. Geophys.*, **38**, 191–219.

Santer, B. D., J. Fyfe, S. Solomon, J. Painter, C. Bonfils, G. Pallotta, and M. Zelinka, 2019: Quantifying stochastic uncertainty in detection time of human-caused climate signals. *Proc. Nat. Acad. Sci.*, **116**, 19 821–19 827, doi:doi.org/10.1038/s41558-019-0424-x.

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

Santer, B. D., and Coauthors, 2011: Separating signal and noise in atmospheric temperature changes: The importance of timescale. *J. Geophys. Res.*, **116**, D22105, doi:10.1029/2011JD016263.

Santer, B. D., and Coauthors, 2013: Human and natural influences on the changing thermal structure of the atmosphere. *Proc. Nat. Acad. Sci.*, **110**, 17 235–17 240, doi:10.1073/pnas.1305332110.

Santer, B. D., and Coauthors, 2017a: Causes of differences between model and satellite tropospheric warming rates. *Nat. Geosci.*, **10**, 478–485.

Santer, B. D., and Coauthors, 2017b: Comparing tropospheric warming in climate models and satellite data. *J. Clim.*, **30**, 3–4.

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

Shine, K. P., and Coauthors, 2003: A comparison of model-simulated trends in stratospheric temperatures. *Q. J. Roy. Met. Soc.*, **129**, 1565–1588, doi:10.1256/qj.02.186.

Simmons, A., and Coauthors, 2020: Global stratospheric temperature bias and other stratospheric aspects of ERA5 and ERA5.1. Technical Memo 859, European Centre for Medium-Range Weather Forecasts, 40 pp.

Solomon, S., 1999: Stratospheric ozone depletion: A review of concepts and history. *Rev. Geophys.*, **37**, 275–316.

Solomon, S., J. S. Daniel, R. R. Neely, 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.

Solomon, S., D. J. Ivy, D. Kinnison, M. J. Mills, R. R. N. III, and A. Schmidt, 2016: Emergence of healing in the Antarctic ozone layer. *Science*, **353**, 269–274, doi:10.1126/science.aae0061.

Solomon, S., P. J. Young, and B. Hassler, 2012: Uncertainties in the evolution of stratospheric ozone and implications for recent temperature changes in the tropical lower stratosphere. *Geophys. Res. Lett.*, **39**, L17706, doi:10.1029/2012GL052723.

Solomon, S., and Coauthors, 2017: Mirrored changes in Antarctic ozone and stratospheric temperature in the late 20th versus early 21st centuries. *J. Geophys. Res.*, **122**, 8940–8950, doi:10.1002/2017JD026719.

Spencer, R. W., J. R. Christy, and W. D. Braswell, 2017: UAH version 6 global satellite temperature products: Methodology and results. *Asia-Pac. J. Atmos. Sci.*, **53**, 121–130, doi:10.1007/s13143-017-0010-y.

Sperber, K. R., H. Annamalai, I.-S. Kang, A. Kitoh, A. Moise, A. Turner, B. Wang, and T. Zhou, 2013: The Asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century. *Cli. Dyn.*, **41**, 2711–2744, doi:10.1007/s00382-012-1607-6.

Steiner, A., and Coauthors, 2020: Observed temperature changes in the troposphere and stratosphere from 1979 to 2018. *J. Clim.*, **33**, 8165–8194, doi:https://doi.org/10.1175/JCLI-D-19-0998.1.

Stone, P. H., and J. H. Carlson, 1979: Atmospheric lapse rate regimes and their parameterization. *J. Atmos. Sci.*, **36**, 415–423.

Swart, N. C., S. T. Gille, J. C. Fyfe, and N. P. Gillett, 2018: Recent Southern Ocean warming and freshening driven by greenhouse gas emissions and ozone depletion. *Nat. Geosci.*, **11**, 836–841.

Tatebe, H., and Coauthors, 2019: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, **12 (7)**, 2727–2765, doi:10.5194/gmd-12-2727-2019.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Thomason, L. W., and Coauthors, 2018: A global space-based stratospheric aerosol climatology: 1979–2016. *Earth Syst. Sci. Data*, **10**, 469–492, doi:doi.org/10.5194/essd-10-469-2018.

Thompson, D. W. J., and Coauthors, 2012: The mystery of recent stratospheric temperature trends. *Nature*, **491**, 692–697, doi:10.1038/nature11579.

Thorne, P. W., J. R. Lanzante, T. C. Peterson, D. J. Seidel, and K. P. Shine, 2011: Tropospheric temperature trends: History of an ongoing controversy. *Wiley Inter. Rev.*, **2**, 66–88.

Trenberth, K. E., J. Fasullo, and L. Smith, 2005: Trends and variability in column-integrated atmospheric water vapor. *Cli. Dyn.*, **24**, 741–758.

Trenberth, K. E., and Coauthors, 2007: Observations: Surface and Atmospheric Climate Change. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Eds., Cambridge University Press, 235–336.

US Senate, 2015: Data or Dogma? Promoting open inquiry in the debate over the magnitude of human impact on Earth's climate. Hearing before the U.S. Senate Committee on Commerce, Science, and Transportation, Subcommittee on Space, Science, and Competitiveness, One Hundred and Fourteenth Congress, first session, December 8. [https://clio.columbia.edu/catalog/12267036].

Wang, S., T. Xu, W. Nie, C. Jiang, Y. Yang, Z. Fang, M. Li, and Z. Zhang, 2020: Evaluation of precipitable water vapor from five reanalysis products with ground-based GNSS bbservations. *Remote Sensing*, **12**, 1817, doi:10.3390/rs12111817.

Wentz, F. J., 2013: SSM/I Version-7 Calibration Report. Tech. Rep. Technical Report 011012, 46 pp. [http://www.remss.com/papers/tech_reports/2012_Wentz_011012_Version-7_SSMI_Calibration.pdf].

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. C. Schabel, 2000: Precise climate monitoring using complementary satellite data sets. *Nature*, **403**, 414–416.

Zelinka, M. D., T. A. Myers, D. T. McCoy, S. Po-Chedley, P. M. Caldwell, P. Ceppi, S. A. Klein, and K. E. Taylor, 2020: Causes of higher climate sensitivity in CMIP6 models. *Geophys. Res. Lett.*, **47**, doi:10.1029/2019GL085782.

Zou, C.-Z., M. D. Goldberg, and X. Hao, 2018: New generation of U.S. satellite microwave sounder achieves high radiometric stability performance for reliable climate change detection. *Sci. Adv.*, **4**, eaau0049.

Zou, C.-Z., and H. Qian, 2016: Stratospheric temperature climate record from merged SSU and AMSU-A observations. *J. Atmos. Ocean. Tech.*, **33**, 1967–1984, doi:10.1175/JTECH-D-16-0018.1.

Zou, C.-Z., and W. Wang, 2011: Inter-satellite calibration of AMSU-A observations for weather and climate applications. *J. Geophys. Res.*, **116**, D23113, doi:10.1029/2011JD016205.

