Understanding model-observation discrepancies in satellite retrievals of atmospheric temperature using GISS ModelE

Madeline C. Casas[1,2], Gavin A. Schmidt[1], Ron. L. Miller[1], Clara Orbe[1], Kostas Tsigaridis[1,3], Larissa S. Nazarenko[1,3], Susanne E. Bauer[1]and Drew T. Shindell[4]

[1]NASA Goddard Institute for Space Studies, New York, NY
[2]Stanford School of Humanities and Sciences, Palo Alto, CA
[3]Center for Climate System Research, Columbia University, New York, NY
[4]Duke University, Durham, NC

Key Points:

- Changes in forcing configurations have significant impact on agreement with satellite data.
- Tropospheric model-observation agreement is linked to ocean heat uptake, aerosols, sensitivity and internal variability.
- Stratospheric discrepancies are mainly related to volcanic aerosol modeling and representations of ozone climatology and trends.

Corresponding author: Gavin A. Schmidt, gavin.a.schmidt@nasa.gov

Abstract

We examine multiple factors in the representation of satellite-retrieved atmospheric temperature diagnostics in historical simulations of climate change during the satellite era (specifically 1979–2021) using GISS ModelE contributions to the Coupled Model Intercomparison Project (Phase 6) (CMIP6). The tropospheric and stratospheric trends in these diagnostics are affected by greenhouse gases (notably carbon dioxide and ozone), coupling with the ocean, volcanic aerosols, solar activity and compositional and dynamic feedbacks. We explore the impacts of internal variability, changing forcing specifications, composition interactivity, the quality of the stratospheric circulation, vertical resolution, and possible impacts of the mis-specification of volcanic aerosol optical depths. Overall temperature trends throughout the satellite period are well captured, but discrepancies at all levels exist and have multiple distinct causes. We find that stratospheric comparisons (using Stratospheric Sounding Unit (SSU) retrievals and successor instruments) are most affected by variations in the representation of ozone depletion and feedbacks, followed by the volcanic signals. Tropospheric skill (using the Microwave Sounding Unit (MSU) retrievals) is affected by the trends in ocean heat uptake and tropospheric aerosols, but also by the representation of stratospheric processes through the impact of the Brewer-Dobson circulation on the height of the tropical tropopause. We demonstrate that no single factor is the dominant cause of the discrepancies and that almost all observations lie within the broad envelope of structural uncertainty.

Plain Language Summary

The assessment of the ability of climate models to match trends and variability in the real world is a key factor in building the credibility of model projections under future scenarios. We focus on satellite-derived atmosphere temperatures from the surface to the stratosphere whose trends at different levels reflect different processes and drivers. The satellite retrievals are weighted averages of atmospheric temperatures and so the vertical structure of the model trends matter in the comparison. We find that overall the trends throughout the atmosphere are well-captured by the GISS models but that discrepancies can occur due to misspecified forcings, chaotic weather variability, and model structure.

1 Introduction

The start of the satellite period (nominally 1979) marked the dawn of a new era in global multi-variate monitoring of climate. Over 40 years of data have been collected since then and has been sufficient not only to refine our knowledge of the Earth's climatology, but also to capture the trends of a changing climate. One suite of important variables is the set of vertically-weighted atmospheric temperature changes seen by the Microwave Sounding Units (MSUs) (Spencer & Christy, 1990; Mears et al., 2003), Stratospheric Sounding Units (SSUs) (Thompson et al., 2012) and their successors, the Advanced Microwave Sounding Units (AMSUs).

The vertical pattern of temperature change in response to increased greenhouse gases has been recognised as a distinct fingerprint since the pioneering work of Manabe and Wetherald (1967). The surface warming, enhanced tropospheric warming, and stratospheric (and above) cooling is unlike the pattern generated by increasing solar activity (which would have more uniform warming through the whole atmosphere), ozone depletion, volcanic activity, or ocean-driven internal variability. However, it wasn't until the satellite era and the development of global atmospheric retrievals using the MSU/SSU/AMSU series of instruments combined with the radiosonde record that the ability to distinguish these vertical fingerprints emerged (Spencer & Christy, 1990; Randel & Cobb, 1994; Santer et al., 1996; Ramaswamy et al., 1996).

The clear trends near the surface and in both the troposphere and stratosphere have thus long been used in detailed comparisons to climate model simulations. Those comparisons have led to the discovery of discrepancies between the satellite retrievals, surface temperatures and models, and therefore understanding them has been a major focus of scientific attention (e.g. Christy & Spencer, 1995; Jones et al., 1997; Hansen et al., 1995; CCSP, 2006; Thorne et al., 2011).

The reasons for any climate model's mismatch to these trends can arise from multiple factors: errors in the model physics; model drivers; observational retrieval errors; or simply from an inappropriate comparison, and all of these possible effects have been encountered over time with respect to the MSU/SSU/AMSU time-series (for instance, Wentz & Schabel, 1998; Santer et al., 1999; Fu et al., 2004; Thompson et al., 2012; Santer et al., 2014; Zou & Qian, 2016). With respect to issues related to the retrievals themselves, there have been multiple updates to the various independent products that have progressively dealt with issues in calibrations, orbits, overlaps, diurnal cycle adjustments etc. (for instance Mears et al., 2003, 2012; Spencer et al., 2017; Zou & Qian, 2016). Comparisons to the multi-model ensembles have been updated as a consequence (Maycock et al., 2018; Seidel et al., 2016), reducing some of the differences, but not all, and not with all observational products.

The use of these datasets for the detection and attribution of climate change has been complicated by the substantial structural uncertainty associated with the retrievals themselves (Hansen et al., 1995; Mears et al., 2003; CCSP, 2006; Thompson et al., 2012; Po-Chedley & Fu, 2012; Zou & Qian, 2016), though as the trend signal has grown and, as successive non-climatic influences have been dealt with, those differences have become less relevant. Nonetheless, the differences in atmospheric trends between models and observations continue to generate substantial discussion (McKitrick & Christy, 2020; Mitchell et al., 2020; Fyfe et al., 2021). Additionally, independent estimates of upper tropospheric and stratospheric temperatures using Global Navigation Satellite System (GNSS) radio-occultation (RO) and radiosondes from 2001 suggest that the MSU-based trends may still be underestimating tropospheric warming (Steiner et al., 2020).

In the papers referenced above, the structural uncertainty in models is often assessed through a sampling of the CMIP ensembles (over many generations of this project). This is a good way to assess some aspects of that variance - for instance, with respect to the treatment of convection, or the sensitivity to variations in climate sensitivity, but the use of an 'ensemble of opportunity' is not a complete assessment of uncertainty, and some real aspects of the uncertainty will not be sampled at all. Within a single model or model family however, we can address some structural variations in a more controlled way and specifically address different sources of uncertainty. Specifically, how sensitive are the comparisons to the real uncertainties in the forcing functions? Or to included interactive composition? We have deliberately added these kinds of model variations to the CMIP6 archive, but note that many papers examining the CMIP6 multi-model ensemble will only use a single model from a particular model family and sometimes only a single ensemble member. This leads to the potential conflation of internal variability with structural variability, underestimating both, and doesn't take advantage of more controlled variations within model families. Thus, single model family analyses should be seen as complementary and orthogonal to analyses that use a multi-model ensemble.

Mitchell et al. (2020) recently compared model trends at specific heights and found little to no improvement in the model ensemble skill as a whole compared to radiosonde data in going from Phase 5 of the Coupled Model Intercomparison Project (CMIP5) (circa 2011) to the 6th phase (CMIP6) (2019–2021). Model trends in the tropical mid-troposphere still seem too large compared to radiosonde trends (McKitrick & Christy, 2020). However, both of these papers only looked at a single ensemble member from each model and so their results may be biased by not looking at the full range of internal variability (Po-Chedley et al., 2021). Within those ensembles, however, there are a number of more structured tests that can help illuminate the reasons for the continued discrepancies. In particular, controlled variations of model structure, initial conditions, forcings and components within a single model family can be used to examine reasons for the remaining discrepancies.

In this paper, we look at the GISS ModelE2.x family of contributions to CMIP6 (Table 1). Model configurations include variations in the ocean component (either observed sea surface temperatures or two different ocean models), the radiative forcings applied, the interactivity of atmospheric composition, vertical resolution and the quality of the stratospheric representation (Table 2). Each configuration has multiple ensemble members (either 5 or 10 members). Additionally, we make use of a suite of single forcing ensemble experiments (5 members each) that highlight the vertically varying fingerprints of specific forcings to help diagnose the changes.

2 Observational Data Sources

Multiple groups have independently analysed the raw MSU, SSU and AMSU data retrievals to create time-series of atmospheric temperatures, notably the University of Alabama Huntsville (UAH) group and Remote Sensing Systems (RSS) (for the MSU data), and the NOAA Center for SaTellite Applications and Research (NOAA STAR) (both MSU and SSU products). We use the latest versions that are publicly accessible (UAH v6, RSS v4, NOAA STAR v4.1 (MSU) and v3.0 (SSU)). We use the differences between them as an indication of the structural uncertainty in the retrieved trends, though we recognise this is possibly an underestimate. The structural uncertainty of the SSU diagnostics is unclear, though reduced compared to previous versions (Thompson et al., 2012). We also focus on the global means, with the understanding that the vertical signal of trend variability in the troposphere is dominated by the moist-lapse-rate controlled tropical regions. All datasets are used through to the end of 2021.

In comparing Surface Air Temperature (SAT) trends in the models to the observations, we are mindful that trends in a blended product of SST and SAT anomalies (the Land-Ocean Temperature Index (LOTI)), such as produced by GISTEMP, HadCRUT5 or the upcoming NOAAGlobalTemp Interim product (Lenssen et al., 2019; Vose et al., 2021; Morice et al., 2021), may be systematically different from the pure SAT trends (Richardson et al., 2018). For instance, in the 10 simulations with the GISS E2.1-G `f2` configuration (see below), the 1979–2014 global mean SAT trends are 0.036 [0.028,0.044] °C/dec (95% range) greater than the LOTI trends. Thus as an alternative measure, we also use the SAT trends from the European Centre for Medium Range Weather Forecasts (ECMWF) Reanalysis version 5 (ERA5) (Hersbach et al., 2020; Simmons et al., 2021), which is perhaps a more appropriate comparison, although in practice it is similar.

3 GISS ModelE simulations

We analyse model simulations performed for the Coupled Model Intercomparison Project (Phase 6) (CMIP6) using various configurations of GISS ModelE, namely GISS-E2.1-G, GISS-E2.1-H (Kelley et al., 2020) and GISS-E2.2-G (Rind et al., 2020). The E2.1 model is an update of the GISS-E2 simulations that were used in CMIP5 (Schmidt et al., 2014; Miller et al., 2014) with the same basic resolution (2.5°×2° in the atmosphere, $\approx 1°$ in the ocean), but with multiple fixes and improvements in tuning. The -G and -H versions differ in the ocean model while the AMIP simulations use SST as a boundary forcing (PCMDI-AMIP-v1.1, based on HadISSTv1.1) (Taylor et al., 2000). The E2.2 versions have a higher model top (0.002 hPa compared to 0.1 hPa) and twice the number of model levels in the atmosphere than E2.1. The E2.2 versions have been designed to greatly improve stratospheric circulation and variability. There are also some atmospheric retunings that were made that affect the base climatology and variability (Orbe

