The vertical profile of recent tropical temperature trends: Persistent model biases in the context of internal variability

Dann M. Mitchell[1], Y. T. Eunice Lo[1], William J. M. Seviour[1,2], Leopold Haimberger[3] and Lorenzo M. Polvani[4]

[1]Cabot Institute for the Environment, and School of Geographical Sciences, University of Bristol, Bristol, UK
[2]Global Systems Institute, and Department of Mathematics, University of Exeter, Exeter, UK
[3]Department of Meteorology and Geophysics, University of Vienna, Austria
[4]Department of Applied Physics & Applied Mathematics, and Lamont-Doherty Earth Observatory, Columbia University, New York, NY USA

E-mail: `d.m.mitchell@bristol.ac.uk`

April 15, 2020

Abstract. Tropospheric and stratospheric tropical temperature trends in recent decades have been notoriously hard to simulate using climate models, notably in the upper troposphere. Aside from the warming trend itself, this has broader implications, e.g. atmospheric circulation trends depend on latitudinal temperature gradients. In this study, tropical temperature trends in the CMIP6 models are examined, from 1979 to 2014, and contrasted with trends from the RICH/RAOBCORE radiosondes, and the ERA5/5.1 reanalysis. As in earlier studies, we find considerable warming biases in the CMIP6 modeled trends, and we show that these biases are linked to biases in surface temperature (these models simulate an unrealistically large global warming). We also uncover previously undocumented biases in the lower-middle stratosphere: the CMIP6 models appear unable to capture the time evolution of stratospheric cooling, which is non-monotonic owing to the Montreal Protocol. Finally, using models with large ensembles, we show that their standard deviation in tropospheric temperature trends, which is due to internal variability alone, explains $\sim$50% ($\pm$20%) of that from the CMIP6 models.

Keywords: temperature trends, troposphere, stratosphere, models, CMIP6, bias

Submitted to: *Environ. Res. Lett.*

1. Introduction

Since the pioneering work of Manabe & Wetherald (1975) climate models have consistently shown greater warming in the tropical upper troposphere than near the surface in response to increased CO_2 concentrations. This robust differential warming is understood to result from convection which, at low latitudes, tends to adjust the temperature profile to a moist adiabat (Manabe & Stouffer 1980, Santer et al. 2005). In this context, the first paper to analyze atmospheric temperature trends inferred from satellite-based microwave sounders (Spencer & Christy 1990) came as a great surprise, as it reported a lack of warming in the free troposphere over the decade 1979-1988, questioning the reliability of climate models and radiosonde observations. That study generated a great deal of controversy, giving rise to dozens of papers, and two expert panel reports. The reader is referred to Thorne et al. (2011) for the latest exhaustive, if not completely updated, review.

In brief, soon after that controversial paper it became clear that both satellite and radiosonde derived temperature trends suffered from considerable biases (see, e.g. Karl et al. 2006). A large effort, therefore, has gone into producing "homogenized" data sets, from which instrumental artifacts are carefully and methodically removed. Nonetheless, much uncertainty remains as to the vertical structure of the observed temperature trends in the free-atmosphere since 1979, notably in the tropics. A more complete discussion can be found in the relevant section of the Fourth and Fifth Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC, see Hegerl et al. 2007, Hartmann et al. 2013, respectively).

In tandem with the effort to put the observed trends on more solid grounds, climate models have greatly evolved since the early IPCC assessment reports. In the last two decades, most state-of-the-art climate models discretize the atmosphere with dozens of vertical levels, have an accurate representation of the stratosphere, and are coupled to dynamic ocean, sea ice, and other components. In spite of these improvements, however, substantial discrepancies remain – between models and observations – in the vertical structure of atmospheric temperature trends in the tropics. For models participating in Phases 3 and 5 of the Coupled Model Inter-comparison Project (CMIP3 and CMIP5), these discrepancies have been reported in numerous papers (see, e.g., Fu et al. 2011, Po-Chedley & Fu 2012, Santer et al. 2013, 2017).

In particular, it is worth recalling the findings of Mitchell et al. (2013), hereafter referred to as M13. While reporting a considerable discrepancy between radiosonde and modeled trends over the period 1979-2008, that study highlighted the fact that an important source of the discrepancy rested in the modeled surface warming, which was larger than the observed one. M13 showed that the discrepancy between models and observations is greatly reduced in the atmosphere-only CMIP5 model simulations, in which surface temperatures are prescribed from observations.

Building on M13, the goal of this paper is to analyze the recently completed simulations performed under Phase 6 of the Coupled Model Inter-comparison Project

(CMIP6, Eyring et al. 2016), and to explore whether the tropical temperature trends in these models are closer to observations than those of the CMIP5 models. We also address two novel aspects of the problem. First, mindful that the trends in atmospheric concentrations of many ozone depleting substances has peaked shortly before the turn of the century (as a consequence of the Montreal Protocol), we separately compute trends before and after the year 1998, seeking to document the role of ozone depletion on atmospheric temperature trends. Second, in the spirit of Hawkins & Sutton (2009), we take advantage of large ensembles of individual CMIP6 model simulations (as opposed to a single run from each model), and seek to document what fraction of the large spread across the CMIP6 models can be attributed to internal atmospheric variability, as opposed to inter-model differences.

2. Methods

To report the observed atmospheric temperature trends, we make use of three different data sets: two radiosonde data sets, the Radiosonde Innovation Composite Homogenization (RICH, v1.5) and the RAdiosone OBservation COrrection using REanalyses (RAOBCORE, v1.5) products (Haimberger 2007, Haimberger et al. 2012), and one reanalysis data set, ERA5 (Hersbach et al. 2020). Note that ERA5 assimilates the radiosonde data used here, as well as many other data sources. For simplicity, throughout this manuscript we will refer to these three data sets, collectively, as "the observations", even though we are well aware that ERA5 is a reanalysis, with observations assimilated into an underlying model.

The difference between the two radiosonde data sets resides in the procedures used for the homogenization; these are fully detailed in Haimberger et al. (2012). Both radiosonde data sets have been updated to cover the period 1979-2019, at a resolution of $10°\times10°$ in horizontal directions, and 13 levels extending from 850 hPa to 10 hPa in the vertical direction. While temperature data are available at monthly resolution, we here construct annual averages, with the proviso that if more than 3 months of data are missing at a grid point in a given year, we count the entire year as missing. We note that both radiosonde data sets have the same resolution, and the same missing data. Fig. S1 shows the coverage of available data at three of the pressure levels that we focus on in this study.

The ERA5 data set is a high resolution reanalysis produced by the European Centre for Medium-Range Weather Forecasts (Hersbach et al. 2020). Its horizontal resolution is 0.28° (in both latitude and longitude), with data available on 137 pressure levels from the surface to 0.01 hPa. Since ERA5 is at higher spatial resolution than the radiosonde data, we regrid it to the same resolution as the radiosondes using bilinear interpolation, and apply the same missing data mask used for the radiosondes. ERA5 data is available over the period 1979-2018; however, in this study, we substitute the years 2000-2006 with an updated product, ERA5.1. This is necessary as an error was identified in the original ERA5 lower stratospheric temperatures, due to an incorrect specification of the error covariance matrix in the assimilation scheme (Simmons et al. 2020).

The primary model data used in this study consists of the historical simulations performed under CMIP6, which extend from 1979-2014. As this period is common amongst all the data sources, we use it for the bulk of our analysis. CMIP6 represents the current state-of-the-art in climate modeling, so most of the participating modeling groups have provided output from fully comprehensive earth-system models. It is important to stress that the historical simulations analyzed here were performed under identical greenhouse gas (GHG), aerosol, and natural forcings (Eyring et al. 2016). As for ozone, some models use prescribed concentrations (Checa-Garcia et al. 2018), while others include interactive chemistry schemes. As in the case of the ERA5 data, we have regridded the model output to the lower resolution grid of the radiosonde data sets, and applied the same missing data mask. A few CMIP6 models have missing data in the lower atmosphere as they do not interpolate below the ground level which, in some mountainous regions, is higher than the lower atmospheric pressure levels.