LIST OF FIGURES

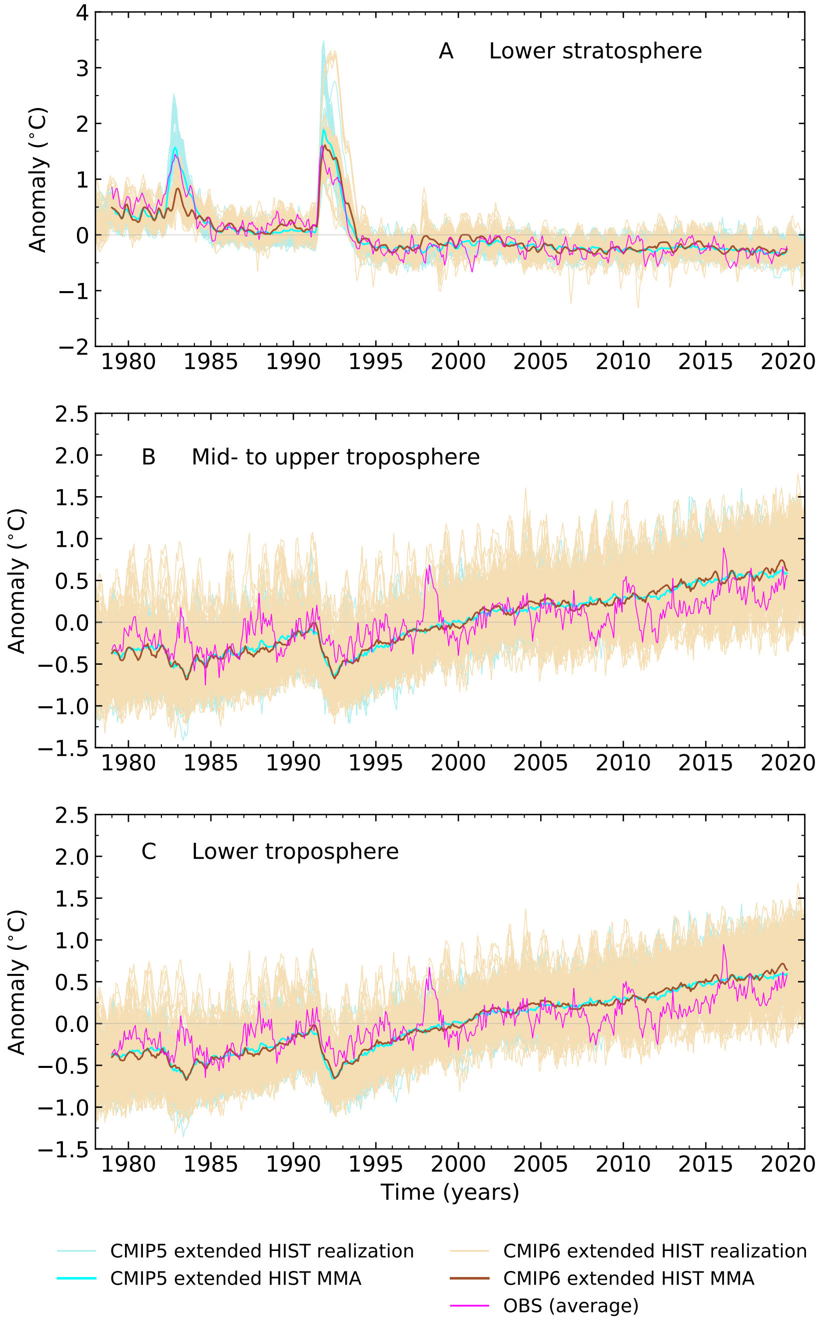

FIG. 1: Time series of monthly-mean near-global averages of the temperature of the lower stratosphere (TLS; panel A), the mid- to upper troposphere (TMT; panel B), and the lower troposphere (TLT; panel C). For TLS and TMT, observations are the average of the RSS "baseline", STAR, and UAH satellite data sets and the ERA 5.1 reanalysis. Since STAR does not produce a TLT data set, the observational average for TLT was calculated with RSS "baseline", UAH, and ERA5.1 only. CMIP5 synthetic satellite temperatures were computed from 123 realizations of historical climate change ("extended HIST") performed with 28 models. For CMIP6, 116 extended HIST realizations were used for TLS and 166 realizations for TMT and TLT (performed with 21 and 22 models, respectively). All temperature changes are defined as anomalies relative to climatological monthly means over 1979 to 2019. TMT is adjusted for the contribution it receives from stratospheric cooling (see Appendix B). Calculation of the multi-model average (MMA) involves first averaging over realizations of an individual model, then averaging over models.