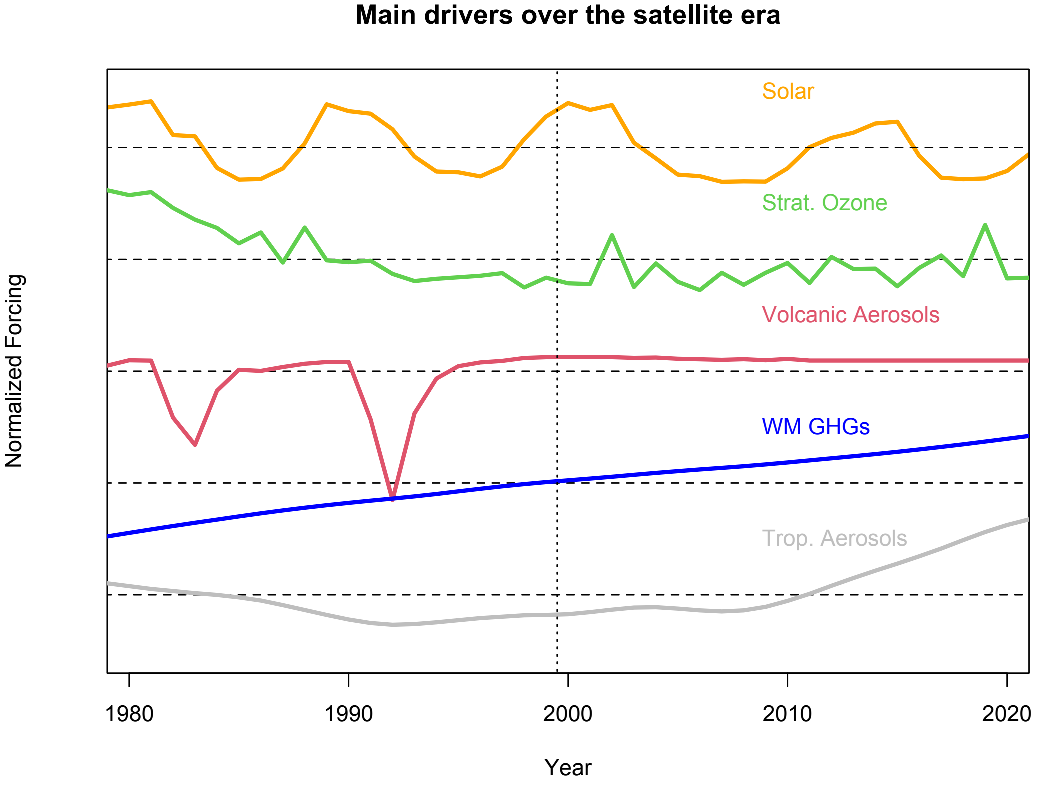


Figure 1. Schematic of the important drivers of atmospheric change over the satellite period. Each variable is plotted as a normalized index (with zero mean and unit standard deviation over the period 1979–2021) in order to highlight the pattern of variance over time. Data sources: solar irradiance (NRLTSI2) (Coddington et al., 2016), ozone hole area (Kramarova et al., 2014), volcanic stratospheric aerosol optical depth (Sato et al., 1993), well-mixed greenhouse gases and tropospheric aerosol radiative forcing (from the E2.1-G `f2` simulations) (Miller et al., 2021). The vertical dotted line distinguishes the 'ozone depletion' and 'ozone recovery' periods for the stratospheric analyses.

et al., 2020), notably there is an overall cold bias but a more realistic spectrum and magnitude of ENSO variability.

Each model configuration has options for the interactivity of atmospheric composition (specifically gas phase chemistry and aerosol physics). The versions denoted `physics_version=1 (p1)`, (NINT) have non-interactive composition, with three-dimensional seasonality (monthly means) and trends in radiatively active components (ozone and aerosols) taken from the interactive `physics_version=3 (p3)` versions that use the One-Moment Aerosol (OMA) scheme and whole-atmosphere chemistry (Bauer et al., 2020). The aerosol number concentrations that impact clouds are obtained from the aerosol mass (Menon & Rotstayn, 2006). Additionally, `physics_version=5 (p5)` uses the MATRIX aerosol module with the same chemistry (Bauer et al., 2008). In MATRIX the number of cloud activating particles is based on an aerosol activation parameterization which treats multimodal and multicomponent aerosols and provide the activated fraction of the number and mass concentration for each population, based on the population composition and the cloud updraft velocity (Abdul-Razzak et al., 1998; Abdul-Razzak & Ghan, 2000). GISS-E2.1 includes only the first indirect effect, which is the effect of aerosols on cloud droplet number concentration and thereby on cloud albedo, cloud effective radii and radiation (Menon et al., 2008, 2010). Miller et al. (2021) has a fuller description of the differences among the physics versions.

We focus on the 'historical' simulations (from 1850–2014) driven by a suite of climate drivers, the Shared Socio-Economic Pathway (SSP) scenario 2-4.5 (ssp245) runs (from 2015 onward) (Nazarenko et al., 2022), and supplemented by various single-forcing simulations for the historical period (Fig. 1). We have more simulations for the historical period than for the SSPs, but because of the vagaries of El Niño/La Niña cycles, only looking at the trends to 2014 might bias the comparisons. Thus where we have configurations that were run for SSP2-4.5, we also track trends over the longer period. The varying composition forcings for E2.1 used in the non-interactive (NINT) cases are derived from our interactive (OMA) runs. The NINT `f1` ozone and aerosol forcings came from our initial AMIP runs with E2.1 (OMA). However, the discovery of an error in the coding for stratospheric ozone chemistry led us to later rerun these simulations to generate the `f2` suite of forcings (which differ mainly in the ozone trends in the stratosphere) (Miller et al., 2021). These `f2` runs also have a complete suite of results with individual forcings to 2014, with some simulations going to 2018 or continued using the SSP2-4.5 drivers. Finally, we have a set of forcings `f3` that are interpolated from the higher vertical resolution E2.2-G (OMA) model that had a noticeably improved stratospheric-tropospheric exchange and circulation, which impacted the ozone climatology, variability and trends. Note that changes in stratospheric water vapor associated with solar-cycle related photolysis changes are not included in any of the NINT runs. In total, we examine nine separate configurations, with over 50 individual simulations.

The ozone and aerosol forcings in each individual configuration may thus be different from the schematic in Figure 1, but the overall transient pattern is close. The MATRIX runs have a faster decline of the magnitude of the tropospheric aerosol effects than those in which the aerosols are derived from the OMA runs (Bauer et al., 2020). Similarly, the exact timing of the ozone hole and stabilization is different in the E2.2 models than in E2.1 (Orbe et al., 2020). We will return to these issues in the discussion.

There is a subtle difference between the net anthropogenic forcing in the E2.1-G `f3` (NINT) runs and the E2.2-G (OMA) runs, from which the ozone and aerosols were derived, related to the indirect aerosol effect. In the non-interactive composition model configurations, the aerosol indirect effects are tuned so that year 2000 forcing, given the aerosol distribution, is around -1 W/m2 (Miller et al., 2021). This tuning is slightly different for the E2.2 and E2.1 model configurations. Thus, when taking the aerosol distribution from E2.2 and using it in the E2.1 model, there is a difference in the aerosol

Model version	Experiment	ripf number	DOI
E2.1-G	amip	r[1-5]i1p1f2	10.22033/ESGF/CMIP6.6984
	historical	r[1-10]i1p[135]f[123]	10.22033/ESGF/CMIP6.7127
	ssp245	r[1-10]i1p[135]f2	10.22033/ESGF/CMIP6.7415
	hist-volc	r[1-5]i1p1f2	10.22033/ESGF/CMIP6.7111
	hist-sol	r[1-5]i1p1f2	10.22033/ESGF/CMIP6.7101
	hist-aer	r[1-5]i1p1f2	10.22033/ESGF/CMIP6.7081
	hist-GHG	r[1-5]i1p1f2	10.22033/ESGF/CMIP6.7079
	hist-totalO3	r[1-5]i1p1f2	N/A
E2.1-H	historical	r[1-5]i1p1f2	10.22033/ESGF/CMIP6.7128
	ssp245	r[1-5]i1p1f2	10.22033/ESGF/CMIP6.7416
E2.2-G	historical	r[1-5]i1p[13]f1	10.22033/ESGF/CMIP6.6951
	ssp245	r[1-5]i1p3f1	10.22033/ESGF/CMIP6.7415

Table 1. Model experiments in CMIP6, simulation identifiers (using standard regular expression format) and DOIs for the ensemble. The `p` variable denotes different treatment of atmospheric composition, with `p1` being non-interactive (NINT), `p3` using whole atmospheric chemistry and the One Moment Aerosol (OMA) module, and `p5` which uses whole atmosphere chemistry and the MATRIX aerosol scheme (Bauer et al., 2020). Note that the definition of the forcing variants are unique to each model and physics variant (so `f1` in the `p1f1` (NINT) simulations is not related to the `f1` in the `p5f1` (MATRIX) simulations). The hist-totalO3 simulations were not submitted as part of the CMIP6 request, but are included in this analysis for completeness.

indirect effect that leads to a decrease in net forcing of 0.27 W/m^2 compared to the `f2` configuration.

We analyse the surface air temperatures (SAT), the Temperature of the Lower Troposphere (TLT) (3km/700 hPa) the Temperature of Mid-Troposphere (TMT)(5 km/500 hPa), the Temperature of the Lower Stratosphere (TLS) (18 km/80 hPa), and SSU channels 1, 2, and 3 (centered on 31, 39, 45 km and 10, 3 and 1.5 hPa respectively). Height and pressures are given for the peak in the atmospheric weighting, but the tails of the weighting functions are quite broad and extend over a wide vertical range, necessitating an appropriate weighted diagnostic in the models for comparison. The MSU and SSU diagnostics within the model are based on a fixed weighting in pressure and, although more complicated forward models can be applied (Shah & Rind, 1995), they do not noticeably impact the global trends (Schmidt et al., 2006).

Santer et al. (2021) (following Fu et al. (2011); Johanson and Fu (2006)) analysed a 'corrected' version of TMT that uses the TLS to correct for differing estimates of lower stratospheric cooling. They found that the trends in the corrected TMT in CMIP5 and CMIP6 models were not statistically distinct from the TLT trends, and so we choose not to additionally analyse the corrected TMT product.

4 Methods

We focus on the annual global anomalies and linear trends in the ensembles for each of the diagnostics described above over the 1979–2014 or 1979–2021 periods. Where needed, we reference anomalies to a 1980–1999 baseline. Ensemble spread is denoted using a 95% confidence interval derived from the 5 or 10 ensemble members.

Using an ordinary least squares linear regression on the annual anomaly data, we calculated the trends in °C per decade for each run and for each variable in the ensem-

Model Configuration	SAT (°C)	ECS (°C)	TCR (°C)	Model Top/Layers
E2.1-G f1 (NINT)	14.3	2.7	1.8	0.1 hPa/40L
E2.1-G f2 (NINT)	14.1	2.7	1.8	"
E2.1-H f2 (NINT)	14.5	3.1	1.9	"
E2.1-G f3 (NINT)	14.2	2.7	1.8	"
E2.1-G (OMA)	14.7	2.6	1.6	"
E2.1-G (MATRIX)	14.8	2.8	1.8	"
E2.2-G (NINT)	12.3	2.4	1.7	0.002 hPa/102L
E2.2-G (OMA)	11.7	2.1	1.4	"
Observations	14.3±0.5	2.0–5.0	1.2–2.4	

Table 2. Selected model characteristics for the different configurations used. Global Mean Surface Air Temperature (SAT, °C) is for the period 1981–2010. ECS and TCR are calculated from the abrupt4xCO2 and 1pctCO2 experiments, respectively. Observations are inferred from Jones et al. (1999), and the 'very likely' sensitivity ranges from the IPCC AR6 report (Masson-Delmotte et al., 2021).

ble and for the ensemble mean. Where relevant, uncertainties are given as the 95% confidence interval on the linear regression on the annual data. We construct density plots using the computed decadal trends for each run in an ensemble using the `density` function in R, following the methods of Sheather and Jones (1991). These plots are used to visualize the spread of decadal trend values within the ensemble and to compare various ensembles for comparison to the observational products for each metric.

In an effort to isolate the effects of stratospheric ozone depletion, we separate our calculation of trends in the stratosphere between the ozone depletion era (1979–1998) and the recovery period (after 1999) following Mitchell et al. (2013) and Seidel et al. (2016). For the TLS data in particular, a single linear trend is not a good fit, and so the separation of these periods can be used to more clearly distinguish the impact of the ozone depletion, as can the inclusion of volcanic and solar predictors in a multiple linear regression.