At the time of this writing, output from 48 models is available for the CMIP6 historical simulations, as listed in Table 1 (with ocean type 'C'). Unless otherwise specified, we take only the first ensemble member of each model as we use individual members as opposed to ensemble means for a like-with-like comparison with observations, and want to ensure equal weighting across the set of models. In addition to the atmosphere-ocean coupled simulations, we also make use of the atmosphere-only version of the historical CMIP6 simulations (see Table 1), which are forced with observed sea surface temperatures (SSTs). To put the CMIP6 models in the context of earlier intercomparisons, we also show results for the CMIP5 models (as listed in M13).

To quantify the relative importance of the major forcings, we also make use of the model output produced by the Detection and Attribution Model Inter-comparison Project (DAMIP, Gillett et al. 2016). At this time a total of 7 models (listed in Table 1) have made available the single-forcing simulations that we analyze here. Specifically, these are the historical 'GHG-only' simulations, forced only with well-mixed greenhouse gases, the 'aerosol-only' simulations, forced only with aerosols (BC, OC, SO_2, SO_4, NO_x, NH_3, CO, NMVOC), and the 'natural-only' simulations, forced only with solar irradiance changes and volcanic aerosols.

Finally, in order to quantify the contribution of internal variability to the spread across the CMIP6 models, we also analyse several "large ensembles" that were performed as part of the CMIP6 historical experiments. We define a large ensemble as having 20 or more members: this allows us to analyze six different large ensembles (see Table 1), ranging in size from 20 (GISS-E2-1-H) to 50 members (CanESM5). Large ensembles are also available for models other than those analyzed here (Deser et al. 2020), but we have chosen to focus on the models that participated in CMIP6 to ensure all models forcings are the same in this study.

Model	Ocean Type	Single Forcings	Large Ensemble Size
ACCESS-CM2	C		
ACCESS-ESM1-5	C/P		
AWI-CM-1-1-MR	C		
BCC-CSM2-MR	C/P	GHG, AER, NAT	
BCC-ESM1	C/P		
CAMS-CSM1-0	C/P		
CanESM5	C	GHG, AER, NAT	50
CESM2	C/P		
CESM2-FV2	C		
CESM2-WACCM	C/P		
CESM2-WACCM-FV2	C		
CIESM	C		
CNRM-CM6-1	C/P	GHG, AER, NAT	29
CNRM-CM6-1-HR	C/P		
CNRM-ESM2-1	C/P		
E3SM-1-0	C		
E3SM-1-1	C		
EC-Earth3	C/P		
EC-Earth3-Veg	C/P		
FGOALS-f3-L	C/P		
FGOALS-g3	C		
FIO-ESM-2-0	C/P		
GFDL-CM4	C/P		
GFDL-ESM4	C/P		
GISS-E2-1-G	C	GHG, AER, NAT	27
GISS-E2-1-G-CC	C		
GISS-E2-1-H	C		20
INM-CM4-8	C		
INM-CM5-0	C		
HadGEM3-GC31-LL	C/P	GHG, AER, NAT	
HadGEM3-GC31-MM	C/P		
INM-CM4-8	C/P		
INM-CM5-0	C/P		
IPSL-CM6A-LR	C/P	GHG, AER, NAT	32
KACE-1-0-G	C		
MIROC6	C/P		
MIROC-ES2L	C		
MPI-ESM-1-2-HAM	C		
MPI-ESM1-2-HR	C/P		
MPI-ESM1-2-LR	C		
MRI-ESM2-0	C/P	GHG, AER, NAT	
NESM3	C/P		
NorCPM1	C/P		30
NorESM2-LM	C/P		
NorESM2-MM	C		
SAM0-UNICON	C/P		
TaiESM1	C		
UKESM1-0-LL	C/P		

Table 1. The CMIP6 models analyzed in this study. C indicates models with a fully-coupled dynamic ocean; P indicates atmosphere only models with prescribed SST; C/P models for which both simulations are available. For the single forcing simulations, GHG refers to greenhouse gas only forcings; AER refers to aerosol only forcings, and NAT refers to natural only forcings (see Gillett et al. 2016, for details)

3. Analysis

In light of the most recent advances in Earth-system modeling and of the improved observational data sets available, we begin by updating the result of M13, and present the vertical profile of zonal mean, annual mean temperature trend from 1979 to 2014. As shown in Fig. 1a, the overall trends consist of a cooling of the stratosphere and a warming of the troposphere, in both models and observations. This pattern is the well-known vertical "fingerprint" of anthropogenic forcings, originally reported by Tett et al. (1996) and Santer et al. (1996). In the stratosphere, the coupled CMIP6 models (red bars) show cooling trends comparable to the observed ones (black lines). In the troposphere, however, the models show considerably larger warming than in the observations.

The warm trends bias in the models is seen throughout the entire troposphere, but is greatest in the upper troposphere (peaking around 200 hPa), where the modeled trends are – on average – 4 to 5 times greater than the observations. We draw attention to the CanESM5 model: it simulates the greatest warming in the troposphere, roughly 7 times larger than the observed trends. We note this model is known to have a high climate sensitivity compared to others (Swart et al. 2019, Forster et al. 2019). Throughout the depth of the troposphere, not a single model realization overlaps all the observational estimates. However, there is some overlap between the RICH observations and the lowermost modelled trend, which corresponds to the NorCPM1 model.

In M13, this considerable bias was attributed to the inability of the models to capture the observed sea surface temperature trends. The same applies to the CMIP6 models, as demonstrated by fact that when the models are forced with prescribed SSTs (blue bars in Fig. 1a) their trends are much closer to the observed values. Nonetheless, one can still see a systematic bias at most tropospheric levels. Between 200 and 100 hPa the differences between the CMIP and AMIP simulations are even more visible. It is important to note that ERA5/5.1 is warmer than the radiosondes in that region; this is likely due to the assimilation of radio occultation data, which shows more warming in the upper troposphere than the radiosondes. As such, the discrepancy in this region may be smaller than reported in previous studies, and very possibly due to observational uncertainty, rather than model biases. A comparison with trends that extend to 2019 is given in Figure S2, with no change in these conclusions.

Now, turning our attention to lower stratospheric trends (100-20 hPa), one may be tempted to conclude – from Fig. 1a – that modeled and observed trends are in good agreement. The story, however, is more complex, and requires a more nuanced analysis.

Recall that, unlike carbon dioxide which has been monotonically increasing since the pre-industrial era, ozone depleting substances, an important and often neglected anthropogenic forcing, exhibit a highly-nonlinear evolution from 1979 to 2014: the usefulness of a single linear trend covering the entire period, therefore, is questionable. The nonlinearity is due to the signing of the Montreal Protocol in 1989: as a consequence of that treaty, the atmospheric concentrations of many ozone depleting substances are no longer increasing. In fact, the trend in "effective equivalent stratospheric chlorine" (EESC, a commonly used metric for the combined concentration of ozone depleting substances) has changed from positive to negative around the very end of the 20th century.

In view of this, following the latest Scientific Assessment of Ozone Depletion (WMO 2018), we split the 1979-2014 period into two parts: the ozone depletion period 1979-1997 (during which EESC was increasing) and the ozone recovery period 1998-2014 (during which EESC was in decline). Separate temperature trends for these two periods are shown in Figs. 1b and c, respectively. It must be emphasized that these two periods are relatively short (less than two decades): hence much caution is called for in any analysis and interpretation.