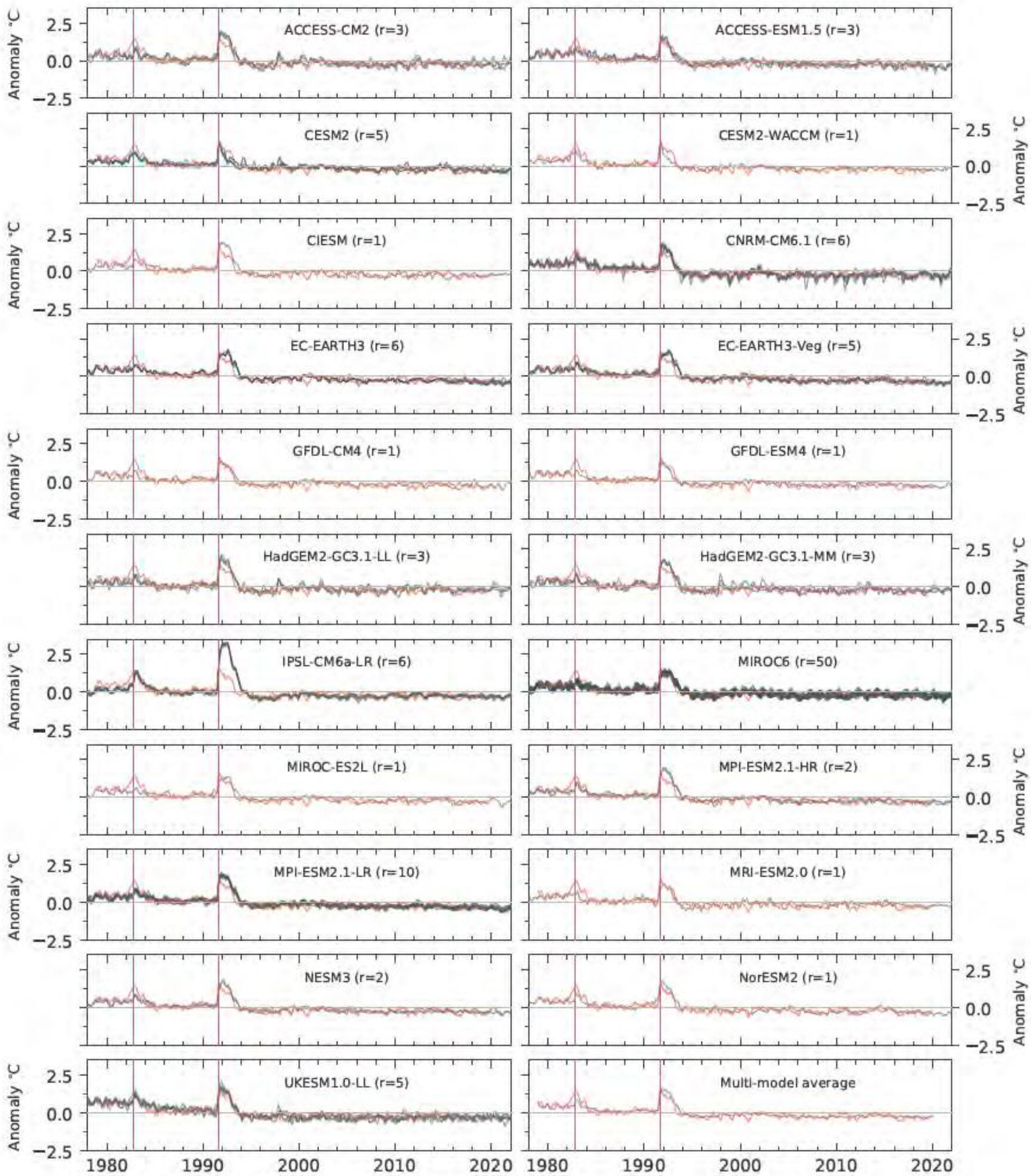

FIG. 2: Time series of monthly-mean anomalies of the temperature of the lower stratosphere (TLS) in CMIP6 extended HIST simulations. Results are for 21 individual CMIP6 models (in grey) and for the RSS "baseline" satellite data (in red). The CMIP6 multi-model average is also shown (bottom right panel). All anomalies are spatially averaged over 82.5°N-82.5°S and are defined relative to climatological monthly means over 1979 to 2019. The number of extended HIST realizations is indicated in parentheses. Vertical lines denote the times of maximum lower stratospheric warming in the RSS "baseline" data after the eruptions of El Chichón and Pinatubo.

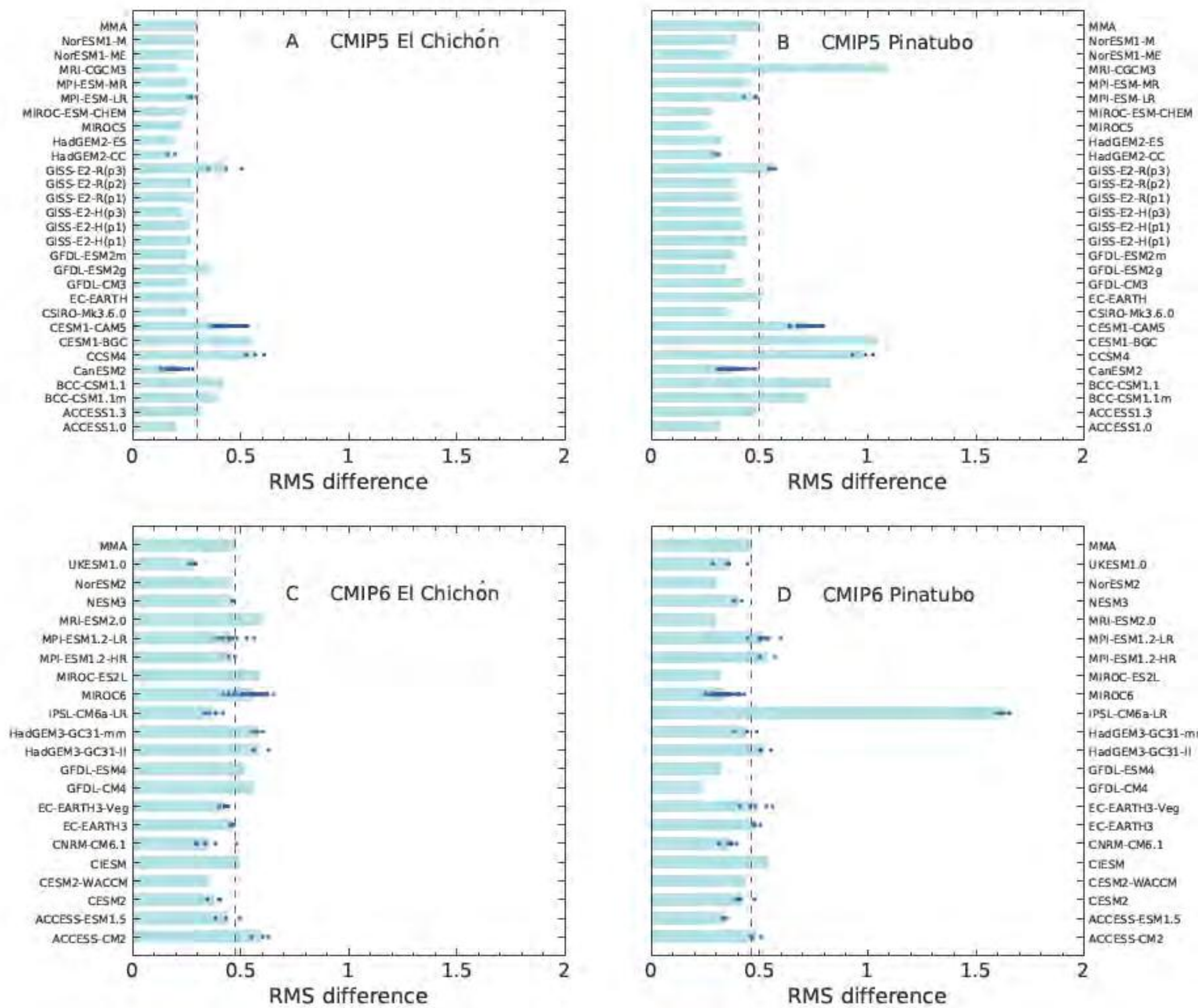

FIG. 3: Root-Mean-Square (RMS) differences between simulated and observed volcanic signals in lower stratospheric temperature in CMIP5 models (panels A, B) and CMIP6 models (panels C, D). RMS differences were calculated for 24-month periods after the 1982 eruption of El Chichón (panels A, C) and the 1991 Pinatubo eruption (panels B, D). The observational target is the RSS "baseline" TLS time series, spatially averaged over 82.5°N-82.5°S. Blue dots denote RMS values from individual realizations of the CMIP5 and CMIP6 extended HIST runs. Horizontal bars are average RMS differences for individual models. The dashed vertical lines are the multi-model average RMS differences, calculated by first averaging RMS values over a model's individual realizations, and then averaging over models.