Consistency of the trends with observational products is assessed in two ways. First, we perform a simple test of whether the ensemble mean from the model configuration is within the 95% confidence interval from one of the observational product(s). This tests whether the observed trend is consistent with our estimate of the forced signal. A better test is whether an observational trend could be plausibly drawn from the model distribution of forced signal plus internal variability. This statistic is calculated following Eqn. 12 in Santer et al. (2008) (assuming a single model) using

$$d = |\overline{T_m} - \overline{T_o}|/\sqrt{s\{< T_m >\}^2 + s\{T_o\}^2} \tag{1}$$

where $s\{< T_m >\}$ is the standard deviation of the ensemble of model temperature trends T_m and $s\{T_o\}$ is the standard deviation in the linear regression the observed temperatures, respectively. This d statistic is assumed to follow a Student's t-distribution, and so the probability of getting as high a value as d can be assessed (assuming the degrees of freedom are one less than the ensemble size). If the probability is less than 95% in a two-tailed test, we conclude that the observations are consistent with the specific model ensemble. Since we are using annual data to compute the trends, the degree of autocorrelation in the residuals is small and neglected here. Inclusion of this effect would lead to slightly broader confidence intervals, and slightly greater consistency, but this does not impact the pattern of results we see nor any conclusions.

In the troposphere, models and observations indicate that the ratio of tropospheric trends to the SAT trends (related to the effective lapse rate) is more stable than the trends themselves (Wigley, 2006; Santer et al., 2005, 2017). Thus we also examine the ratios of TLT and TMT data to the SAT products in each configuration.

By contrasting specific pairs or sets of simulations in our archive, we can isolate many different aspects of the drivers and responses. For instance, in the troposphere, we can distinguish the impacts of changes in sea surface temperature (SST) (via different ocean models or observed ocean temperatures) by comparing the E2.1-G `f2`, E2.1-H `f2` and E2.1 (AMIP) simulations. One such difference arises by varying rates and structure of ocean heat uptake in E2.1-H `f2` compared to E2.1-G `f2`. There is more overall ocean heat uptake in E2.1-H `f2`, however the uptake is localized to the upper ocean rather than the deep ocean. This creates a larger SST increase in E2.1-H `f2` than in E2.1-G `f2` simulations with identical forcings (but see Miller et al. (2021) for a more thorough exploration of the ocean heat content changes in the GISS E2.1 simulations). We can also compare E2.2-G to E2.1-G `f3` and `f2` to examine whether there is significant improvement related to the higher vertical resolution model and better resolved stratosphere, and whether changes between simulations relates to the forcings or model structure. We have multiple realizations of the forcing fields (notably, aerosols and ozone) with the same underlying climate model to examine the sensitivity to those fields. Also, within each ensemble, we can estimate the impact of the modeled internal variability on the trends.

The impact of specific large volcanic eruptions in the first half of the satellite period (El Chichon in 1982 and Mt. Pinatubo in 1991) could bias the trend comparisons if there are issues in either the volcanic forcings used in the models or the model radiative response to the volcanic aerosol inputs. Similarly, there is clear evidence of a solar cycle signal in the stratospheric diagnostics. We therefore use stratospheric volcanic aerosol optical depth and solar irradiance estimates in an additional multiple linear regression to reduce the influence of potential errors in natural forcings and/or response.

5 Results

5.1 Tropospheric Trends

We first examine the time-series from a subset of the configurations in figure 2. This shows the contrasting behavior of of various model configurations within the overall warming trend, highlighting the degree of simulated internal variability. As expected, the AMIP configurations (driven with observed SST) have very little spread and are a very good match to the RSS and NOAA STAR TLT and TMT changes, though they diverge from UAH TLT, notably after 2000. However, we should note that the AMIP input files were based on SST estimates from 2016, but recent revisions (compare HadSST2 to HadSST4 for instance Rayner et al. (2006); Kennedy et al. (2019)) dealt with many non-climate discontinuities, and the net effect has been to increase the reported SST trends - by about 50% in the tropics and globally over the 1979–2019 period. These changes will have an impact on future AMIP-style runs (Flannaghan et al., 2014), likely increasing the tropospheric SAT and TLT trends. Assessing the importance of these updates will be the subject of future study.

The various flavors of E2.1-G coupled models have more spread due to over-estimated magnitude of internal variability (principally the frequency of ENSO (Kelley et al., 2020)) but broadly capture the observed trends. The response to volcanoes in 1982 and 1991 is more pronounced in both the `f2` and `f3` ensembles as compared to `f1`, demonstrating the impact of the stratospheric ozone chemistry correction between `f1` and the later iterations, even on surface temperatures. The E2.1-H model (same forcings, but with a

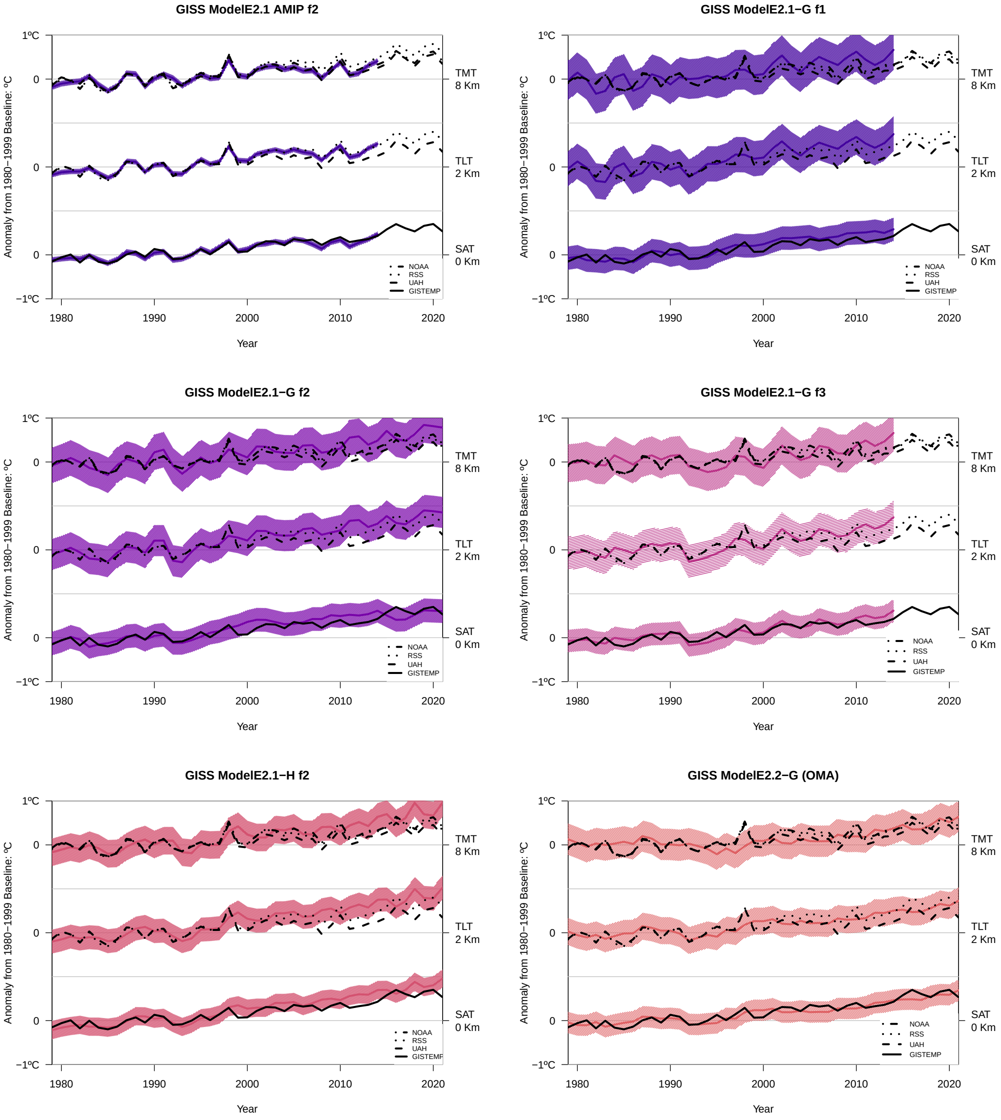


Figure 2. Tropospheric global mean temperature anomaly time series in various configurations showing the SAT, TLT and TMT changes, specifically the E2.1 AMIP; E2.1-G `f1`, `f2`, and `f3`; E2.1-H `f2`; and E2.2-G configurations for 1979–2014 (or to 2021 where available). The three diagnostics are offset by 1°C for clarity. The spread is the 95% confidence interval of the envelope of individual ensemble members. Observations from GISTEMP, UAH, RSS and NOAA STAR are in solid, dashed, dotted and dash-dotted lines, respectively.

	Ensemble Mean Linear Trends (°C/dec)					
	1979–2014			1979–2021		
Configuration	SAT	TLT	TMT	SAT	TLT	TMT
E2.1 AMIP f2	**0.15***	**0.18**^{R}	**0.12**^{R,S}			
E2.1-G f1	0.22^{G}	**0.24**^{R}	**0.20**^{S}			
E2.1-G f2	0.22*	**0.23**^{R}	**0.17**^{R,S}	**0.22***	0.25^{R}	0.21
E2.1-H f2	0.24	0.25^{R}	**0.19**^{S}	0.26^{E}	0.27^{R}	0.23
E2.1-G f3	**0.18***	**0.20**^{R}	**0.14***			
E2.1-G (OMA)	**0.18***	**0.20**^{R}	**0.14**^{R,S}	**0.19***	**0.22**^{R}	**0.18**^{R,S}
E2.1-G (MATRIX)	0.22	0.25^{R}	**0.18**^{S}	0.24	0.27	0.23
E2.2-G	**0.18***	**0.18***	**0.11***			
E2.2-G (OMA)	**0.14***	**0.15***	**0.07***	**0.17***	**0.17***	**0.12***
Observations						
ERA5	0.16±0.03			0.19±0.03		
GISTEMP	0.16±0.03			0.19±0.02		
UAH v6		0.11±0.05	0.07±0.05		0.13±0.04	0.10±0.04
RSS v4		0.19±0.05	0.11±0.05		0.21±0.03	0.14±0.04
NOAA STAR v4.1			0.14±0.05			0.16±0.04

Table 3. Model and observed trends (°C/dec) in tropospheric diagnostics. Model trends are derived from the ensemble mean. Uncertainty in the observational trend is the 95% confidence interval on the linear regression. Trends through to 2021 (using the SSP2-4.5 simulations as a continuation) are available for some configurations. Ensemble mean trends that are consistent with at least one observational product within the observational uncertainty are in bold. Ensembles that provide distributions from which any of the observations might plausibly be drawn (using Eqn. 1) are noted with a *. Where there is a difference depending on the observational product, the consistent product(s) is/are noted (G for GISTEMP, E for ERA5, U for UAH, R for RSS, S for NOAA STAR).