Let us start by considering the observations. It is clear that the stratospheric cooling trends in the ozone-depletion period are greatly reduced in the ozone-recovery

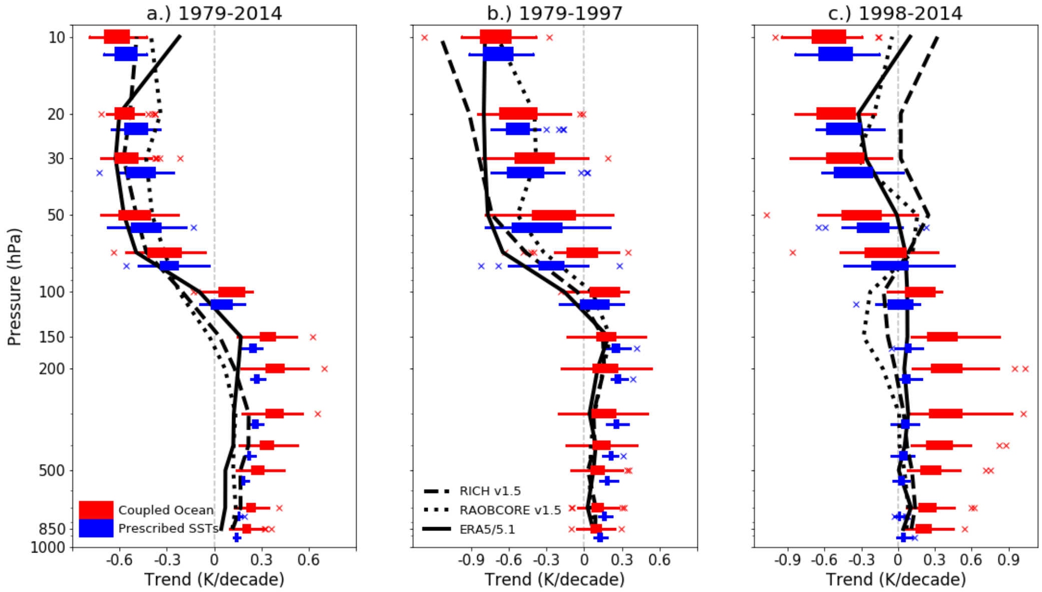

Figure 1. Vertical profiles of tropical (20S-20N) temperature trends for the period 1979-2014. The black lines show the RICH1.5 and RAOBCORE1.5 radiosondes, and ERA5/5.1 reanalysis. The red box-and-whisker bars show trends for ocean-atmosphere coupled CMIP6 models (48 in total); the blue bars shows trends for CMIP6 models with prescribed sea surface temperature (28 in total); the red bars are plotted at the correct altitude, but the blue bars are slightly offset downwards to aid comparison; each box shows the lower-to-upper quartile of the modeled trends, and the whiskers show the full range of data up to 1.5 times the inter-quartile range away from the mean, in which case the points beyond are represented by coloured crosses. The model data and ERA5/5.1 data are masked with the same observational mask from RICH, including the variation in time and pressure of the mask. Monthly data are averaged to annual data; if more than 3 months of data are missing in any grid box in a given year, all months for that year are set to missing. Panel a.), b.) and c.) show trends from 1979-2014, 1979-1997 (ozone depletion era), and 1998-2014 (ozone recovery era), respectively. The reanalysis line (solid black line) is constructed using ERA5 from 1979-2000, ERA5.1 from 2000-2006, and ERA5 from 2006-2014.

period. The RICH radiosonde data, in fact, indicates that stratospheric cooling trends have disappeared after 1998, although RAOBCORE and ERA5/5.1 still show a modest cooling. Results over this period are in agreement with the satellite observations, which show a completely flat temperature time series after 1996 in the lower stratospheric channel (the so-called MSU Channel 4), as reported in Mitchell (2016), Seidel et al. (2016). It has been proposed that the near disappearance of cooling trends in the lower stratosphere is a simple consequence of the fact that ozone depletion is no longer occurring (see, e.g., Fig. 3.21 of WMO 2018). Some studies have also pointed to a role for SSTs in recent tropical lower stratospheric temperature trends (e.g Shangguan et al. 2019), although modelling results indicate that this effect is small (Polvani & Solomon

2012).

Turning now to the modeled trends, Figs. 1b and c reveal a considerable discrepancy with the observations. In the stratosphere, the majority of CMIP6 models cool too little in the ozone depletion period when compared with RICH and ERA5, although there is good agreement with RAOBCORE. For the ozone recovery period the models all cool too much, with the inter-quartile range not encompassing any observational product, and the total range not encompassing RICH at all. We suspect that these temperature biases might be due to a poor representation of stratospheric ozone forcing in the CMIP6 models. To this date, the methodology used to construct the ozone forcing for CMIP6 remains undocumented in the peer-reviewed literature, although Checa-Garcia et al. (2018) have shown considerable uncertainty in the radiative forcing associated with ozone in the CMIP6 models. We have also not explored whether biases in stratospheric temperature trends are smaller for CMIP6 models with interactive ozone chemistry. It is also possible that the CMIP6 biases in stratospheric temperature trends stem from other sources, e.g. circulation changes that are inaccurately simulated in models (e.g. Garfinkel et al. 2013). We note, however, that models with a realistic simulation of stratospheric ozone, and a good vertical resolution in the stratosphere, are perfectly capable of reproducing the observed stratospheric trends between 100 and 20 hPa over both periods separately (see, e.g., Fig. A3 of Randel et al. 2017).

It is also instructive to contrast the tropospheric temperature trends in the ozone-depletion and ozone recovery period. Forster et al. (2007) – on the basis of a purely radiative calculation with a fixed dynamical heating assumption – suggested that ozone depletion in the tropical stratosphere may lead to cooling in the tropical upper troposphere, due to a reduction in downwelling longwave radiation from the ozone above. However, using an atmospheric general circulation model with prescribed ozone concentrations, Polvani & Solomon (2012) showed that effects of stratospheric ozone depletion on tropical temperature trends do not extend much below the 100 hPa level. Given the observational uncertainties, it is difficult to discern a significant difference between Figs. 1b and 1c in the observed tropospheric trends.

As for the modeled tropospheric trends, however, the discrepancy with observations is much larger after 1998. We suspect that one cause of this discrepancy is related to the fact that the 1998-2014 period corresponds to the occurrence of the so called “global warming hiatus” (see Fyfe et al. 2016, for a recent update of this debate). If the hiatus is indeed related to an increased heat uptake by the oceans, as suggested by some studies (Meehl et al. 2011, 2014), it cannot be considered an externally forced process: therefore, one would not expect it to be captured in the models over the same time period. Another contributing factor is that 1997/1998 had one of the largest El Niño events on record, which, given the short period the trend is calculated over, becomes important. Indeed, if the analysis is repeated for the 1999-2014 period (i.e. missing the large El Niño year), the tropospheric observational trends are higher, and in better agreement with the coupled model estimates (Figure S3). Be that as it may, we note here for the record that from 1998 to 2014, the CMIP5 models warm, on average 4 to

5 times faster than the observations, and in one model the warming is 10 times larger than the observations.