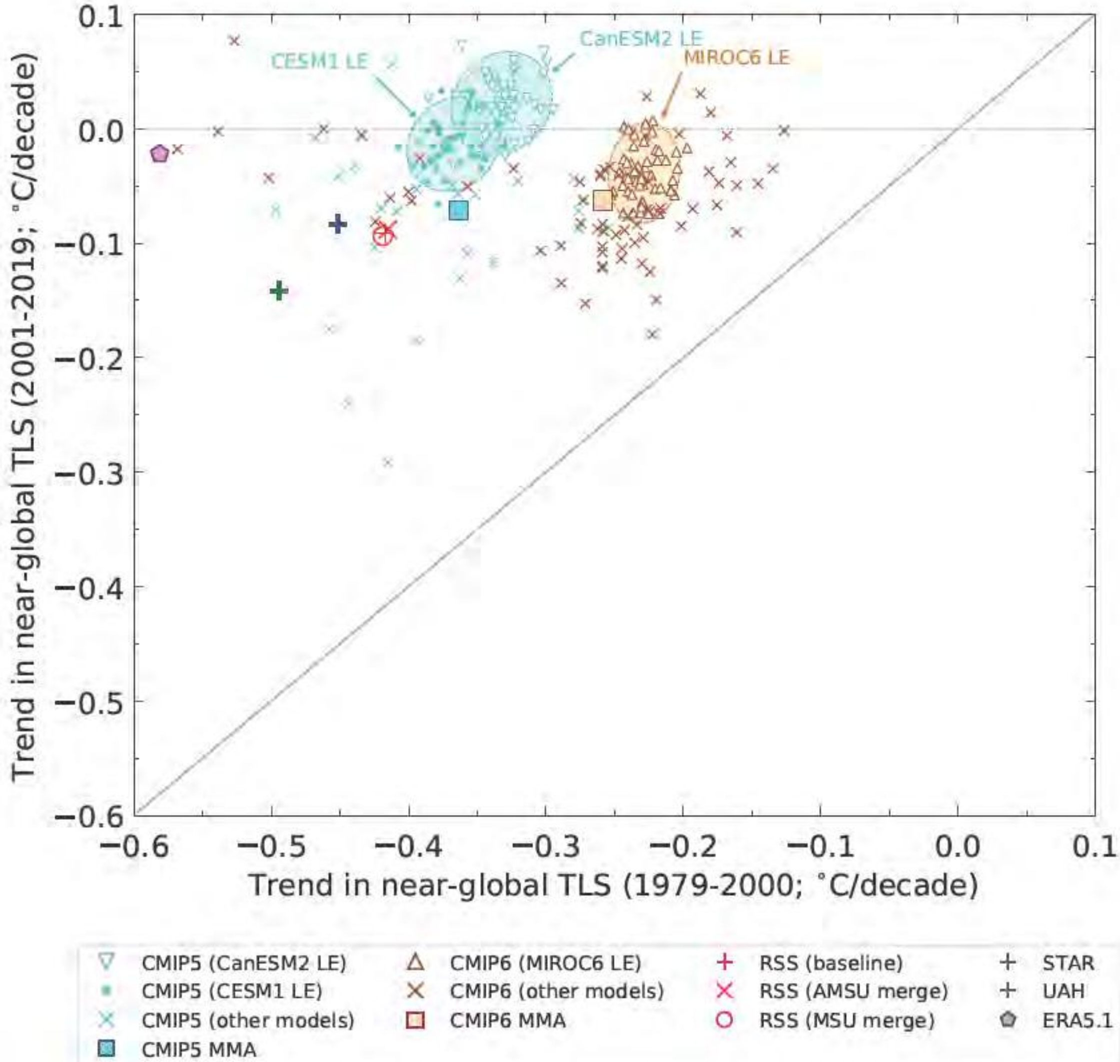

FIG. 4: Least-squares linear trends in near-global average lower stratospheric temperature over ozone depletion and ozone recovery periods (1979 to 2000 and 2001 to 2019, respectively). Model results are from 123 and 116 extended HIST simulations performed with 28 different CMIP5 and 21 different CMIP6 models (respectively). CMIP5 trends include results from the 40-member CESM1 and 50-member CanESM2 large ensembles (LEs). CMIP6 trends incorporate the 50-member MIROC6 LE. Observed estimates of TLS trends rely on satellite data (RSS, STAR, and UAH) and the ERA5.1 reanalysis. Three different versions of the RSS data are shown. The 1:1 line (with trends of equal size over the the ozone depletion and ozone recovery periods) is marked in purple. For both periods, the CMIP5 multi-model average TLS trend is closer to the satellite data results. No individual CMIP5 or CMIP6 realization has larger lower stratospheric cooling in the ozone recovery period than in the ozone depletion period. This underscores the fact that the non-linear behavior of TLS over the satellite era is dominated by the response to ozone forcing, not by multi-decadal internal variability (Solomon et al. 2017). The shaded ellipses are the 2σ confidence intervals for each of the three LEs. For information on spatial averaging and calculation of multi-model averages, refer to Fig. 1.

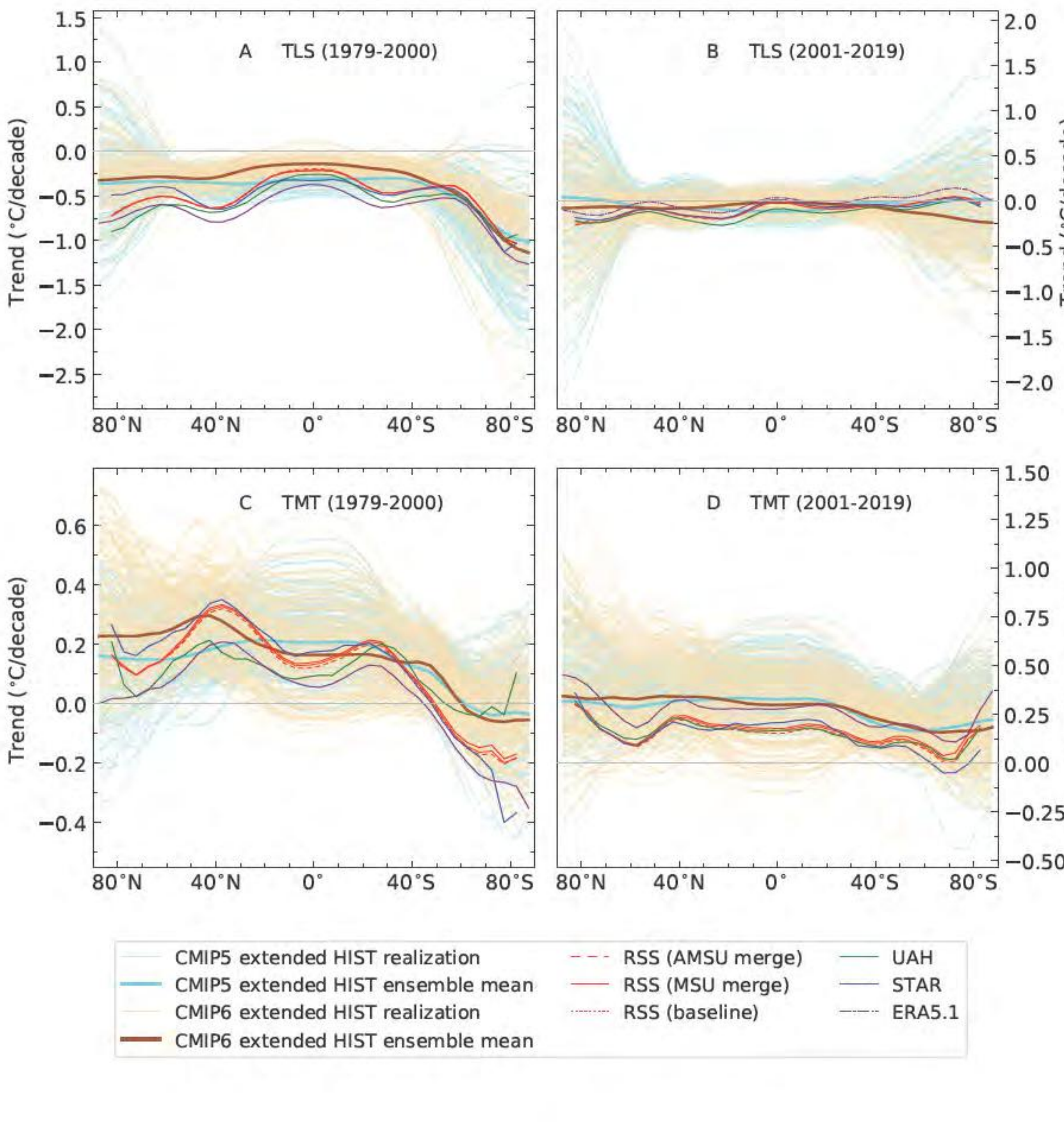

FIG. 5: Zonal-mean trends in monthly-mean lower stratospheric temperature (panels A,B) and in corrected mid- to upper tropospheric temperature (panels C,D). Results are for ozone depletion and ozone recovery periods (left and right columns, respectively). For information regarding the numbers of CMIP5 and CMIP6 models and extended HIST realizations, calculation of multi-model averages, spatial averaging, and observational data, refer to Fig. 1.