	Ensemble Trend Ratios			
	1979–2014		1979–2021	
Configuration	TLT/SAT	TMT/SAT	TLT/SAT	TMT/SAT
E2.1 AMIP f2	1.04 [0.96,1.20]	0.83 [0.76,0.88]		
E2.1-G f1	1.11 [1.03,1.21]	0.92 [0.82,1.01]		
E2.1-G f2	1.08 [1.01,1.14]	0.81 [0.71,0.89]	1.11 [1.06,1.18]	0.93 [0.88,1.00]
E2.1-H f2	1.05 [0.96,1.12]	0.80 [0.71,0.87]	1.04 [0.98,1.10]	0.69 [0.56,0.75]
E2.1-G f3	1.10 [1.01,1.16]	0.73 [0.63,0.80]		
E2.1-G (OMA)	1.11 [0.98,1.20]	0.76 [0.61,0.87]	1.13 [1.05,1.20]	0.85 [0.74,0.98]
E2.1-G (MATRIX)	1.12 [1.05,1.20]	0.83 [0.76,0.89]	1.14 [1.06,1.21]	0.95 [0.86,1.02]
E2.2-G	1.03 [0.82,1.21]	0.59 [0.41,0.78]		
E2.2-G (OMA)	1.04 [0.96,1.20]	0.45 [0.21,0.61]	1.03 [0.97,1.07]	0.69 [0.56,0.75]
Observations				
ERA5+UAH	0.77 [0.64,0.90]	0.55 [0.41,0.70]	0.74 [0.66,0.81]	0.55 [0.47,0.64]
GISTEMP+UAH	0.76 [0.62,0.90]	0.53 [0.38,0.69]	0.75 [0.66,0.83]	0.56 [0.46,0.65]
ERA5+RSS	1.23 [1.14,1.31]	0.82 [0.70,0.95]	1.13 [1.08,1.19]	0.77 [0.69,0.84]
GISTEMP+RSS	1.23 [1.13,1.32]	0.81 [0.67,0.94]	1.16 [1.10,1.21]	0.78 [0.69,0.86]
ERA5+NOAA STAR		0.95 [0.84,1.06]		0.86 [0.79,0.93]
GISTEMP+NOAA STAR		0.94 [0.81,1.06]		0.87 [0.79,0.95]

Table 4. Model and observed trend ratios in tropospheric diagnostics. Ensemble trends are the average of the trends from individual simulations, with the 95% envelope from the ensemble. Uncertainty in the observational ratios is the 95% confidence interval on the linear regression through the origin of the two annual timeseries. Trend ratios through to 2021 (using the SSP2-4.5 simulations as a continuation) are available for some configurations.

different ocean model) and the E2.2-G model (with higher vertical resolution) have slightly lower (and more realistic) estimates of internal variability. Note that in the coupled models, the timing of ENSO variability will not in general be correlated with the observations; ENSO is effectively stochastic in these coupled simulations, and therefore has no requirement to line up with the observed timeseries. In the AMIP simulations, the timing is forced to be the same.

Quantitative comparisons of the trends (in the historical period 1979–2014, and also in the extended period to 2021 for those ensembles that were continued under SSP2-4.5) can be seen in Table 3 for all nine model configurations. These results demonstrate more finely that there are notable differences in the trends among the configurations (and also the observational products) even within a broad qualitative agreement. For reference, differences in the ensemble mean trends that are greater than about ±0.02 are statistically significant.

Before we address why specific ensembles have different trends, it's worth noting that even for a multi-decadal trend in the global mean temperature, there is significant spread across the individual ensemble members. For E2.1-G `f2`, for which we have ten simulations, the 1979–2021 SAT trends range from 0.18°C/dec to 0.26°C/dec, so even with 43 years of data, the spread can be important (see also Fyfe et al. (2021)). The analogous range for the E2.1-H and E2.2-G (OMA) configurations are [0.22, 0.30] and [0.15, 0.18]°C/dec respectively.

In comparing the model ensemble to the real world signal, the appropriate consistency test is whether the real world trend is a plausible draw from the modelled distribution. Thus even if the trend in the ensemble mean is outside the confidence interval for the observed trend, the real world changes are potentially still consistent with the modeled distribution (Santer et al., 2008). Bolding in Table 3 is based on whether the ensemble mean is within the uncertainty of the observed trend (a test of whether the observed trend is consistent with our estimate of the forced trend), while the superscripts denote whether each observational product can be considered a plausible draw from the specific modeled ensemble.

Most configurations have a reasonable ensemble mean estimate of the tropospheric trends, and are consistent with ERA5, GISTEMP SAT and RSS TLT products. Only two configurations are also consistent with the UAH TLT estimates (both versions of E2.2-G model). Given the wide spread in estimated TMT trends across the observations, this diagnostic is less discriminating, though notably again, the E2.2-G configurations are consistent with the UAH trend.

Three configurations have ensemble SAT trends significantly greater than observations (and for two of them, this is also true for the longer 1979–2021 trend): E2.1-G `f2`, E2.1-H `f2`, and the E2.1-G (MATRIX) configurations. For E2.1-H, the Transient Climate Response (TCR) is higher than for the other configurations (Table 2) due to a reduced uptake of heat into the ocean (compared to the E2.1-G configurations), while the E2.1 (MATRIX) simulations have a more rapid decrease of (negative) aerosol forcings than with the OMA and non-interactive versions (Bauer et al., 2020). Additionally, there is a slightly significant correlation ($r^2 = 0.66$) between the TLT trends and the climate sensitivity of the coupled models.

Figure 3 show the 1979–2014 trends for the configurations. For each ensemble we show the density plot for each ensemble, with the exception of E2.1 (AMIP) which has a very narrow distribution. Across the configurations there is a wide range of trends for each diagnostic, skewed slightly higher than the observations, but on the whole mostly consistent with the RSS and NOAA STAR products. There are significant differences in the spread for specific configurations.

To better assess reasons for the spread in the trends, it's useful to also calculate the ratios of trends in the troposphere which removes the issues of overall global forcing and global mean temperature response, and allows for a focus on the global mean lapse rate, which is possibly more sensitive to model convective processes and other parameterizations (Santer et al., 2017, 2021). Table 4 gives the TLT to SAT, and TMT to SAT, ratios for each ensemble for the historical period and for the extended period to 2021.

The ratios of the trends in the troposphere are relatively stable and similar to those seen in the CMIP5 multi-model ensemble (Santer et al., 2017). Whether these ratios in the models are calculated using the ensemble mean, or the mean of the individual trends for each ensemble member, or from a linear regression passing through the origin and through the individual annual points from each ensemble member, the values are basically the same. There is some suggestion that the trends get larger over time as the climate change signal increasingly dominates over the internal variability.

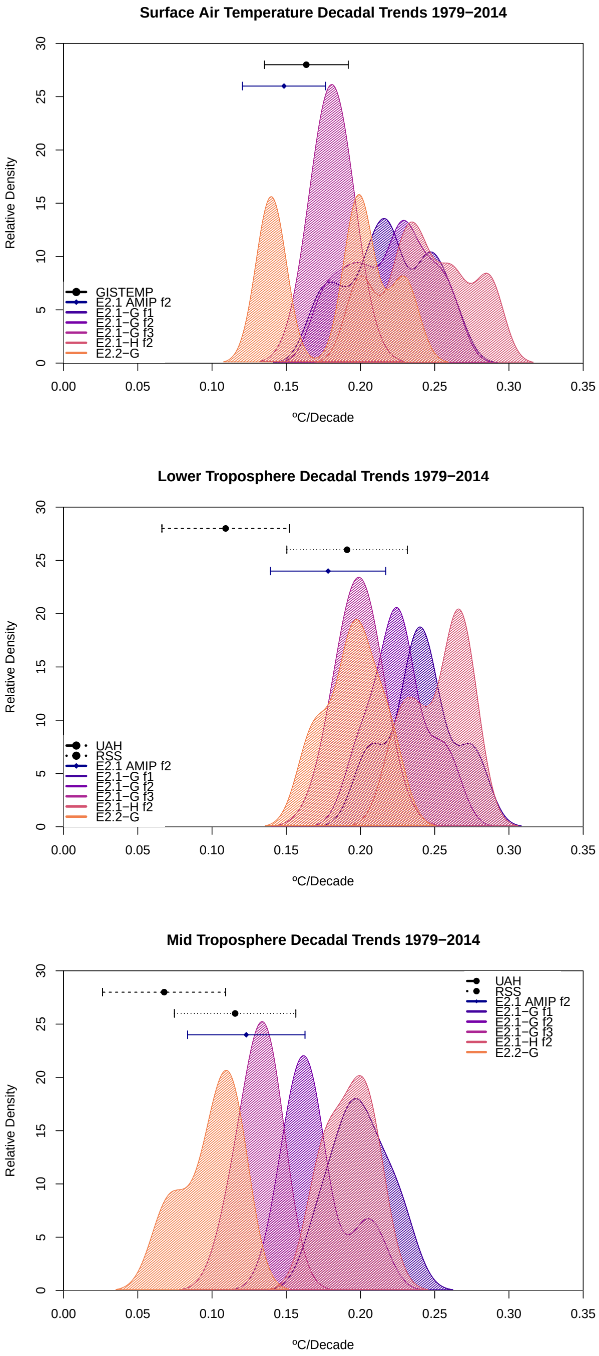
The observation-based trend ratios depend strongly on whether we use the UAH data or the RSS/NOAA STAR data. With respect to UAH, the TLT/SAT trend ratio is around 0.75, which is contrary to our basic physical understanding of the lapse rate, indicating that there is likely a systematic problem in either all the SAT products or specifically the UAH TLT product. With respect to the other data products, the ratio is around 1.2, significantly greater than one (as expected). The TMT trends (and hence ratios) are smaller because of the greater influence of the (cooling) stratosphere, but vary widely across the products from around 0.55 (using UAH), 0.8 (using RSS) or 0.9 (using NOAA STAR).

In the models, the TLT to SAT ratios are all greater than one, slightly less than observational trend ratio using RSS, but entirely inconsistent with the trend ratio using UAH. There is little spread in this value across different time periods, model structure or forcing. However, there is a much wider spread in the trend ratios for the TMT to SAT ratio, which vary by a factor of two between E2.1-G `f1` and E2.2-G (OMA). This result is tied to the spread in the TLS trends (see below), with the model configurations with the largest cooling trends in the lower stratosphere having the smallest TMT/SAT trend ratios. This underscores the importance of ozone, volcanic aerosols and possibly even solar forcing in affecting the TMT trends and the TMT/SAT trend ratio. The E2.1-G `f1` configuration had an error in the stratospheric ozone chemistry and the smallest TLS cooling (to 1998), the `f2` runs were corrected and had a more realistic stratospheric cooling and a TMT/SAT trend close to that seen with the RSS products.

E2.1-G `f2` visibly shows clearer consistency with historical observations (particularly for TMT and TLT) than E2.1-G `f1` (Fig. 3). While the `f2` improvements are noticeable in the period from the 1980's through the 90's, there is also a notable discrepancy in predicted and observed warming in the late 2000's and early 2010's, though this may be related to specific pattern of ENSO in the real world.

The contrast between the E2.1-G `f2` and `f3` configurations is also instructive. These runs differ only in the aerosol and ozone fields, and show that the tropospheric trends are quite sensitive to plausible changes in the aerosol forcing in particular. The TMT/SAT trend ratios however show more difference even though the underlying model processes are identical. This is plausibly connected with a lower tropical tropopause (by about 4 hPa/100m or so). This arises because of a slower Brewer-Dobson circulation in the in the E2.2 simulations from which the `f3` forcings are drawn (Orbe et al., 2020). This implies a slightly greater stratospheric contribution to the TMT trends and is supported by the observation that the 'corrected' TMT trends (Johanson & Fu, 2006) are closer in the two ensembles than for the standard TMT trends (e.g. the `f3` trends are 17% smaller than the `f2` trend for corrected TMT, compared to 25% smaller in TMT). Additionally, while the trends in ozone are driven mainly by chemistry, acceleration in the lower branch of the Brewer-Dobson circulation will have differing impacts on ozone trends given different gradients in the climatology.

The E2.1-G f2 ensemble runs hindcast more warming in the 2000's than the AMIP ensemble does. As the AMIP runs in use in these figures is the same as E2.1-G f2 runs except for its reliance on observational surface ocean temperature data, this indicates that, at least in part, the atmospheric-ocean dynamics in the coupled ensemble are contributing to more warming. However, it should be noted that the 95% confidence envelope of E2.1-G f2 does still overlap with the satellite observations in the majority of the temperature anomaly charts even as the model approaches the present.



Surface Air Temperature Decadal Trends 1979−2014

Lower Troposphere Decadal Trends 1979−2014

Mid Troposphere Decadal Trends 1979−2014

Figure 3. Trend analysis across the non-interactive configurations for the tropospheric diagnostics for the period 1979–2014. Uncertainties on the observational trends are the 95% confidence intervals. E2.1 (AMIP) results are plotted as a 95% spread. All other model ensembles are plotted as a density plot.