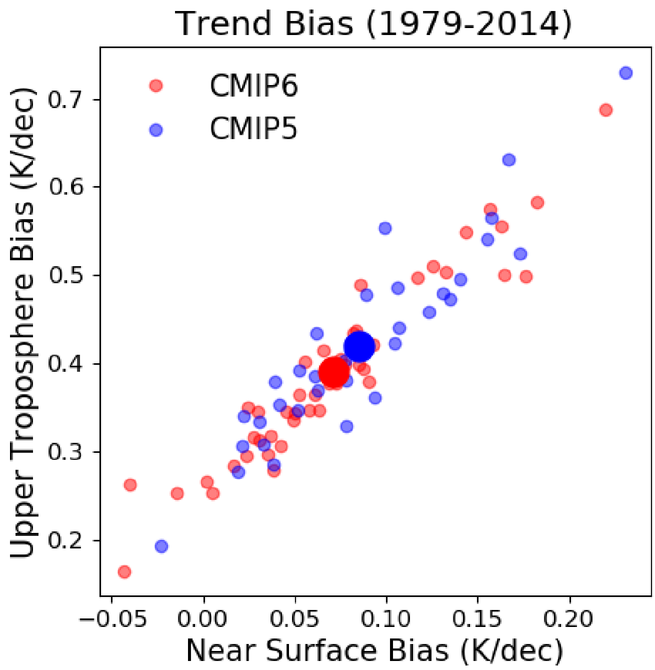

Figure 2. Upper tropospheric biases (at 200 hPa) vs. near surface biases (at 850 hPa) in the historical simulations of the coupled CMIP6 (red) and CMIP5 (blue) models. The discrepancy (or bias) is defined as the 1979-2014 trend difference between each model and the RICH v1.5 radiosonde value as the same level. The larger circles show the CMIP5 and CMIP6 multi-model mean. CMIP5 data have had RCP4.5 simulations added on to bring the end date to 2014.

To better quantify the relationship between the near surface and the upper tropospheric biases, which was already noted in M13, we illustrate their correlation in Fig. 2. For the CMIP6 models (red dots) the upper tropospheric (200 hPa) bias is very highly correlated with the near surface (850 hPa) bias, over 1979-2014: the Spearman correlation coefficient is $r = 0.95$. A similar number, $r = 0.91$, is calculated for the older CMIP5 models (blue), and the multi-model means are very close too. This indicates that there has not been any substantial improvement, in terms of tropical tropospheric temperature trends, between CMIP5 and CMIP6.

Next, we examine the source of the large spread in tropical temperature trends across the CMIP6 models. In particular, we examine separately the forced response and the internal variability. Starting with the former, the impacts of the different forcings on tropical atmospheric temperature trends is studied by analyzing the single-forced experiments that have been carried out under the Detection and Attribution Model Inter-comparison Project (DAMIP, Gillett et al. 2016). Specifically, we make use of three separate experiments: the GHG-only simulations, the aerosol-only simulations, and the natural-only simulations (for more details, see Section 2 of this paper, or Gillett et al. 2016).

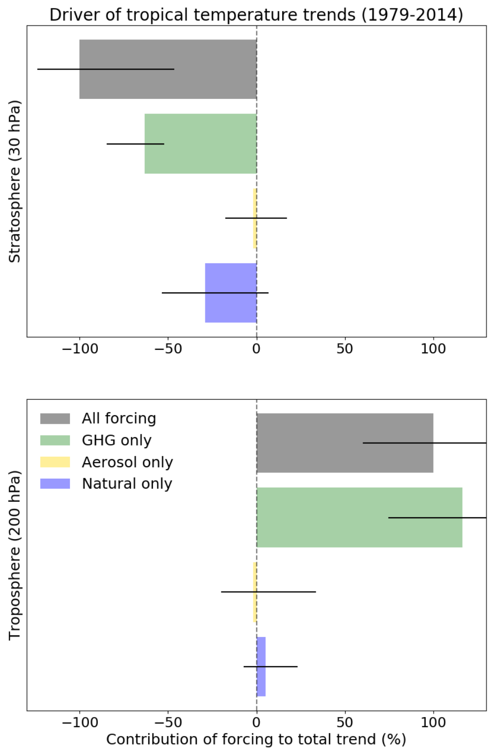

Figure 3. Percentage contribution of individual forcings to the total trend for the (top) mid-lower stratosphere identified as the 30 hPa level, and (bottom) upper troposphere identified as the 200 hPa level. Only a subset of CMIP6 models which include the single forcing simulations are used (see Table 1). The bars represent the multi-model mean contribution to the trend, normalized by the total trend in the historical (All forcing) simulations; the grey bars are equal to 100%, by definition. The horizontal black lines show the individual model spread of the ensemble means, again, normalized by the gray bar. Positive/negative values represent warming/cooling.

In Fig. 3 we illustrate how single forcings contribute to the total modeled trends, in both the stratosphere (30 hPa, top panel) and upper troposphere (200 hPa, bottom panel). By definition, the total trend (gray bars) is equal to 100%. Recall that, owing to data availability, only a subset of the CMIP6 models in Fig. 1 are used for this analysis (see Table 1). The stratospheric cooling trend is dominated by the GHG forcings, but also with a sizable component coming from natural forcings, most likely a cooling trend from volcanic emissions. Stratospheric ozone is prescribed to a pre-industrial climatology in these single forcing simulations, so cooling from ozone depletion is only present in the all-forcing (i.e. historical) simulations, and cannot be separately estimated using these specific DAMIP simulations.

In the upper troposphere (Fig. 3, bottom panel) GHGs are the overwhelming driver of temperature trends, with negligible contributions from aerosols and natural forcings. For the aerosol forcing, we note that only one model (MRI-ESM2-0) shows warming, and this warming trend skews the results considerably, providing a large positive error bar. Without that one model, the aerosol cooling is more substantial at this height. Needless to say, the ensemble size (7) is relatively small, and we hope more models will soon become available. Also, we note that these single-forcing simulations are not expected to sum to 100%, i.e. the sum of the green, yellow and blue bars to equal the gray bar, because 1) other forcings may be important, e.g. tropospheric and stratospheric ozone, 2) there may exist some non-linear interactions between different forcings, and 3) one cannot precisely estimate the forced signal with these DAMIP runs since a large ensemble of single-forcings simulations for each model is not available (the ensemble means would be estimates of the forced signal in each model). So, the results in Fig. 3 are contaminated by internal variability.

However, internal variability can be estimated by exploiting the fact that six CMIP6 models have made available large ensembles of historical integrations (see Table 1). In each panel in Fig. 4 we plot the upper tropospheric vs. the near the surface temperature trends for two sets of runs: one set consists of the the first simulations of each of the 48 different CMIP6 model (red dots), and the other set consists of all members of each of the 6 models with historical large ensembles (blue circles). The crosses of the corresponding colour indicate mean of each set, and the accompanying dashed lines show the accompanying linear regression. The observations are shown with black symbols.

Two theoretical lines are also plotted in each panel in Fig. 4: the dry and moist adiabatic lapse rates (DALR and MALR, respectively), plotted as black lines. The MALR is computed using the following approximation (taken from Bakhshaii & Stull 2013)

$$\frac{\mathrm{d}T}{\mathrm{d}p} = \left(\frac{1}{p}\right) \frac{R_d T + L_v r_s}{c_{pd} + \frac{L_v^2 r_s \epsilon}{R_d T^2}} \tag{1}$$

where c_{pd} is the specific heat capacity for dry air at constant pressure, R_d is the gas constant for dry air, L_v is the latent heat of vaporisation, r_s is the saturation mixing ratio, and $\epsilon = R_d/R_v$ is the ratio of ideal gas constants for dry air and water vapor. The MALR profiles are calculated by integrating this equation vertically, starting at 850 hPa, and using $T(850\ \text{hPa}) = 291$ K, which we take from the ERA5 reanalysis. The DALR is obtained from the same formula, setting $r_s = 0$, which the reduces to more common $\mathrm{d}T/\mathrm{d}z = g/c_{pd}$.