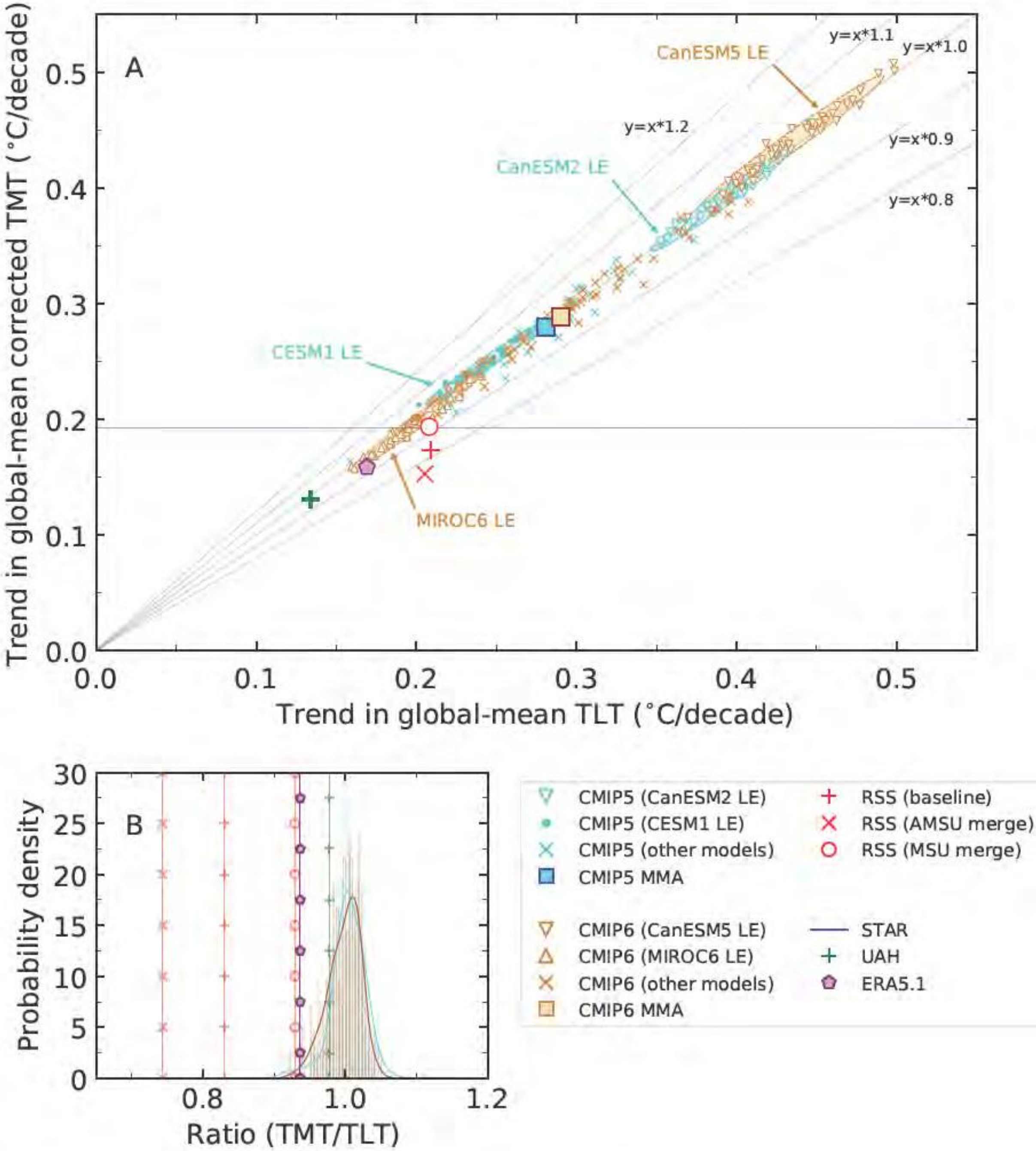

FIG. 6: Scatter plot (panel A) of linear trends in near-global mean lower tropospheric temperature (TLT) and mid- to upper tropospheric temperature (TMT) and histograms of the TMT/TLT trend ratio (panel B). All trends are over 1979 to 2019. TMT is corrected for lower stratospheric cooling. The multi-model averages include information from the 50- and 40-member CanESM2 and CESM1 LEs (for CMIP5) and from the 50-member CanESM5 and MIROC6 LEs (for CMIP6). The shaded ellipses in panel A are the 2σ confidence intervals for each LE. Because TLT is not produced by STAR, the STAR TMT trend is plotted as a horizontal line in panel A. Selected isopleths of equal values of the TMT/TLT trend ratio are denoted by dashed grey lines in panel A. For further details of CMIP5 and CMIP6 realizations and models, calculation of multi-model averages, spatial averaging, observational data sources, and fits to histograms, refer to caption of Fig. 1 and SM.