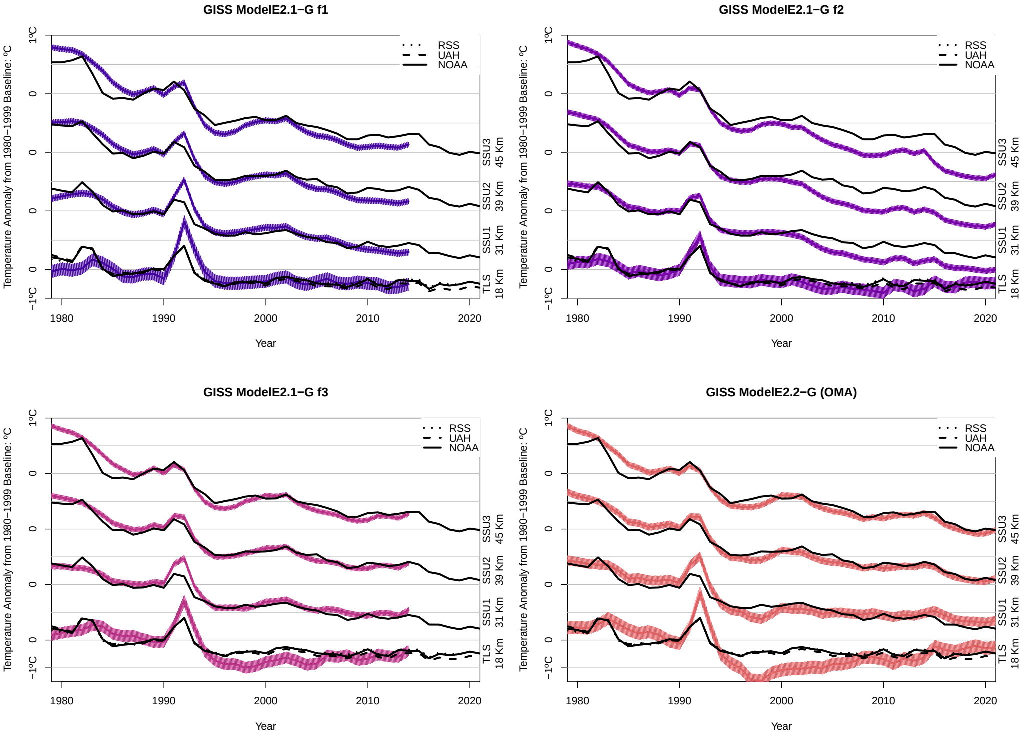


Figure 4. Stratospheric time-series for MSU TLS and SSU products for E2.1-G `f1`, `f2` and `f3`, and for E2.2-G compared to the observations (each offset by 1°C for clarity). Observations from RSS, UAH and NOAA STAR are dotted, dashed and solid respectively.

					Linear trends (°C/dec)							
	1979–1998				1999–2014				1999–2021			
Configuration	TLS	SSU-1	SSU-2	SSU-3	TLS	SSU-1	SSU-2	SSU-3	TLS	SSU-1	SSU-2	SSU-3
E2.1 AMIP f2	**-0.25***	**-0.90***	-1.28	-1.60	-0.25	-0.76	-0.82	-0.77				
E2.1-G f1	-0.10	-0.64	**-1.13***	-1.59	**-0.14***	-0.69	-0.77	-0.74				
E2.1-G f2	**-0.28***	**-0.90***	-1.28	-1.61	**-0.20**^{U,S}	-0.76	-0.83	-0.77	**-0.01***	-0.58	-0.72	-0.85
E2.1-H f2	**-0.26***	**-0.92***	-1.31	-1.63	**-0.20**^{U,S}	-0.76	-0.84	-0.79	**-0.00***	-0.59	-0.73	-0.86
E2.1-G f3	**-0.56***	**-0.76***	**-1.14***	-1.55	0.23	**-0.33***	**-0.52***	**-0.61***				
E2.1-G (OMA)	**-0.36***	**-0.95***	-1.34	-1.65	**-0.03***	**-0.53***	-0.65	**-0.69***	**0.10**^{R,S}	**-0.39***	**-0.55***	**-0.62***
E2.1-G (MATRIX)	**-0.42***	**-0.98***	-1.35	-1.66	**0.04***	**-0.44***	**-0.58***	**-0.65***	0.16	-0.28	**-0.44***	**-0.54***
E2.2-G	**-0.61***	**-0.80***	**-1.15***	-1.51	0.24	**-0.31***	**-0.49***	**-0.57***				
E2.2-G (OMA)	-0.85^U	**-1.01***	-1.26*	-1.54	0.34*	-0.14*	**-0.41***	**-0.53***	0.42	-0.18*	**-0.46***	**-0.58***
Observations												
UAH v6	-0.49±0.27				-0.09±0.11				-0.10±0.06			
RSS v4	-0.41±0.28				-0.04±0.11				-0.05±0.06			
NOAA STAR	-0.46±0.28	-0.86±0.22	-1.02±0.23	-1.11±0.29	-0.07±0.12	-0.37±0.13	-0.44±0.16	-0.51±0.15	-0.06±0.06	-0.40±0.08	-0.51±0.09	-0.56±0.09

Table 5. Model and observed trends (°C/dec) in stratospheric diagnostics. Ensemble mean and spread are given for two distinct periods pre- and post-1998 (with trends ending in 2014 or 2021). Uncertainty in the observational trend is the 95% confidence interval in the linear regression. For the extended period, we use up to 2021 for all channels. Ensemble trends that are consistent with at least one observational product within the observational uncertainty are highlighted in bold. Ensembles that provide distributions from which any of the observations might plausibly be drawn are noted with a *. Where there is a difference depending on the observational product, the consistent product(s) is/are noted (U for UAH, R for RSS, S for NOAA STAR).

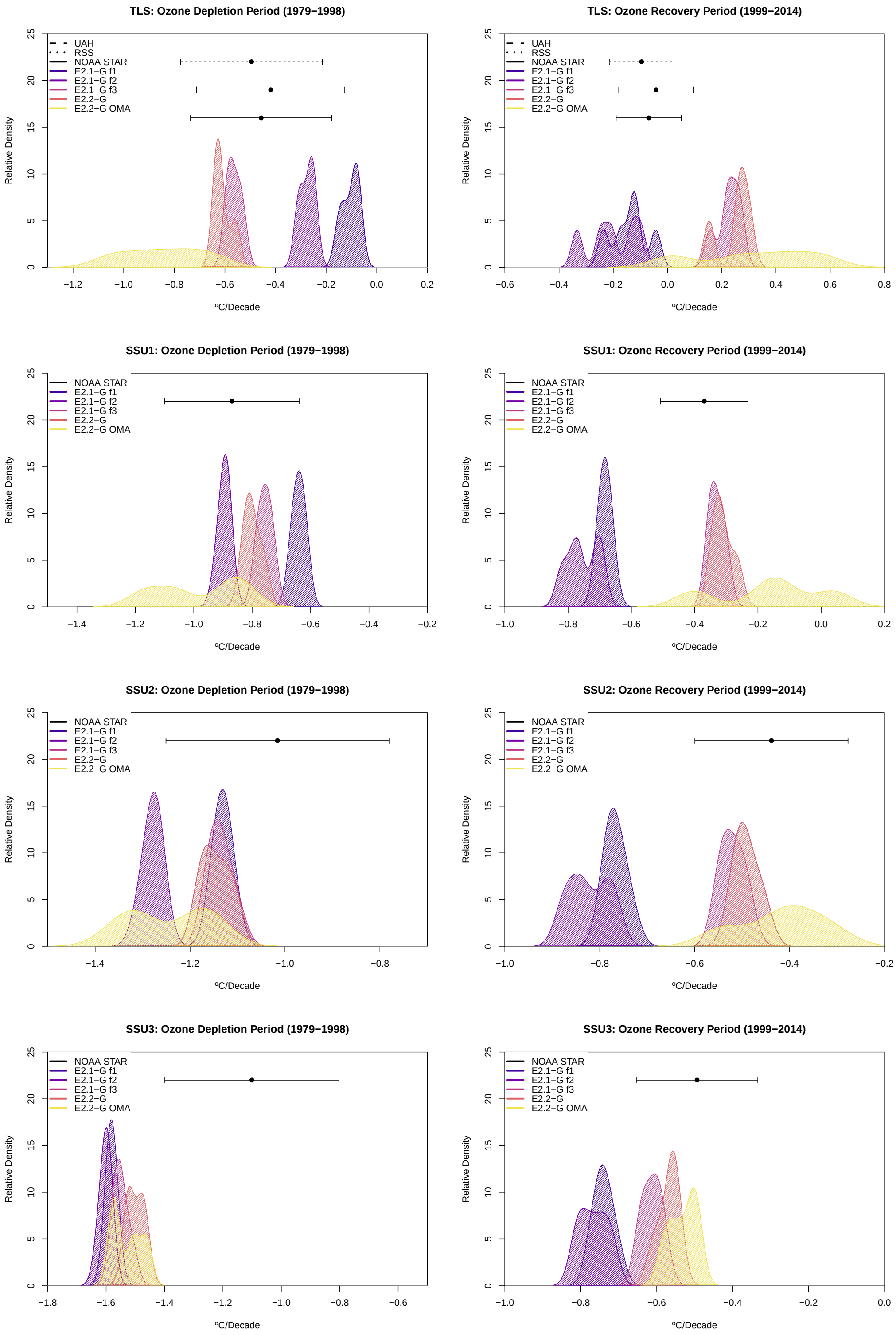


Figure 5. Ensemble trends and observations for the TLS and SSU channels in the stratosphere. Uncertainties in the observations are the 95% confidence interval on the linear regression.

5.2 Stratospheric Trends

Figure 4 shows a selection of model anomalies in the stratospheric diagnostics, TLS and the three SSU channels. The effect of the ocean module is not significant, so the results of the E2.1-H or E2.1 (AMIP) configurations are omitted for clarity. Overall cooling trends increase as a function of height as the CO_2 impact increases, while the impact of volcanic aerosols decreases. It's clear that a single linear trend is not a good fit for these diagnostics over this period because of the impacts of volcanic aerosols early in the period, the changing impact of ozone depletion (strong in the first 20 years of the record, and less important subsequently, see fig. 1) and the impact of solar cycles. The overall structure of the changes is reasonable in all configurations, but there are clear discrepancies. The difference in the volcanic response between the E2.1-G `f1` and `f2` simulations is directly related to the correction of bug in the stratospheric chemistry, which clearly improved the simulations. However, in all cases, the response to El Chichon in 1982 is muted compared to that for Mt. Pinatubo (1991) which appears to be too large. The other clear difference is the between the `f2` and `f3` runs which have markedly different upper stratospheric trends. The high-top simulations using E2.2-G have a much closer match to the SSU observations than E2.1-G `f2`, which is mimicked by the results in E2.1-G `f3` which uses the E2.2-G (OMA) ozone climatology and trends.

More quantitatively, the ensemble mean linear trends in the stratosphere are given in Table 5. The ozone depletion signal is dominant in the TLS trends so, following Seidel et al. (2016), we separate the stratospheric linear analysis into two periods, 1979–1998 (the 'ozone depletion' period) and the subsequent evolution 1999–2014 (or with a continuation to 2021). In the lower stratosphere, the different E2.1-G configurations behave similarly in the early period. However, trends in the recovery period vary as a function of the specific forcings. The `f2` forcings indicate more cooling than other configurations in the SSU channels. This is not the case in the lower stratosphere, though, which is driven more by the ozone trends. E2.1-G `f2` results track the TLS observations more closely than E2.1-G `f3` or E2.2 do. Notably, in the decadal trend diagrams (fig. 5), the E2.1-G `f3` configuration has the widest difference between the ozone depletion and recovery period while `f2` is more centered around the observational data without significant differentiation between the two periods. The improvements in model agreement from the E2.1-G `f2/f3` to E2.2 models configurations is increasingly pronounced in the higher altitudes.

While the internal variability in the stratosphere channels is generally muted compared with the troposphere, there is significant spread in each ensemble for the two periods (fig. 5) due mainly to the different forcings. The model ensembles using the high-top model (or the ozone fields derived from it) have distinct trends from the other ensembles, doing better in the ozone depletion period (except for SSU-3) and in the ozone recovery period (except for MSU-TLS). The difference between E2.1 f1 and f2 ensembles is mostly clearly seen in the ozone depletion period trends, where the E2.1 f2 models cool more which is a better match to the observations. Almost all model cooling trends in SSU-3 are too large.