Several interesting points should be noted in Fig. 4. First, there is a very strong correlation between the near surface and upper tropospheric trends, in all seven of the sets of models/ensembles: this confirms that the spread in upper tropospheric warming trends, in all cases, can be traced back to the spread in surface temperature. Second, the regression curves are close but not coincident with the theoretical moist

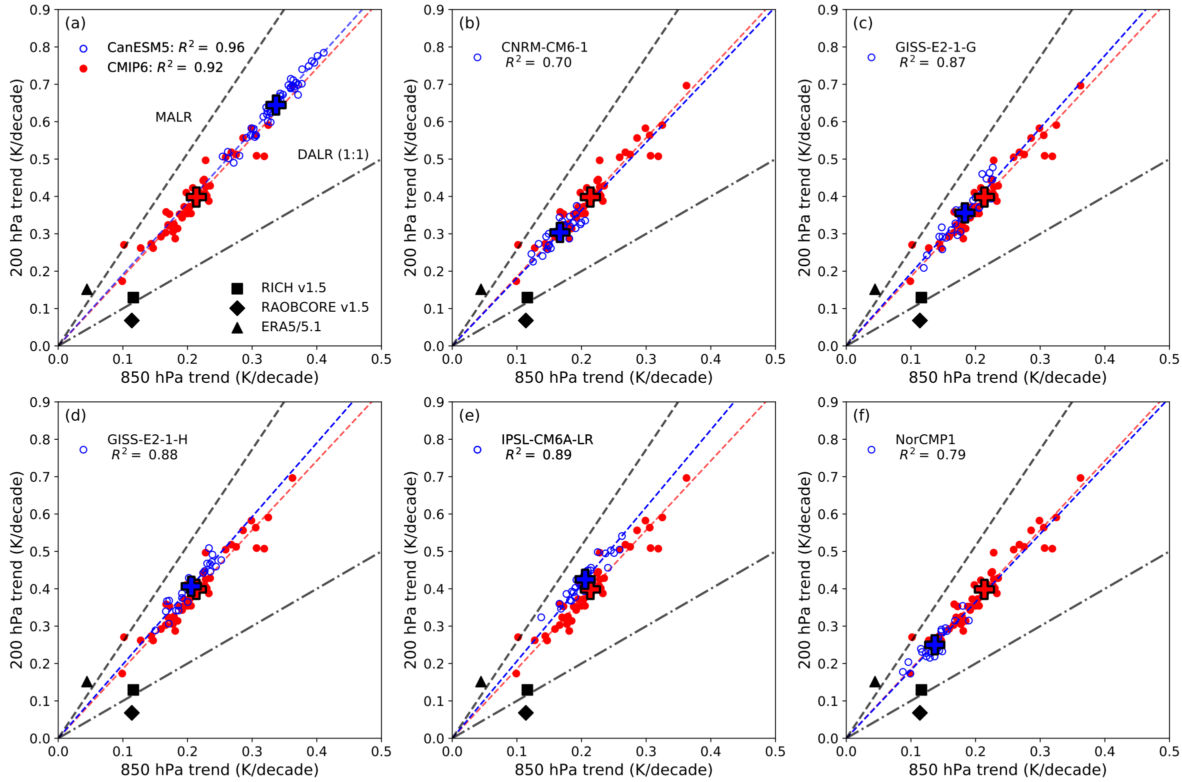

Figure 4. Regression between tropical (20S-20N) temperature trends (1979-2014) at 200 hPa and 850 hPa for the first historical simulation of each CMIP6 model (red dots) and each of the 6 models with large ensembles of historical simulations (blue circles), as well as 3 observationally-derived trends (black symbols). Crosses show ensemble means. Linear least-squares regression lines are also shown for both sets, with the corresponding R^2 given in the legend. The black dashed and dot-dashed lines show the relationships given by the moist and dry adiabatic lapse rates, respectively.

adiabatic line: this indicates that the popular idea that tropical temperature profiles follow moist adiabats may not be quantitatively correct at these levels, at least not for the temporal- spatial-averaged sea surface temperature response considered here. For instance, Flannaghan et al. (2014) show that tropical temperature trends only follow a moist adiabat once you appropriately weight the near surface temperature trends toward regions of deep convection, since it is the deep convecting regions that ultimately influence the upper troposphere. Third, contrasting the red dots and blue circles one gets the distinct impression that the spread across each large ensembles is comparable to the spread across the entire CMIP6 set. This is especially clear for CanESM5, which provided 50 distinct runs of the same model (panel a), and suggests that a large fraction of the CMIP6 spread may actually come from internal variability. Finally, we note that the means of the large ensembles (blue plus symbols), which represent the forced trends in each model, can be found at both ends of the CMIP6 range (red dots): contrast, for instance, panels a and f. This indicates that the spread in forced trends across the models can be almost as large as the range spanned by the CMIP6 models.

In order to more clearly illustrate this point, i.e. to quantitatively compare the spread of the entire vertical profile of temperature trends across both the CMIP6 ensemble and the large ensembles, some thought is required. As seen in Fig. 4, the large ensembles have an average surface warming which is often different from the CMIP6 set. This implies that the lapse rates are higher for the large ensembles with higher surface warming than CMIP6, notably CanESM5, and lower for the large ensembles with lower surface warming (e.g. NorCPM1). Thus, some rescaling is needed for a meaningful comparison. Exploiting the tropospheric lapse rates across the large ensembles (the blue dashed lines in Fig. 4 panels a-f), we construct the relationship

$$T_t(n,p) = \alpha(p)T_t(n, 850\text{ hPa}) + c(p), \tag{2}$$

where $T_t(n,p)$ is the temperature trend at level p for large ensemble member n, and the values of $\alpha(p)$ and $c(p)$ are derived from linear regression across ensemble members at each level p. Above 200 hPa the regression is not strong, so we do not apply this trend scaling beyond that level. Now, to quantitatively compare the spread in trends across the large ensembles and across the CMIP6 ensemble, we scale the individual large ensemble members so that their ensemble mean, at 850 hPa, is the same as the CMIP6 ensemble. The scaled trends $T'_t(n,p)$ are defined by the expression:

$$T'_t(n,p) = \alpha(p)(T_t(n, 850\text{ hPa} - O) + c(z), \tag{3}$$

where $O = \langle T_t(n, 850)\rangle_{\text{Large-ensemble}} - \langle T_t(n, 850)\rangle_{\text{CMIP6}}$ is the difference between the ensemble means at 850 hPa.

Fig. 5 shows the scaled spread, as per Eqn. 3, in the CMIP6 models (red) and each of the large ensembles (blue) for two different pressure surfaces. To be clear: the red boxes here are identical to those in Fig. 1a at 850 and 200 hPa. The mean trend for each large ensemble at 850 hPa is, by construction, identical to the mean of the CMIP6 ensemble. The standard deviation of the scaled CanESM5 ensemble (lightest blue) encompasses ∼70% of the CMIP6 range, whereas for CNRM-CM6-1 (second lightest blue) it only encompasses ∼30%. All the other large ensembles are found somewhere between these extremes and, on average, the large ensembles explain ∼50% of the total CMIP6 variability. Note that the number of models (or ensemble members) in each spread is different. To test if this matters we repeat our analysis with only 20 samples for each of datasets (the lowest common denominator), but our results remain similar, and so we conclude there is little sensitivity to sample sizes greater than 20. Given this result, the clear indication here is that internal variability may be responsible for around 50% of the CMIP6 standard deviation, at least for trends over intervals spanning 3-4 decades (in our case, the trends are 35 years long). Finally, we note that while the large ensemble spreads are approximately Gaussian, the spread in CMIP6 models has a heavy upper tail, in line with the skewed range of climate sensitivities within this ensemble (Forster et al. 2019).