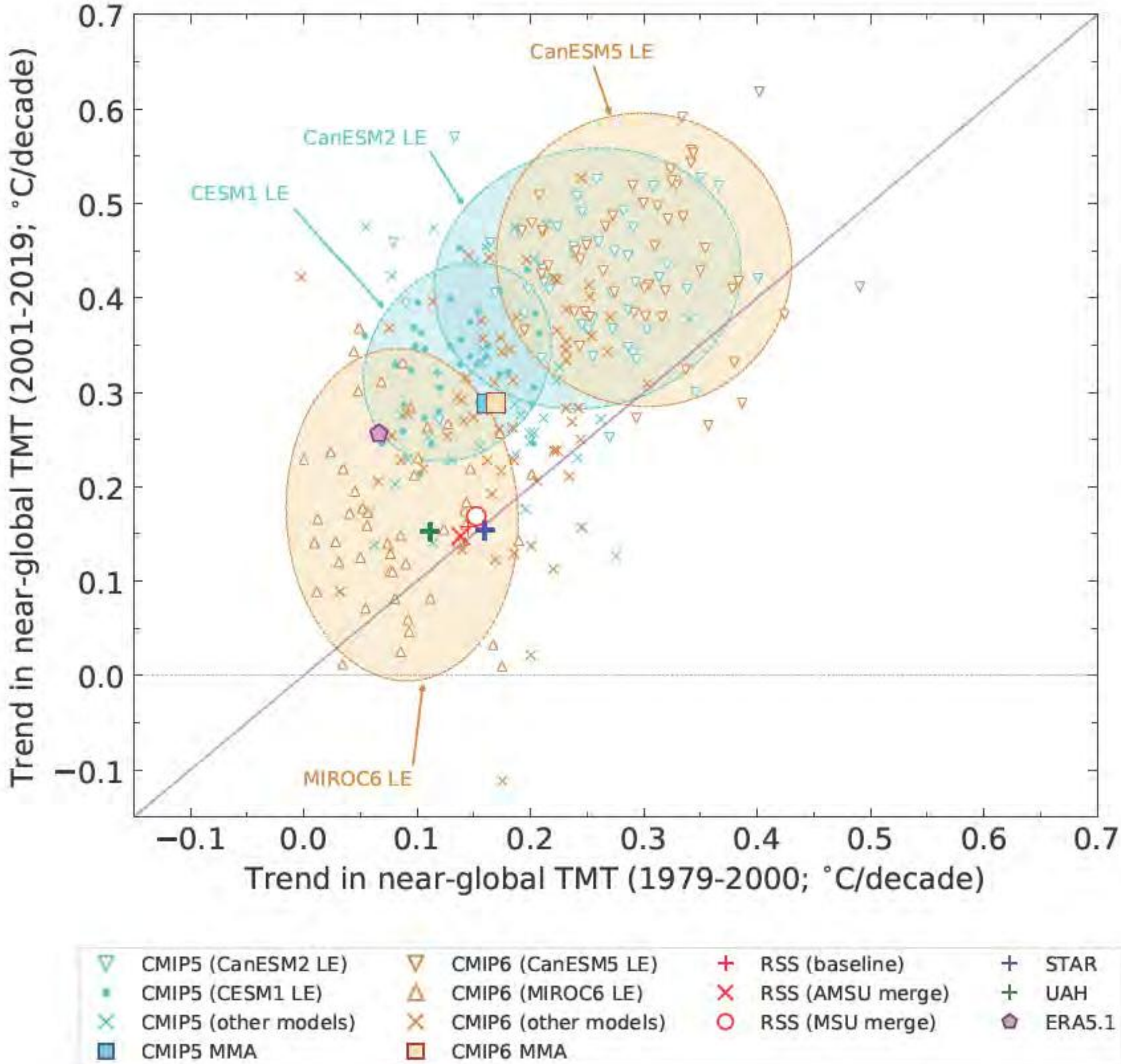

FIG. 7: As for Fig. 4 but for linear trends in near-global average mid- to upper tropospheric temperature (TMT) over 1979 to 2000 (x-axis) and over 2001 to 2019 (y-axis). TMT is corrected for the influence of lower stratospheric cooling. While Fig. 4 excluded TLS results from the 50-member CanESM5 LE because of anomalous TLS variability, TMT trends from the CanESM5 LE are minimally affected by this anomalous variability and are included here. The 1:1 line (with TMT trends of equal size over the two periods) is marked in purple. Simulated TMT trends are larger in the second analysis period in approximately 90% of the realizations. In satellite data, trends in the two periods are of roughly equivalent size.

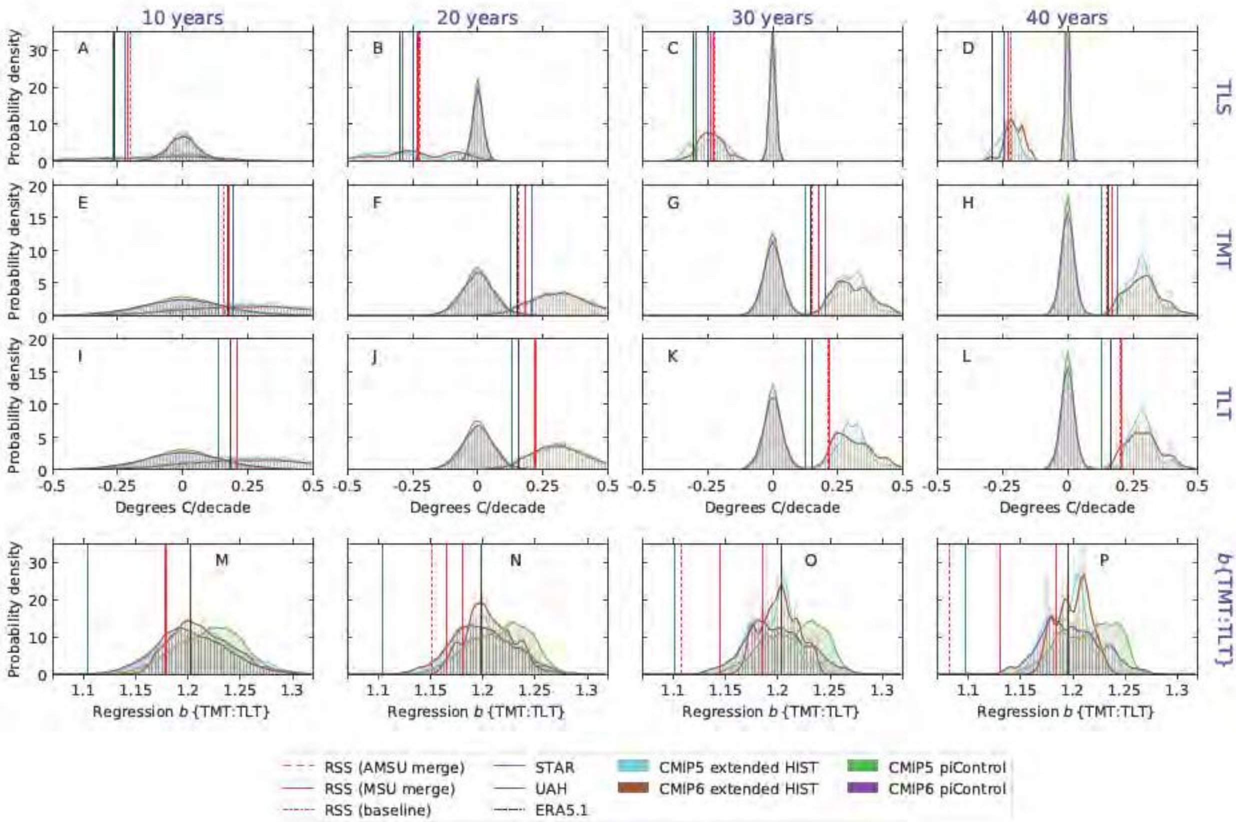

FIG. 8: Trends and regression coefficients in CMIP5, CMIP6, and observations. Maximally overlapping L-year trends were calculated from time series of monthly-mean, near-global spatial averages of TLS, TMT, and TLT (panels A-D, E-H, and I-L, respectively). The regression coefficient $b_{\{TMT:TLT\}}$, a measure of amplification of warming in the tropical troposphere, was computed with maximally overlapping L-year time series of monthly-mean TMT and TLT, spatially averaged over ocean areas between 20°N-20°S (panels M-P). The four selected timescales shown here are 10, 20, 30, and 40 years (columns 1-4, respectively). Histograms of these L-year trends and regression coefficients are shown for CMIP5 and CMIP6 extended HIST simulations and for pre-industrial control runs. Histograms are weighted to account for model differences in the number of extended HIST simulations or in control run length. For each histogram, results are normalized by the total number of trend or regression coefficient samples. Fits to the model trend and $b_{\{TMT:TLT\}}$ distributions were performed with kernel density estimation (see SM). The vertical lines for the observed trends and regression coefficients are the averages across the maximally overlapping L-year analysis periods. For trends in TMT, the RSS "MSU merge" and STAR results are almost identical.