Some insight into the reasons behind the trend disparities can be gained by looking at the ozone trends in the different ensembles (fig. 6). We look at the 60°S–60°N mean total column ozone (since the satellite observations don't completely capture the changes at the high latitudes). The model ensembles are slightly depleted in total ozone in the early 1980's (by 7–9 DU) compared to the SBUV (v8.7) data, but all show a steady decline over the 'ozone depletion' period in line with the observations at least through to 1994. In the E2.1 configurations, the depletion period lasts longer than observed (by 5 years or so), while in the E2.2 configuration there is a greater perturbation associated with Mt Pinatubo and a deeper depletion towards the the end of the 1990's.

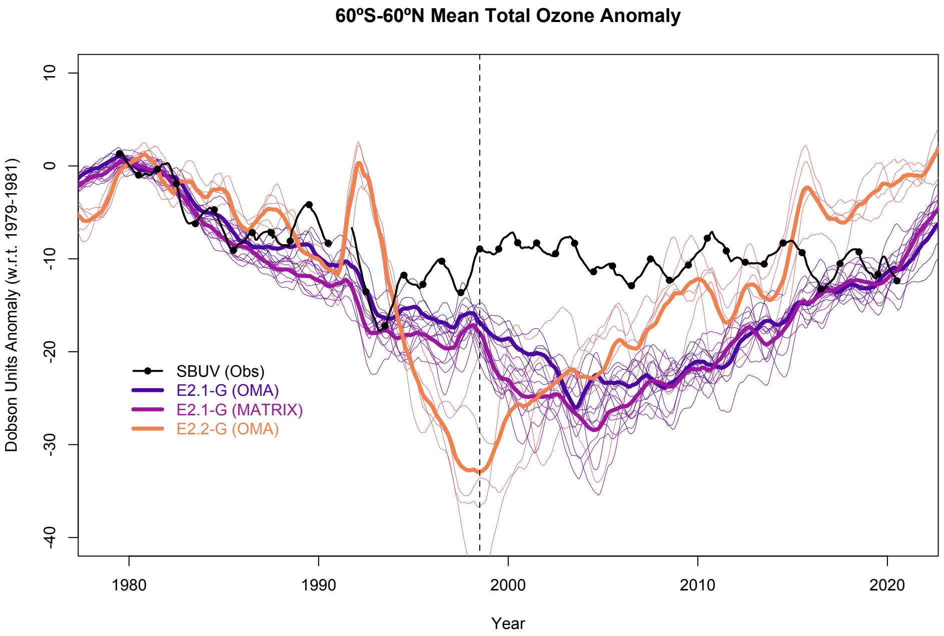


Figure 6. Time-series of the 12-month running mean total column ozone (DU) (60°S–60°N) for the ensembles with interactive composition from 1979–2021. Thick lines are the ensemble mean, with individual members in the thin lines. The vertical dashed line denotes the separation of the 'ozone depletion' and 'ozone recovery' periods. Observations are from SBUV v8.7 (McPeters et al., 2013, and updates).

In these ModelE simulations, stratospheric chlorine loading (and hence overall ozone depletion) was set using a relationship based on the concentrations of CFC-11 and CFC-12. In the real world, the Equivalent Effective Stratospheric Chlorine (EESC) (Newman et al., 2007) depends on many lower concentration gases which are not explicitly tracked in the GCM. While the parameterized EESC is a good approximation up to about 2000, there is an increasing divergence with the real world afterwards, with the real EESC reducing more rapidly than in the model. Notably, the EESC peak in the real world was around 2001 with a decrease of 14% from its peak by 2020, while in our parameterization it does not occur until 2005, and has only decreased 8%. Thus in all the model configurations, the ozone depletion period extends a few more years than observed, and does not recover as quickly. Given the cooling impact of ozone depletion on the TLS and SSU channels, this issue can explain a portion of the mismatched trends in the ozone recovery period.

5.2.1 Multiple linear regression using volcanic and solar predictors

The clear impacts of volcanic eruptions and solar cycles in the stratospheric diagnostics increase the non-linearity of the temperature trends in the stratosphere. We therefore redo the linear regressions including predictors for these effects to assess whether the longer term trends are being affected by potential errors in the modeling of the volcanic or solar responses. If those errors are significant, we should see a better match in the modeled and observed linear trends.

We use a volcanic predictor based on the aerosol optical depth history (Sato et al., 1993, and updates to 2012 with constant values in 2013 and 2014) and a solar index derived from a historical total solar irradiance dataset (Coddington et al., 2016). We highlight the results with respect to SSU-2 in figure 7 and for the ensemble means for all diagnostics in figure 8. As expected, the linear trends for both the models and the observations in the ozone depletion period are both smaller in magnitude and less uncertain when the volcanic and solar predictors are excluded. The added predictors make the most difference in the TLS trends where the resulting linear trends are all more consistent. Note that the linear trends over this time period will include effects from both CO_2 and ozone depletion.

For the SSU channels the impacts are more muted, but there aren't any major shifts in the consistency with the observations. Notably, the E2.1-G `f2` simulations are show notably stronger stratospheric cooling than observed regardless of which additional predictors are included. The E2.1-G `f3` and E2.2-G results for the SSU-3 channel get closer to the observed trends, but are still too strong, suggesting that further investigation of the time-series of ozone depletion is warranted.

There are some interesting aspects of these regressions in the ozone depletion period. This is illustrated for SSU Channel 2 in Table 6. First, the solar regressions (effectively over two solar cycles are in line with the inference from the observations, as are the volcanic effects. It's noticeable that the volcanic signal is stronger in the E2.2-G (OMA) and E2.1-G `f3` configurations than in the other two configurations and the observations (though all coefficients are consistent with the observations). In all cases the linear trends are now more coherent with the observations, but collectively they are still a little too steep (except for E2.1-G `f3` which is just about compatible). The slightly enhanced solar effects in E2.1-G `f3` may arise from the lack of solar-cycle related changes in photolysis which would causes upper stratospheric water vapor to decrease at solar maxima, damping the temperature impact.

SSU-2 Multiple Linear Regression Coefficients 1979–1998

Model Configuration	Intercept	Linear trend (°C/dec)	Volcanic AOD	Solar TSI
E2.1-G f2	-460	-1.15	4.4	0.48
E2.1-G f3	-563	-0.98	5.6	0.56
E2.2-G	-463	-1.02	5.1	0.49
E2.2-G (OMA)	-445	-1.13	5.8	0.49
NOAA STAR	-549 ± 240	-0.87±0.11	4.9±1.9	0.53±0.17

Table 6. Multiple linear regression results for the SSU-2 diagnostics for selected ensemble mean model configurations and observations for the 'ozone depletion' period. All coefficients are significant at the 95% level. Uncertainties on the regression of the NOAA STAR observations are the 95% confidence levels.

SSU-2 Linear Trends 1979–1998 w/ and w/o Solar and Volcanic Predictors

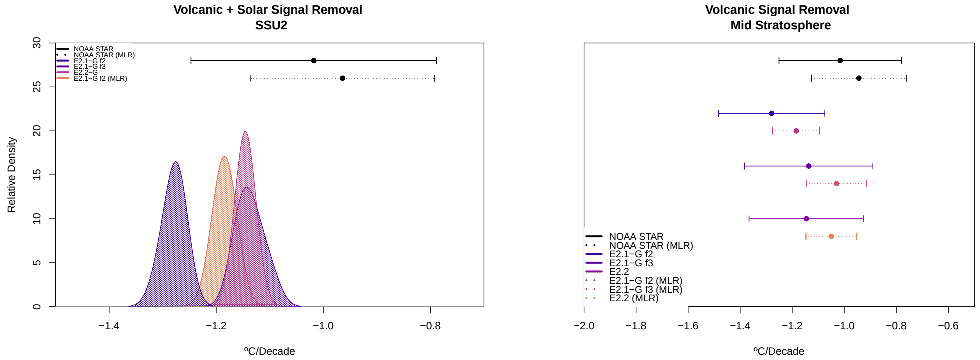


Figure 7. Ensemble linear trends with and without volcanic aerosol and solar predictors in the mid-stratosphere (SSU-2) (1979–1998). Uncertainties on the observations represent a 95% confidence interval on the linear regression. a) The linear trends in each ensemble in the multiple linear regression and, b) trends in the ensemble mean for each configuration with and without the MLR.

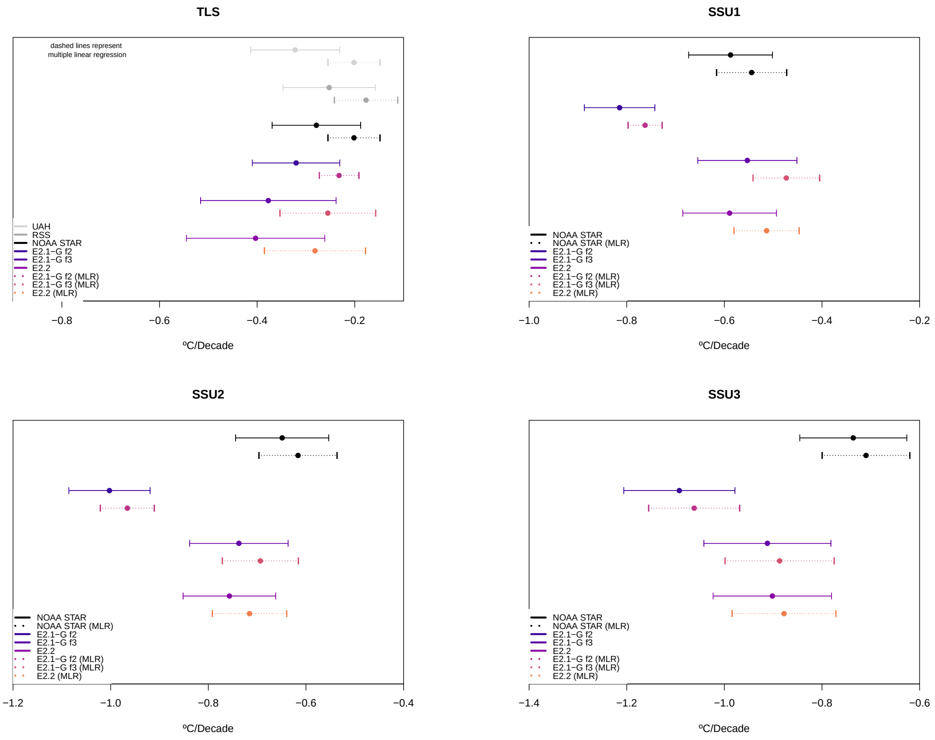


Figure 8. Ensemble linear trends with and without using volcanic and solar predictors for the stratospheric diagnostics (1979–2014). Error bars on the observations represent a 95% confidence interval on the linear regression.