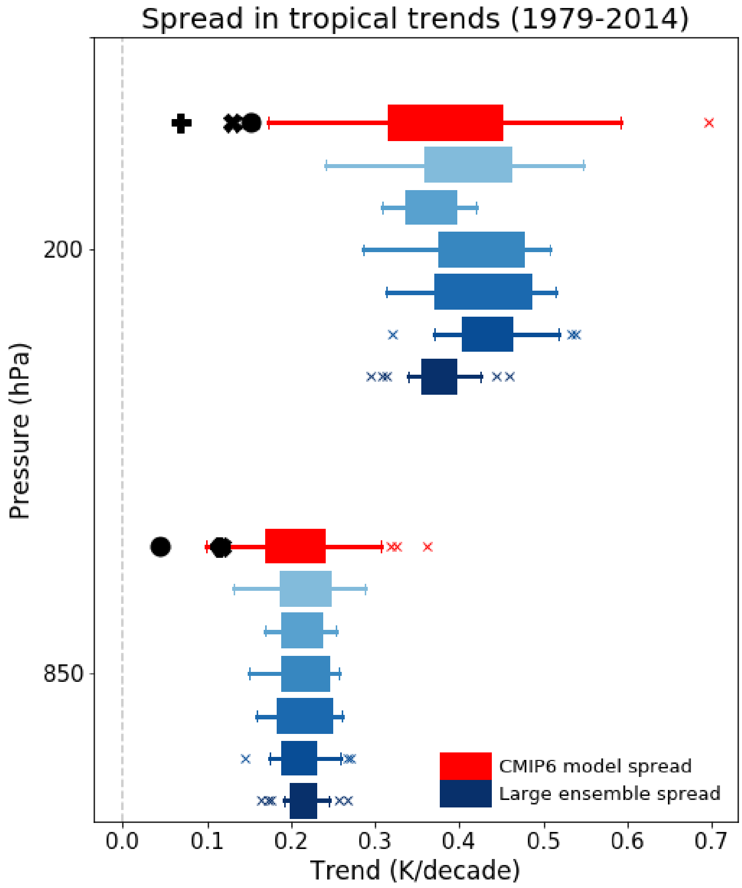

Figure 5. A comparison of the spread in tropical temperature trends for the CMIP6 models (red) and individual large ensembles (blue) at two different pressure levels, 850 hPa and 200 hPa. Each shade of blue represents a different large ensemble. From light-blue to dark-blue they are: CanESM5, CNRM-CM6-1, GISS-E2-1-G, GISS-E2-1-H, IPSL-CM6A-LR and NorCPM1 (see Table 1). The large ensembles have been scaled so as to be centered on the CMIP6 ensemble profile, see text for details. The observations at these two levels are marked by a black cross (RICH), plus (RAOBCORE) and circle (ERA5/5.1). Note that at 850 hPa RICH and RAOBCORE overlap. The box-and-whiskers display the same statistics as in Figure 1.

4. Conclusions

We have compared the modeled and observed tropical temperature trends, over the period 1979-2014, from 850 hPa to the mid-stratosphere. Focusing on the CMIP6 models, we have confirmed the original findings of Mitchell et al. (2013): first, the modeled tropospheric trends are biased warm throughout the troposphere (and notably in the upper troposphere, around 200 hPa) and, second, that these biases can be linked to biases in surface warming. As such, we see no improvement between the CMIP5 and the CMIP6 models.

In addition, we have here uncovered substantial model biases in tropical stratospheric trends. From 100 to 20 hPa (the lower to middle stratosphere), the CMIP6 models do not simulate the observed cooling during the ozone depletion period (1979-1997) compared with 2 of the 3 observational products used, and then simulate too much cooling in the ozone recovery period (1998-2014) compared with all observational products. Unfortunately, these biases cancel when one computes a single trend over the entire 1979-2014 period, giving the impression that the CMIP6 simulations of stratospheric temperature are accurate. We stress the importance of computing separate trends before and after the year 1998, which has become common practice in recent Ozone Assessment Reports (see, e.g. WMO 2018), as the forcing from ozone and halocarbons is not monotonic owing to the signing of the Montreal Protocol in 1989.

Finally, analyzing six CMIP6 models which provided relatively large ensembles (from 20 to 50 members), we have been able to quantify the fraction of the CMIP6 model spread due to internal variability, as opposed to model differences. We find that the standard deviation of the large ensembles, which is due to internal variability alone, is 30-70% of that of CMIP6 models for the period 1979-2014, with a central estimate of 50%. This result highlights the importance of using large ensembles when evaluating trend differences across the CMIP6 models.

Acknowledgments

We thank the editor and two anonymous reviewers for fast and in-depth reviews, Bill Bell from ECMWF for providing us with early access to the ERA5.1 data, and all modelling centres that produced data as part of CMIP6. DM was funded by a NERC fellowship (NE/N014057/1). EL was funded under the NERC HAPPI-Health project (NE/R009554/1). WJMS was funded by a University of Bristol Vice Chancellor's fellowship. LMP is funded, in part, by an award (#1914569) from the US National Science Foundation to Columbia University.

Data Availability

The data that support the findings of this study are openly available at $https : //esgf - index1.ceda.ac.uk/projects/cmip6 - ceda/$. ERA5 data are available from ECMWF. Radiosonde data are available from Leopold Haimberger. Our code is freely available at $https : //github.com/BrisClim/$.

417 ## Supplementary Information

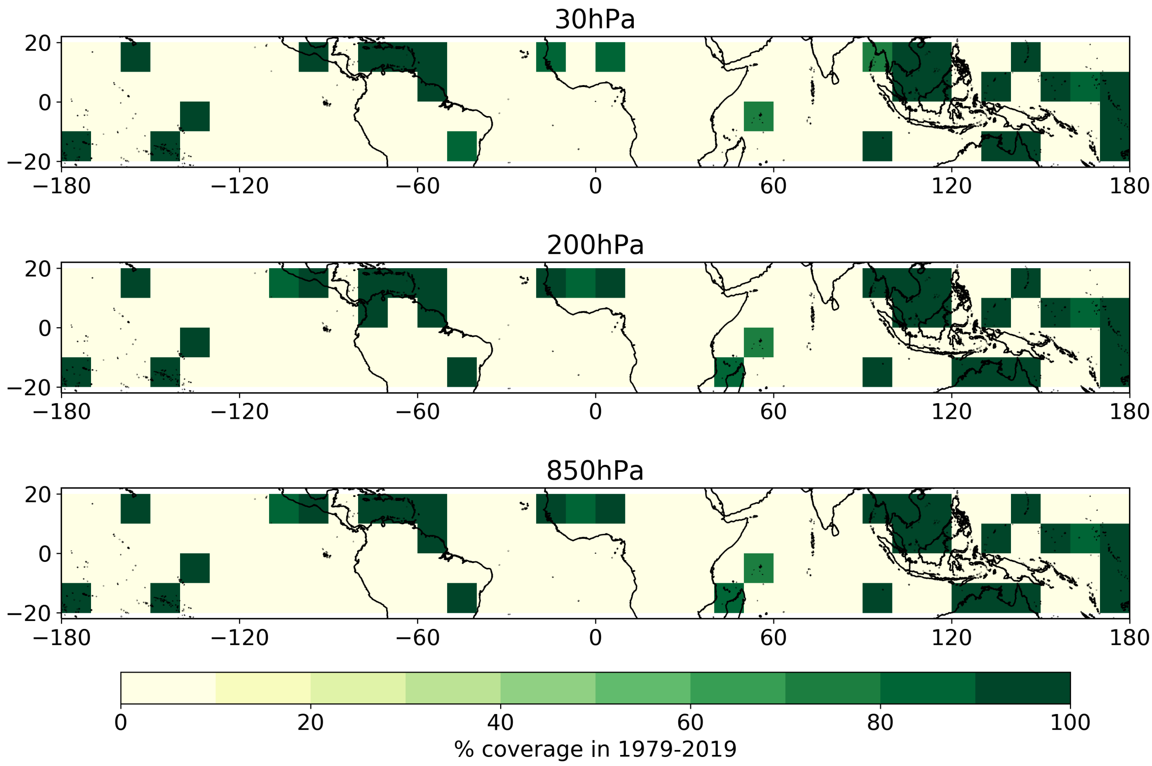

Figure S1. Percentage of months between 1979–2019 that RAOBCORE v1.5 data are available in the tropics at 850, 200 and 30 hPa. All months in the years that have more than 3 months of data missing are set to missing.