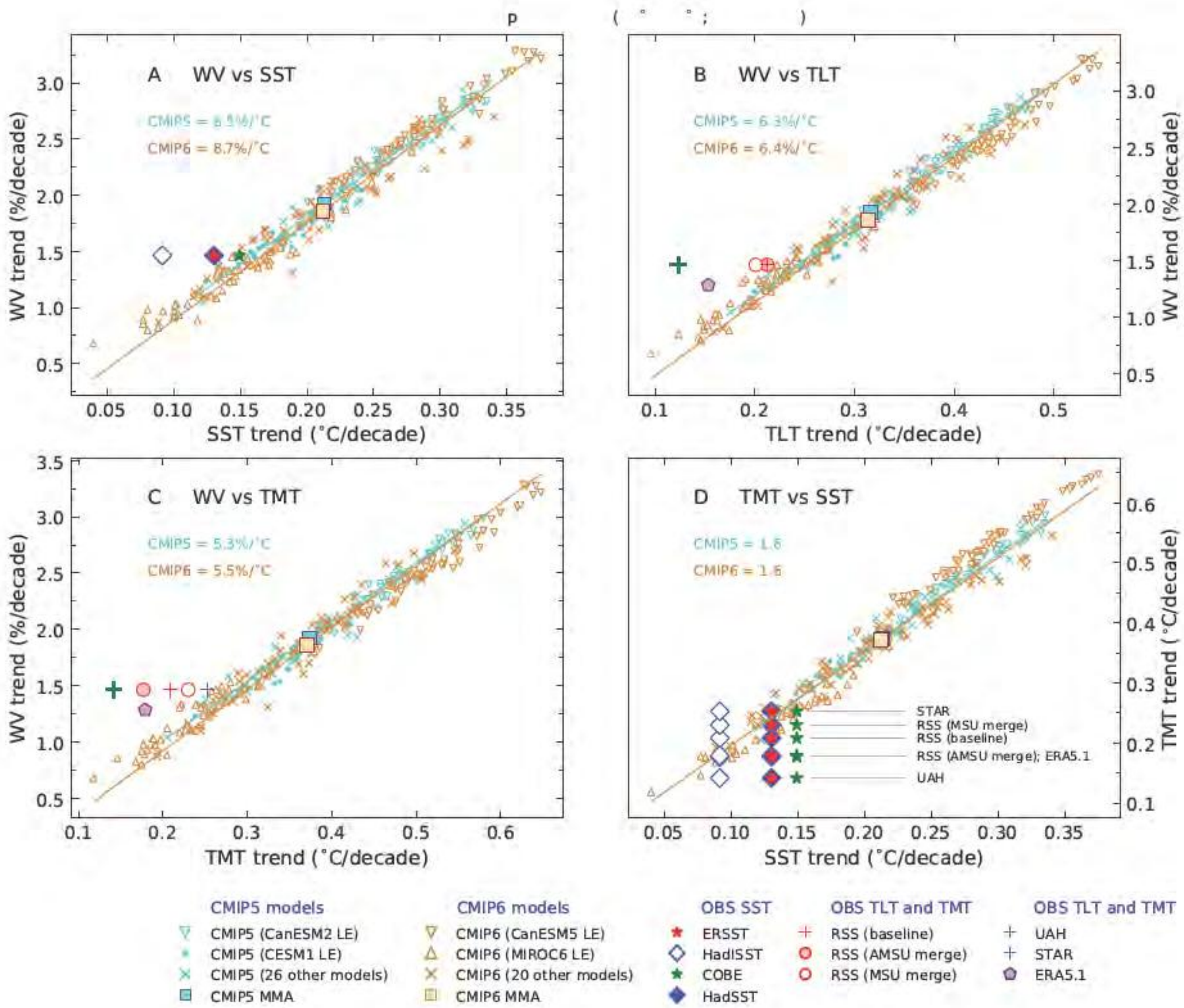

FIG. 9: Scatter plot of tropical trends in WV and SST (panel A), WV and TLT (panel B), WV and corrected TMT (panel C), and corrected TMT and TLT (panel D). Trends are over 1988 to 2019, the period of availability of observed WV data from 7 different microwave radiometers (Mears et al. 2018), and were calculated with WV, TLT, TMT and SST data averaged over tropical oceans (20°N-20°S). Before computing WV trends, monthly-mean WV anomalies were expressed as percentages with respect to climatological monthly means. Because satellite-derived WV is produced by RSS only, all satellite TLT and TMT trends in panels B and C are plotted against the RSS WV trend. ERA5.1 TLT and TMT trends are plotted against the WV trend from the reanalysis. Since there are 4 different observed SST data sets and 6 different observed TMT data sets, there are 4×6 combinations of SST and TMT trends in panel D. The x-axis position of observational symbols in panel D reflects the observed SST trend; the y-axis position depends on the observed TMT trend. The CMIP5 multi-model average trend in each panel include results from the CanESM2 and CESM1 LEs; the CMIP6 multi-model average trend include results from the CanESM5 and MIROC6 LEs. The regression fits and slopes were estimated with Orthogonal Distance Regression and are given separately for CMIP5 and CMIP6 results (see SM).