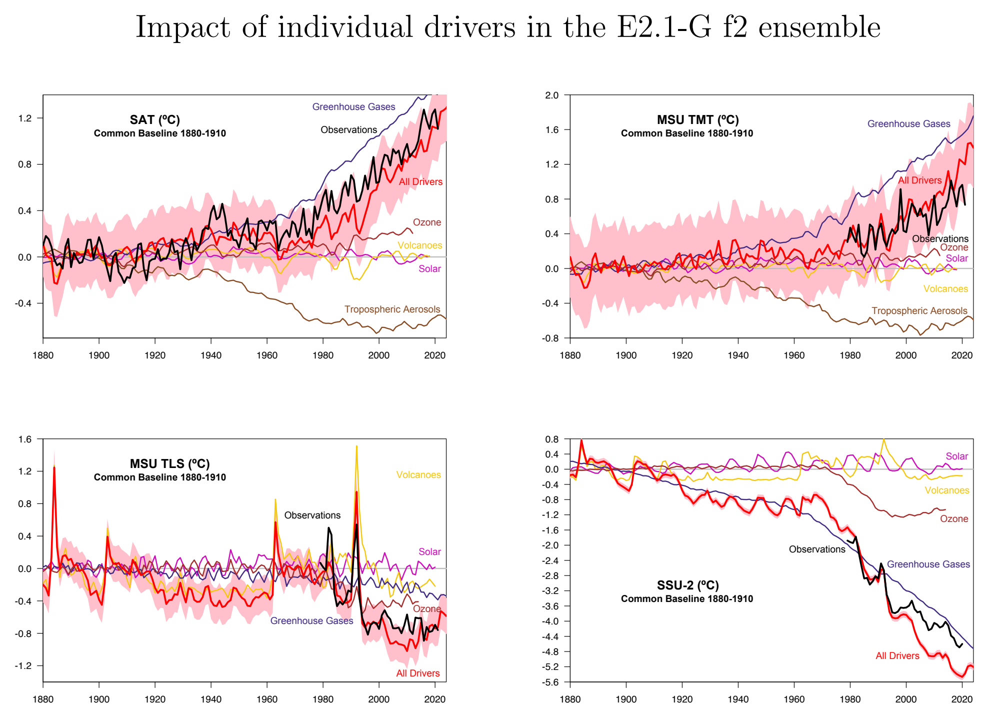


Figure 9. Breakdown of the ensemble mean SAT, TMT, TLS, and SSU-2 anomalies for E2.1-G `f2` as a function of relevant single forcings from 1880–2021 with respect to a baseline of 1880–1910. The uncertainty on the 'All drivers' line is the derived from the 95% confidence interval from the pre-industrial control run, which in practice is indistinguishable from the envelope of the individual ensemble member spread. For the tropospheric diagnostics, we apply a 4-year running mean filter to reduce the 'weather' noise that still remains in the ensemble mean for each single forcing (which only used 5 ensemble members). Illustrative observations are the GISTEMP LOTI, RSS MSU, and the NOAA-STAR SSU. SAT observations are plotted with the same baseline as the models, but for the satellite era diagnostics we align them so that their mean is equal to the model 'All drivers' ensemble mean over 1980–1999.

6 Single-forcing results

For the E2.1-G `f2` (NINT) configuration, we performed a complete set of single forcing simulations for the historical period (5 ensemble members each). Thus for each diagnostic, we can illustrate the modeled response to each of the drivers individually (fig. 9). Because of the way the historical composition files were derived (from an OMA simulation), the solar-only and volcanic-only forcing simulations include compositional responses (notably in ozone) which is a new feature compared to how similar exercises were done in previous iterations (Marvel et al., 2015). Some forcings (such as orbital forcing or land-use/land-cover change) don't have a significant expression in the global mean diagnostics (very close to zero for orbital forcing, and slightly negative for land-use/land-cover on SAT) and are not included in the figures. The impact of 'Tropospheric Aerosols' is only significant for the tropospheric diagnostics, though there is a very slight stratospheric warming associated with them (not shown). The ozone-only results used composition files from anthropogenic-only simulations (with no natural drivers), and thus include both tropospheric ozone increases and stratospheric ozone decreases, driven by emissions of chemical precursors and ozone-depleting substances. The Greenhouse Gas (GHG)-only simulations include the radiative impacts of CO_2, CH_4, N_2O and CFCs (but not any chemistry related impacts).

For the SAT, TMT and SSU2 diagnostics illustrated in figure 9, the dominance of GHGs in driving the long-term trend is clear, however, other forcings (tropospheric aerosols, ozone, volcanic aerosols) all play key roles, though their importance varies through the atmosphere. Volcanic and ozone forcings are relevant throughout the atmosphere, while solar forcing increases in importance with height. GHG, volcanic and ozone impacts all change sign in going from the troposphere to the stratosphere. The breakdown for TLT is similar to that for SAT, and the two other SSU channels (not shown) resemble SSU-2. For TLS, the ozone changes are the dominant factor in recent decades (c.f. Ramaswamy et al. (1996)), although the GHG (CO_2) impact is increasing in importance. The variations across the other model configurations, particularly in the stratosphere, can be thought of as being driven by small changes in the secondary components - notably stratospheric ozone and the volcanic response.

7 Discussion

The vertical profile of atmospheric trends over recent decades is a key metric in assessing the fidelity of climate models, and ultimately in understanding why the current climate is changing. While the overall patterns are robust and clear - warming in the troposphere, cooling in the stratosphere, punctuated by volcanic effects, and modulated by solar activity - there are sufficient quantitative discrepancies between models and observations and among observational products to merit closer attention.

Among the dozens of simulations with the GISS E2.x suite of Earth System Models in nine different configurations, there is sufficient structural and stochastic variety to help us more easily identify the factors that have impacted the model-observation comparisons than when looking across the whole multi-model ensemble.

Most obviously, even for trends over four decades, there is substantial intra-ensemble spread due to the different realizations of internal variability in the troposphere which is an essential consideration when comparing a model ensemble to the single real world realization (Santer et al., 2008; Po-Chedley et al., 2021).

Secondly, two factors for which there is still substantial uncertainty - the tropospheric aerosol forcing changes and the rate and manner of heat uptake into the ocean - still make a notable difference in the troposphere. Model configurations with less deep ocean heat uptake and those with a faster decrease in aerosols have stronger surface trends than those without. Additionally, while the spread of climate sensitivity in these configurations (2.1°C to 3.1°C) is not as wide as in the broader CMIP6 ensemble (1.8–5.6°C) (Zelinka et al., 2020), it is sufficient for the trends under similar forcings to diverge. Unfortunately, the intersection of these factors means that it is hard to constrain one of them alone using these global metrics.

Thirdly, it is likely that there will be further refinements and better estimates of structural uncertainty for the satellite retrievals themselves that might better reconcile the independent estimates of upper tropospheric warming (Steiner et al., 2020). If so, the conclusions here may need to be revisited.

Nonetheless, there are robust conclusions that can be drawn. There is clearly more work to be done related to the response of these models to volcanic eruptions. The discrepancies seen in the magnitude and time-evolution of volcanic signal suggest that either the input aerosol fields are not accurate, and/or that the model response (perhaps in the heterogeneous chemistry) is flawed. More first principles modeling via injection of volcanic gases and subsequent aerosol modeling (LeGrande et al., 2016) and the results of the VolMIP exercise might lead to more coherent and hopefully more accurate impacts (Timmreck et al., 2018; Zanchettin et al., 2022).

Our results underline the importance of ozone climatology and chemistry responses. The difference in SSU trends in the E2.1-G `f2` and `f3` configurations can only be due to the different ozone inputs. There are two facets to these differences, a more accurate climatology, with a lower altitude ozone layer (consistent with a more accurate (weaker) Brewer-Dobson circulation and older stratospheric age-of-air in the E2.2 models (Orbe et al., 2020)) and slightly different trends over time. We find that the decadal trends of E2.2-G agree more closely with ozone depletion and recovery observations than those of E2.1-G, though there remains observational disagreement in the time series of mean total ozone anomalies for both configurations. Our results show that the magnitude of the SSU cooling trends (driven mainly by the CO_2 forcing) are mediated by the ozone response in the models. Ozone will likely respond differently in the high-top versus low-top models because of the changes in the Brewer-Dobson circulation and because the slower circulation in the high-top versions changes the response of ozone to increased CO_2 though the details of this will be the subject of future research. Ongoing improvements in the modeling of stratospheric halogen loads will also likely make a difference to the trends in the 'ozone recovery' period.

It should be noted that, in agreement with other CMIP6 models, GISS ModelE output tends to agree more with RSS and NOAA STAR tropospheric observations than with the UAH data. The results from the AMIP results suggest that the UAH results start to anomalously deviate from expectations around the year 2000. Updates to the SST inputs for the AMIP simulations (Kennedy et al., 2019) are likely to worsen the comparison further. Also, since the ratio of TLT to SAT trends is a robust metric across configurations, independent of the climate sensitivity, vertical resolution or ocean component, the fact that this is not consistent with any UAH/SAT trend ratio is suggestive of a systematic problem.

To summarise, it is too simplistic to attribute all model discrepancies with the MSU and SSU observational to a single dominant cause. This analysis has demonstrated that even within a single model family, multiple factors are at play: ozone chemistry and climatology, and feedbacks are clearly important to the TMT and stratospheric channels; correct simulations of volcanic aerosol and consequent compositional changes are also important. Both are targets for future development. However, internal variability and structural uncertainty in the observations are essential components to address in any analysis. Attempts to classify the responses in the multi-model ensemble by using only single ensemble members from each model or model family will, simply by chance, conflate internal variability with structural uncertainty (e.g. McKitrick and Christy (2020); Mitchell et al. (2020)) and may give misleading results.

Open Research

We use observed MSU/SSU data products from Mears and Wentz (2016); Spencer et al. (2017); Zou and Qian (2016), and surface temperature data from GISTEMP (Lenssen et al., 2019) and ERA5 (Hersbach et al., 2020). The indices of various drivers in Fig. 1 are ozone hole area (*NASA Ozone Watch*, 2022), total column ozone (McPeters et al., 2013) (updated to v8.7), volcanic aerosol optical depth (Sato et al., 1993), Total Solar Irradiance (Coddington et al., 2016), and radiative forcing (Miller et al., 2021). The GISS ModelE data are available from the *Earth System Grid Federation* (2022) and also from the *NASA Center for Climate Simulation* (2022) portal (including non-CMIP6 simulations and derived data (such as the model MSU and SSU diagnostics).

Acknowledgments

Climate modeling at NASA GISS is supported by the NASA Modeling Analysis and Prediction program, and resources supporting this work were provided by the NASA High-End Computing (HEC) Program through NASA Center for Climate Simulation (NCCS) at Goddard Space Flight Center (GSFC). M.C.C. was supported by the Climate Change Research Initiative (CCRI) at GISS funded by the GSFC Office of STEM Engagement. Comments from two anonymous referees are greatly appreciated.

References

Abdul-Razzak, H., & Ghan, S. J. (2000). A parameterization of aerosol activation: 2. Multiple aerosol types. *Journal of Geophysical Research: Atmospheres*, *105*(D5), 6837–6844. doi: 10.1029/1999jd901161

Abdul-Razzak, H., Ghan, S. J., & Rivera-Carpio, C. (1998). A parameterization of aerosol activation: 1. Single aerosol type. *Journal of Geophysical Research: Atmospheres*, *103*(D6), 6123–6131. doi: 10.1029/97jd03735

Bauer, S. E., Tsigaridis, K., Gao, Y. C., Faluvegi, G., Kelley, M., Lo, K. K., ... Wu, J. (2020). Historical (1850–2014) aerosol evolution and role on climate forcing using the GISS ModelE2.1 contribution to CMIP6. *Journal of Advances in Modeling Earth Systems*. doi: 10.1029/2019MS001978

Bauer, S. E., Wright, D. L., Koch, D., Lewis, E. R., McGraw, R., Chang, L.-S., ... Ruedy, R. (2008). MATRIX (Multiconfiguration Aerosol TRacker of mIXing state): an aerosol microphysical module for global atmospheric models. *Atmos. Chem. Phys.*, *8*, 6003–6035.

CCSP. (2006). *Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences.* Asheville, NC, USA: National Oceanic and Atmospheric Administration, National Climatic Data Center. (Karl, T. R. et al., eds, 164 pp.)