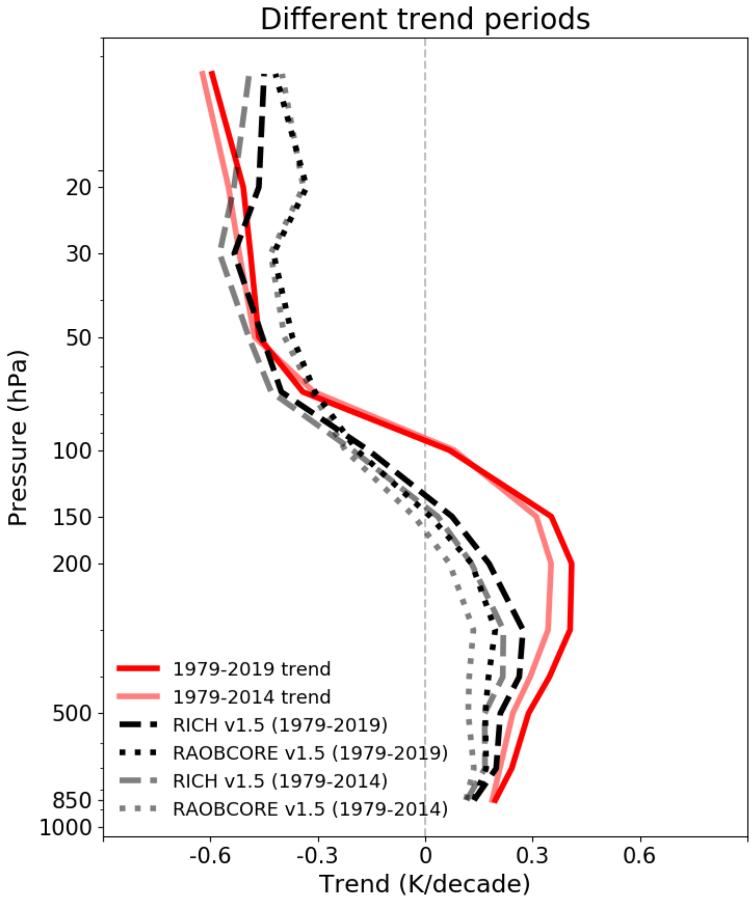

Figure S2. As in Figure 1a but only showing the mean response of the CMIP6 models (reds). The darker lines show the trend calculated from 1979-2019, and the lighted lines show it calculated from 1979-2014 (as in the main body of the paper). CMIP6 historical data end in 2014, so SSP2RCP4.5 data are added to allow for trends to be calculated to 2019. Note we do not currently have the ERA5 data for the latter part of 2019, so it is excluded from this figure.

References

Bakhshaii, A. & Stull, R. (2013), 'Saturated pseudoadiabats – A noniterative approximation', *Journal of Applied Meteorology and Climatology* **52**(1), 5–15.

Checa-Garcia, R., Hegglin, M. I., Kinnison, D., Plummer, D. A. & Shine, K. P. (2018), 'Historical tropospheric and stratospheric ozone radiative forcing using the cmip6 database', *Geophysical Research Letters* **45**(7), 3264–3273.

Deser, C., Lehner, F., Rodgers, K., Ault, T., Delworth, T., DiNezio, P., Fiore, A., Frankignoul, C., Fyfe, J., Horton, D. et al. (2020), 'Insights from earth system model

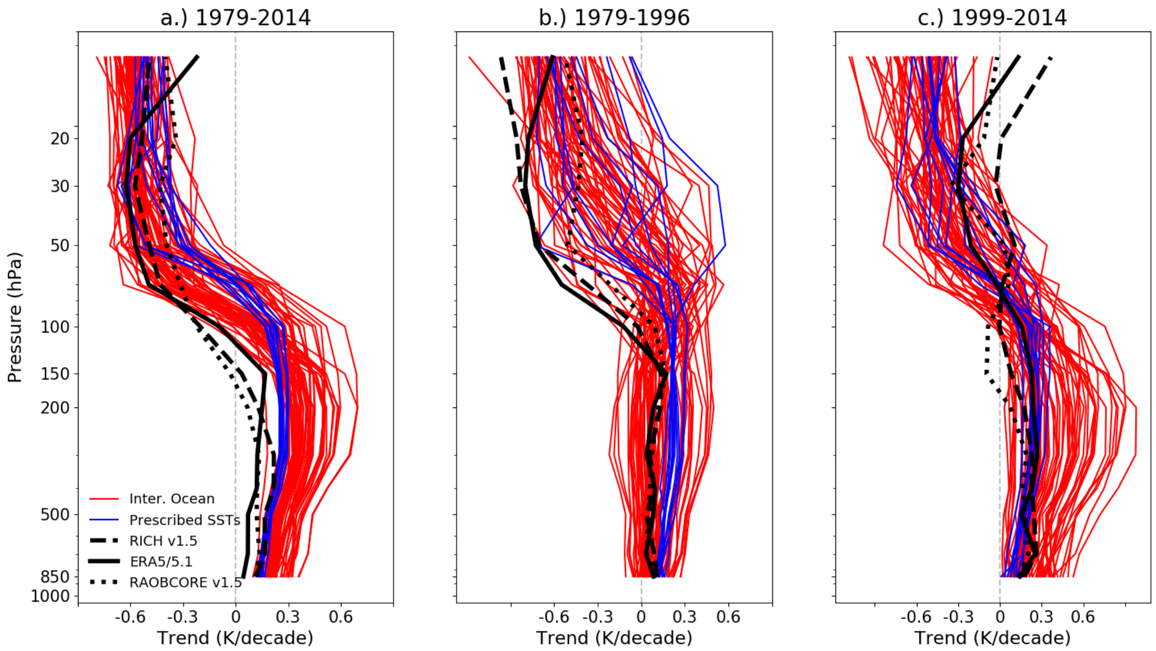

Figure S3. As in Figure 1a but with different trend averaging periods for panel b and c.

initial-condition large ensembles and future prospects', *Nature Climate Change* pp. 1–10.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J. & Taylor, K. E. (2016), 'Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6): Experimental design and organization', *Geoscientific Model Development (Online)* **9**(LLNL-JRNL-736881).

Flannaghan, T., Fueglistaler, S., Held, I. M., Po-Chedley, S., Wyman, B. & Zhao, M. (2014), 'Tropical temperature trends in atmospheric general circulation model simulations and the impact of uncertainties in observed ssts', *Journal of Geophysical Research: Atmospheres* **119**(23), 13–327.

Forster, P. M., Bodeker, G., Schofield, R., Solomon, S. & Thompson, D. (2007), 'Effects of ozone cooling in the tropical lower stratosphere and upper troposphere', *Geophysical Research Letters* **34**(23).

Forster, P. M., Maycock, A. C., McKenna, C. M. & Smith, C. J. (2019), 'Latest climate models confirm need for urgent mitigation', *Nature Climate Change* pp. 1–4.