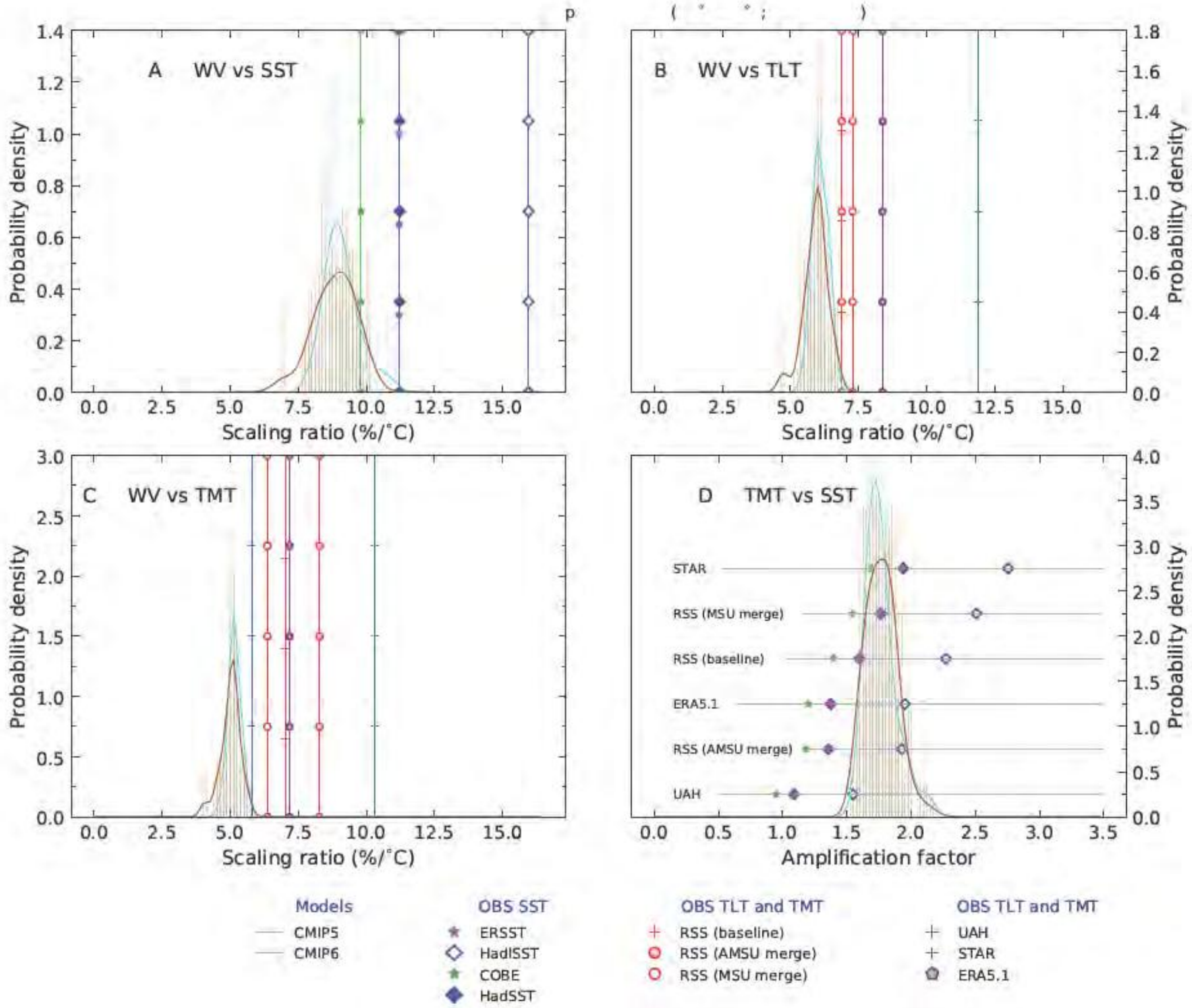

FIG. 10: Histograms of the ratios between the model trends plotted in each of the four panels of Figure 9. Results are for $R_{\{WV/SST\}}$, $R_{\{WV/TLT\}}$, $R_{\{WV/TMT\}}$, and $R_{\{TMT/SST\}}$ (panels A-D, respectively). Observational trend ratios in panels A-C are plotted as vertical lines. Each satellite TMT data set in panel D can be paired with 4 different observed SST trends, yielding 4 different observed values of $R_{\{TMT/SST\}}$ (see Fig. 9 caption). Observed $R_{\{TMT/SST\}}$ values in panel D are plotted in six rows, one row per satellite TMT data set. The vertical spacing and y-axis location of rows is nominal; the vertical ordering of rows reflects the size of the observed tropical TMT trend over 1988 to 2019. The largest TMT trend (in the STAR data set) has the largest y-axis offset in panel D. For details regarding fits to the model histograms and histogram weighting, refer to SM.

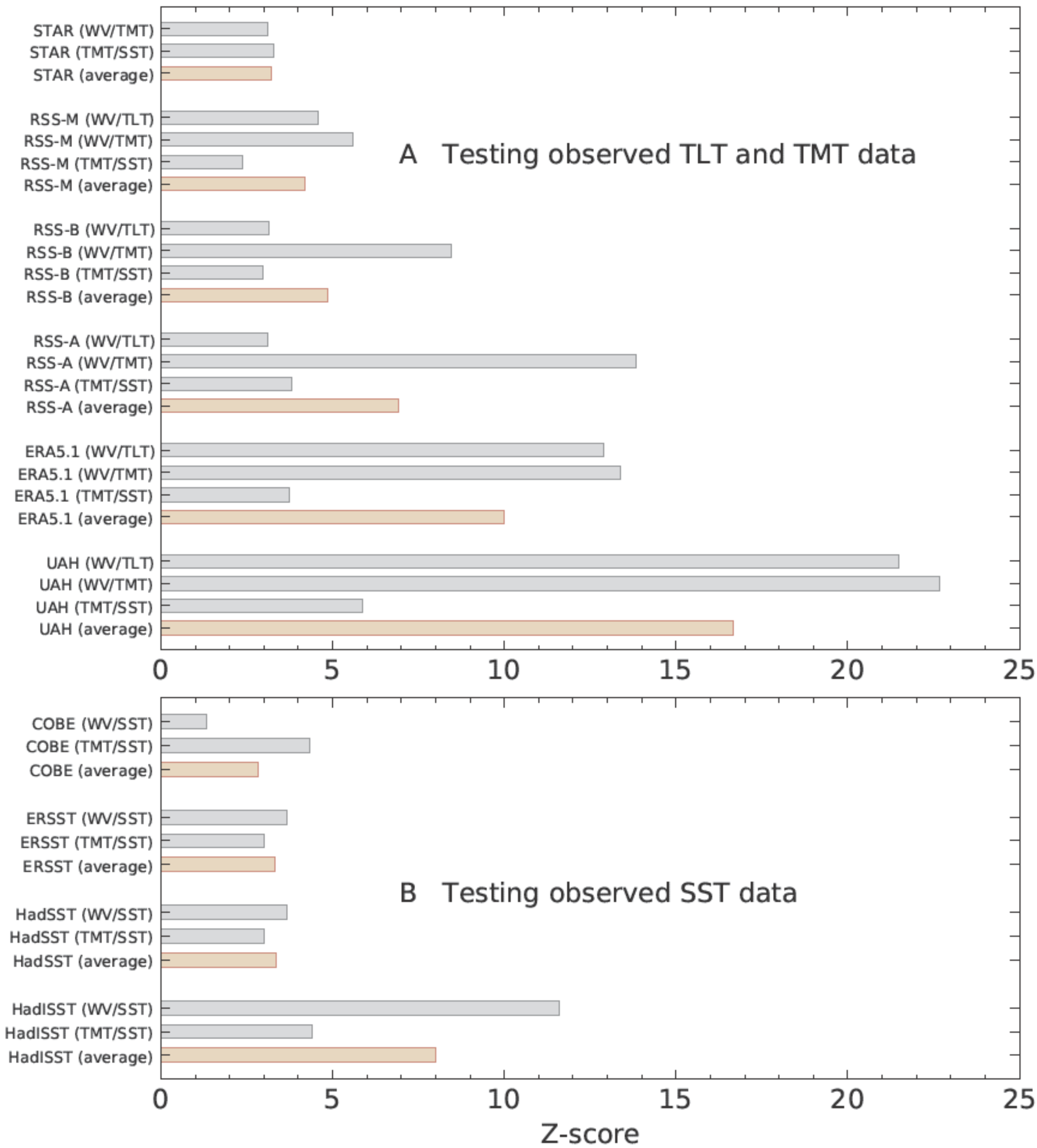

FIG. 11: Normalized differences (Z-scores) between observed scaling ratios and the mean of model scaling ratio distributions. Results in panel A are for tests of $R_{\{WV/TLT\}}$ ratios based on 5 different observed TLT data sets and for tests of $R_{\{WV/TMT\}}$ and $R_{\{TMT/SST\}}$ ratios based on 6 different observed TMT data sets. Panel B involves tests of $R_{\{WV/SST\}}$ and $R_{\{TMT/SST\}}$ with 4 different observed SST data sets. All Z-scores were calculated with the scaling ratio data in Fig. 10. For each ratio tested, the observed ratio is subtracted from the mean of the CMIP5 or CMIP6 sampling distribution of the ratio. These differences are normalized by the CMIP5 or CMIP6 standard deviation of the ratio's sampling distribution; CMIP5 and CMIP6 Z-scores are then averaged. For the $R_{\{TMT/SST\}}$ ratios in panel A, there is an additional averaging step: each observed TMT data set can be paired with 4 different observed SST data sets, yielding 4 different Z-scores (see rows in Fig. 10D). We average these 4 values per TMT data set. Likewise, each observed SST data set in panel B can be paired with 6 different TMT data sets, yielding 6 different values of $R_{\{TMT/SST\}}$ (see columns in Fig. 10D). We average these 6 values per SST data set. The brown bars are average Z-scores for different types of scaling ratio.