Fu, Q., Manabe, S. & Johanson, C. M. (2011), 'On the warming in the tropical upper troposphere: Models versus observations', *Geophysical Research Letters* **38**(15).

Fyfe, J. C., Meehl, G. A., England, M. H., Mann, M. E., Santer, B. D., Flato, G. M., Hawkins, E., Gillett, N. P., Xie, S.-P., Kosaka, Y. et al. (2016), 'Making sense of the early-2000s warming slowdown', *Nature Climate Change* **6**(3), 224.

Garfinkel, C. I., Waugh, D. W. & Gerber, E. P. (2013), 'The effect of tropospheric jet latitude on coupling between the stratospheric polar vortex and the troposphere', *Journal of climate* **26**(6), 2077–2095.

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., Santer, B. D., Stone, D. & Tebaldi, C. (2016), 'Detection and attribution model intercomparison project (damip)', *Geoscientific Model Development* **9**(10), 3685–3697.

Haimberger, L. (2007), 'Homogenization of radiosonde temperature time series using innovation statistics', *Journal of Climate* **20**(7), 1377–1403.

Haimberger, L., Tavolato, C. & Sperka, S. (2012), 'Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations', *Journal of Climate* **25**(23), 8108–8131.

Hartmann, D. L., Tank, A. M. K., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y. A. R., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A. et al. (2013), Observations: atmosphere and surface, *in* 'Climate change 2013 the physical science basis: Working group I contribution to the fifth assessment report of the intergovernmental panel on climate change', Cambridge University Press, pp. 194–200.

Hawkins, E. & Sutton, R. (2009), 'The potential to narrow uncertainty in regional climate predictions', *Bulletin of the American Meteorological Society* **90**(8), 1095–1108.

Hegerl, G. C., Zwiers, F. W., Braconnot, P., Gillett, N. P., Luo, Y., Marengo Orsini, J., Nicholls, N., Penner, J. E. & Stott, P. A. (2007), Understanding and attributing climate change, *in* 'Climate change 2007 the physical science basis: Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change', Cambridge University Press, pp. 699–701.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. & Thépaut, J.-N. (2020), 'The ERA5 Global Reanalysis', *Quart. J. Roy. Meteorol. Soc.* **146**. in press.

Karl, T., Hassol, S., Miller, C. & (Eds.), W. M. (2006), 'Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences', *Synth. Assess. Prod. 1.1* . U.S. Clim. Change Sci. Program, Washington, DC.

Manabe, S. & Stouffer, R. J. (1980), 'Sensitivity of a global climate model to an increase of CO_2 concentration in the atmosphere', *Journal of Geophysical Research: Oceans* **85**(C10), 5529–5554.

Manabe, S. & Wetherald, R. T. (1975), 'The effects of doubling the CO_2 concentration on the climate of a general circulation model', *Journal of the Atmospheric Sciences* **32**(1), 3–15.

Meehl, G. A., Arblaster, J. M., Fasullo, J. T., Hu, A. & Trenberth, K. E. (2011), 'Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods', *Nature Climate Change* **1**(7), 360.

Meehl, G. A., Teng, H. & Arblaster, J. M. (2014), 'Climate model simulations of the observed early-2000s hiatus of global warming', *Nature Climate Change* **4**(10), 898.

Mitchell, D. (2016), 'Attributing the forced components of observed stratospheric temperature variability to external drivers', *Quarterly Journal of the Royal Meteorological Society* **142**(695), 1041–1047.

Mitchell, D., Thorne, P., Stott, P. & Gray, L. (2013), 'Revisiting the controversial issue of tropical tropospheric temperature trends', *Geophysical Research Letters* **40**(11), 2801–2806.

Po-Chedley, S. & Fu, Q. (2012), 'Discrepancies in tropical upper tropospheric warming between atmospheric circulation models and satellites', *Environmental Research Letters* **7**(4), 044018.

Polvani, L. M. & Solomon, S. (2012), 'The signature of ozone depletion on tropical temperature trends, as revealed by their seasonal cycle in model integrations with single forcings', *Journal of Geophysical Research: Atmospheres* **117**(D17).

Randel, W. J., Polvani, L., Wu, F., Kinnison, D. E., Zou, C.-Z. & Mears, C. (2017), 'Troposphere-stratosphere temperature trends derived from satellite data compared with ensemble simulations from WACCM', *Journal of Geophysical Research: Atmospheres* **122**(18), 9651–9667.

Santer, B. D., Fyfe, J. C., Pallotta, G., Flato, G. M., Meehl, G. A., England, M. H., Hawkins, E., Mann, M. E., Painter, J. F., Bonfils, C. et al. (2017), 'Causes of differences in model and satellite tropospheric warming rates', *Nature Geoscience* **10**(7), 478.

Santer, B. D., Painter, J. F., Mears, C. A., Doutriaux, C., Caldwell, P., Arblaster, J. M., Cameron-Smith, P. J., Gillett, N. P., Gleckler, P. J., Lanzante, J. et al. (2013), 'Identifying human influences on atmospheric temperature', *Proceedings of the National Academy of Sciences* **110**(1), 26–33.

Santer, B. D., Taylor, K., Wigley, T., Johns, T., Jones, P., Karoly, D., Mitchell, J., Oort, A., Penner, J., Ramaswamy, V. et al. (1996), 'A search for human influences on the thermal structure of the atmosphere', *Nature* **382**(6586), 39.

Santer, B. D., Wigley, T. M., Mears, C., Wentz, F. J., Klein, S. A., Seidel, D. J., Taylor, K. E., Thorne, P. W., Wehner, M. F., Gleckler, P. J. et al. (2005), 'Amplification of surface temperature trends and variability in the tropical atmosphere', *Science* **309**(5740), 1551–1556.

Seidel, D. J., Li, J., Mears, C., Moradi, I., Nash, J., Randel, W. J., Saunders, R., Thompson, D. W. & Zou, C.-Z. (2016), 'Stratospheric temperature changes during the satellite era', *Journal of Geophysical Research: Atmospheres* **121**(2), 664–681.

Shangguan, M., Wang, W. & Jin, S. (2019), 'Variability of temperature and ozone in the upper troposphere and lower stratosphere from multi-satellite observations and reanalysis data', *Atmospheric Chemistry and Physics* **19**(10), 6659–6679.

Simmons, A., Soci, C., Nicolas, J., Bell, B., Berrisford, P., Dragani, R., Flemming, J., Haimberger, L., Healy, S., Hersbach, H., Horányi, A., Inness, A., Munoz-Sabater, J., Radu, R. & Schepers, D. (2020), 'Global stratospheric temperature bias and other stratospheric aspects of era5 and era5.1', (859).
URL: *https://www.ecmwf.int/node/19362*

Spencer, R. W. & Christy, J. R. (1990), 'Precise monitoring of global temperature trends from satellites', *Science* **247**(4950), 1558–1562.

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S. et al. (2019), 'The Canadian Earth System Model version 5 (CanESM5. 0.3)', *Geoscientific Model Development* **12**(11), 4823–4873.

Tett, S. F., Mitchell, J. F., Parker, D. E. & Allen, M. R. (1996), 'Human influence on the atmospheric vertical temperature structure: Detection and observations', *Science* **274**(5290), 1170–1173.

Thorne, P. W., Lanzante, J. R., Peterson, T. C., Seidel, D. J. & Shine, K. P. (2011), 'Tropospheric temperature trends: History of an ongoing controversy', *Wiley Interdisciplinary Reviews: Climate Change* **2**(1), 66–88.

WMO (2018), 'World Meteorological Organization, Scientific Assessment of Ozone Depletion: 2018', pp. Geneva, Switzerland, 588 pp.