Christy, J. R., & Spencer, R. W. (1995). Assessment of precision in temperatures from the microwave sounding units. *Climatic Change*, *30*(1), 97–102. doi: 10.1007/bf01093227

Coddington, O., Lean, J. L., Pilewskie, P., Snow, M., & Lindholm, D. (2016). A solar irradiance climate data record. *Bulletin of the American Meteorological Society*, *97*(7), 1265–1282. doi: 10.1175/bams-d-14-00265.1

Earth System Grid Federation. (2022). Dept. of Energy, USA. Retrieved from `https://esgf-node.llnl.gov/search/cmip6/[Dataset]`

Flannaghan, T. J., 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). doi: 10.1002/2014jd022365

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

Fu, Q., Manabe, S., & Johanson, C. M. (2011). On the warming in the tropical upper troposphere: Models versus observations. *Geophysical Research Letters*, *38*. doi: 10.1029/2011gl048101

Fyfe, J. C., Kharin, V. V., Santer, B. D., Cole, J. N. S., & Gillett, N. P. (2021). Significant impact of forcing uncertainty in a large ensemble of climate model simulations. *Proceedings of the National Academy of Sciences*, *118*(23). doi: 10.1073/pnas.2016549118

Hansen, J., Wilson, H., Sato, M., Ruedy, R., Shah, K., & Hansen, E. (1995). Satellite and surface temperature data at odds? *Climatic Change*, *30*, 103–117. doi: 10.1007/BF01093228

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal* *of the Royal Meteorological Society*, *146*(730), 1999–2049. (Global mean surface temperature was downloaded from `https://climate.copernicus.eu/sites/default/files/ftp-data/temperature/2021/12/ERA5_1991-2020/ts_1month_anomaly_Global_ERA5_2T_202112_1991-2020_v01.csv`. [Dataset]) doi: 10.1002/qj.3803

Johanson, C. M., & Fu, Q. (2006). Robustness of tropospheric temperature trends from MSU channels 2 and 4. *Journal of Climate*, *19*(17), 4234–4242. doi: 10.1175/jcli3866.1

Jones, P. D., New, M., Parker, D. E., Martin, S., & Rigor, I. G. (1999). Surface air temperature and its variations over the last 150 years. *Revs. Geophys.*, *37*, 173–199.

Jones, P. D., Osborn, T. J., Wigley, T. M. L., Kelly, P. M., & Santer, B. D. (1997). Comparisons between the microwave sounding unit temperature record and the surface temperature record from 1979 to 1996: Real differences or potential discontinuities? *Journal of Geophysical Research: Atmospheres*, *102*(D25), 30135–30145. doi: 10.1029/97jd02432

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., ... Yao, M.-S. (2020). GISS-E2.1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems*, *12*. doi: 10.1029/2019ms002025

Kennedy, J. J., Rayner, N. A., Atkinson, C. P., & Killick, R. E. (2019). An ensemble data set of sea surface temperature change from 1850: The Met Office Hadley Centre HadSST.4.0.0.0 data set. *Journal of Geophysical Research: Atmospheres*, *124*(14), 7719–7763. doi: 10.1029/2018jd029867

Kramarova, N. A., Nash, E. R., Newman, P. A., Bhartia, P. K., McPeters, R. D., Rault, D. F., ... Labow, G. J. (2014). Measuring the Antarctic ozone hole with the new Ozone Mapping and Profiler Suite (OMPS). *Atmospheric Chemistry and Physics*, *14*(5), 2353–2361. doi: 10.5194/acp-14-2353-2014

LeGrande, A. N., Tsigaridis, K., & Bauer, S. E. (2016). Role of atmospheric chemistry in the climate impacts of stratospheric volcanic injections. *Nature Geoscience*, *9*(9), 652–655. doi: 10.1038/ngeo2771

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., & Zyss, D. (2019). Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres*, *124*, 6307–6326. doi: 10.1029/2018jd029522

Manabe, S., & Wetherald, R. T. (1967). Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *J. Atmos. Sci.*, *24*, 241–259.

Marvel, K., Schmidt, G. A., Miller, R. L., & Nazarenko, L. S. (2015, dec). Implications for climate sensitivity from the response to individual forcings. *Nature Climate Change*, *6*(4), 386–389. doi: 10.1038/nclimate2888

Masson-Delmotte, V., et al. (Eds.). (2021). *Climate Change 2021: The physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

Maycock, A. C., Randel, W. J., Steiner, A. K., Karpechko, A. Y., Christy, J., Saunders, R., ... Zeng, G. (2018). Revisiting the mystery of recent stratospheric temperature trends. *Geophysical Research Letters*, *45*(18), 9919–9933. doi: 10.1029/2018gl078035

McKitrick, R., & Christy, J. (2020). Pervasive warming bias in CMIP6 tropospheric layers. *Earth and Space Science*. doi: 10.1029/2020ea001281

McPeters, R. D., Bhartia, P. K., Haffner, D., Labow, G. J., & Flynn, L. (2013). The version 8.6 SBUV ozone data record: An overview. *Journal of Geophysical Research: Atmospheres*, *118*(14), 8032–8039. (V8.7 data downloaded from `https://acd-ext.gsfc.nasa.gov/Data_services/merged/data/sbuv_v87_mod.int_lyr.70-20.za.r1.txt`. [Dataset]) doi: 10.1002/jgrd.50597

Mears, C. A., Schabel, M., & Wentz, F. J. (2003). A reanalysis of the MSU Channel 2 tropospheric temperature record. *J. Clim.*, *16*, 3650–3664.

Mears, C. A., & Wentz, F. J. (2016). Sensitivity of satellite-derived tropospheric temperature trends to the diurnal cycle adjustment. *Journal of Climate*, *29*(10), 3629–3646. (Data downloaded from `ftp://ftp.remss.com/msu/graphics/` (free registration required). [Dataset]) doi: 10.1175/jcli-d-15-0744.1

Mears, C. A., Wentz, F. J., & Thorne, P. W. (2012). Assessing the value of Microwave Sounding Unit-radiosonde comparisons in ascertaining errors in climate data records of tropospheric temperatures. *Journal of Geophysical Research: Atmospheres*, *117*. doi: 10.1029/2012jd017710

Menon, S., Koch, D., Beig, G., Sahu, S., Fasullo, J., & Orlikowski, D. (2010). Black carbon aerosols and the third polar ice cap. *Atmos. Chem. Phys.*, *10*, 4559–4571.

Menon, S., & Rotstayn, L. (2006). The radiative influence of aerosol effects on liquid-phase cumulus and stratiform clouds based on sensitivity studies with two climate models. *Climate Dynamics*, *27*(4), 345–356. doi: 10.1007/s00382-006-0139-3

Menon, S., Unger, N., Koch, D., Francis, J., Garrett, T., Sednev, I., ... Streets, D. (2008). Aerosol climate effects and air quality impacts from 1980 to 2030. *Environ. Res. Lett.*, *3*. doi: 10.1088/1748-9326/3/2/024004

Miller, R. L., Schmidt, G. A., Nazarenko, L., Bauer, S. E., Kelley, M., Ruedy, R., ... Yao, M.-S. (2021). CMIP6 historical simulations (1850-2014) with GISS ModelE2.1. *J. Adv. Model. Earth Syst.*, *13*. (Radiative forcing data downloaded from `https://data.giss.nasa.gov/modelforce/Miller_et_al21/ERFs_SSP245_MillerFig10_2021.txt`. [Dataset]) doi: 10.1029/2019MS002034

Miller, R. L., Schmidt, G. A., Nazarenko, L. S., Tausnev, N., Ruedy, R., Kelley, M., ... Zhang, J. (2014). CMIP5 historical simulations (1850–2012) with GISS ModelE2. *J. Adv. Model. Earth Syst.*, *6*, 441–477. doi: 10.1002/2013MS000266

Mitchell, D. M., Lo, Y. T. E., Seviour, W. J. M., Haimberger, L., & Polvani, L. M. (2020). The vertical profile of recent tropical temperature trends: Persistent model biases in the context of internal variability. *Environmental Research Letters*, *15*. doi: 10.1088/1748-9326/ab9af7

Mitchell, D. M., Thorne, P. W., Stott, P. A., & Gray, L. J. (2013). Revisiting the controversial issue of tropical tropospheric temperature trends. *Geophysical Research Letters*, *40*, 2801–2806. doi: 10.1002/grl.50465

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., ... Simpson, I. R. (2021). An updated assessment of near-surface temperature change from 1850: The HadCRUT5 data set. *Journal of Geophysical Research: Atmospheres*, *126*. doi: 10.1029/2019jd032361

NASA Center for Climate Simulation. (2022). National Aeronautics and Space Administration. Retrieved from `https://portal.nccs.nasa.gov/datashare/giss_cmip6/[Dataset]`

Nasa ozone watch. (2022). National Aeronautics and Space Administration. Retrieved from `https://ozonewatch.gsfc.nasa.gov/statistics/annual_data.txt`

Nazarenko, L. S., Tausnev, N., Russell, G. L., Rind, D., Miller, R. L., Schmidt, G. A., ... Yao, M.-S. (2022). Future climate change under SSP emission scenarios with GISS-E2.1. *J. Adv. Model. Earth Syst.*. doi: 10.1029/2021ms002871

Newman, P. A., Daniel, J. S., Waugh, D. W., & Nash, E. R. (2007). A new formulation of equivalent effective stratospheric chlorine (EESC). *Atmospheric Chemistry and Physics*, *7*(17), 4537–4552. doi: 10.5194/acp-7-4537-2007

Orbe, C., Rind, D., Jonas, J., Nazarenko, L., Faluvegi, G., Murray, L. T., . . . Schmidt, G. A. (2020). GISS model E2.2: A climate model optimized for the middle atmosphere—2. Validation of large-scale transport and evaluation of climate response. *Journal of Geophysical Research: Atmospheres*, *125*(24). doi: 10.1029/2020jd033151

Po-Chedley, S., & Fu, Q. (2012). Discrepancies in tropical upper tropospheric warming between atmospheric circulation models and satellites. *Environmental Research Letters*, *7*. doi: 10.1088/1748-9326/7/4/044018

Po-Chedley, S., Santer, B. D., Fueglistaler, S., Zelinka, M. D., Cameron-Smith, P. J., Painter, J. F., & Fu, Q. (2021). Natural variability contributes to model–satellite differences in tropical tropospheric warming. *Proceedings of the National Academy of Sciences*, *118*(13). doi: 10.1073/pnas.2020962118

Ramaswamy, V., Schwarzkopf, M. D., & Randel, W. J. (1996). Fingerprint of ozone depletion in the spatial and temporal pattern of recent lower-stratospheric cooling. *Nature*, *382*(6592), 616–618. doi: 10.1038/382616a0

Randel, W. J., & Cobb, J. B. (1994). Coherent variations of monthly mean total ozone and lower stratospheric temperature. *Journal of Geophysical Research*, *99*(D3), 5433. doi: 10.1029/93jd03454

Rayner, N. A., Brohan, P., Parker, D. E., Folland, C. K., Kennedy, J. J., Vanicek, M., . . . Tett, S. F. B. (2006). Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. *Journal of Climate*, *19*(3), 446–469. doi: 10.1175/jcli3637.1

Richardson, M., Cowtan, K., & Millar, R. J. (2018). Global temperature definition affects achievement of long-term climate goals. *Environmental Research Letters*, *13*(5), 054004. doi: 10.1088/1748-9326/aab305

Rind, D., Orbe, C., Jonas, J., Nazarenko, L., Zhou, T., Kelley, M., . . . Schmidt, G. A. (2020). GISS model E2.2: A climate model optimized for the middle atmosphere. Model structure, climatology, variability and climate sensitivity. *Journal of Geophysical Research: Atmospheres*. doi: 10.1029/2019jd032204

Santer, B. D., Bonfils, C., Painter, J. F., Zelinka, M. D., Mears, C., Solomon, S., . . . Wentz, F. J. (2014). Volcanic contribution to decadal changes in tropospheric temperature. *Nature Geosci.*, *7*(3), 185–189. doi: 10.1038/ngeo2098

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

Santer, B. D., Po-Chedley, S., Mears, C., Fyfe, J. C., Gillett, N., Fu, Q., . . . Zou, C.-Z. (2021). Using climate model simulations to constrain observations. *Journal of Climate*, 1–59. doi: 10.1175/jcli-d-20-0768.1

Santer, B. D., Solomon, S., Pallotta, G., Mears, C., Po-Chedley, S., Fu, Q., . . . Bonfils, C. (2017). Comparing tropospheric warming in climate models and satellite data. *Journal of Climate*, *30*(1), 373–392. doi: 10.1175/jcli-d-16-0333.1

Santer, B. D., Taylor, K. E., Wigley, T. M. L., Johns, T. C., Jones, P. D., Karoly, D. J., . . . Tett, S. (1996). A search for human influences on the thermal structure of the atmosphere. *Nature*, *382*(6586), 39–46. doi: 10.1038/382039a0

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M. L., Lanzante, J. R., . . . Wentz, F. J. (2008). Consistency of modelled and observed temperature trends in the tropical troposphere. *International Journal of Climatology*, *28*(13), 1703–1722. doi: 10.1002/joc.1756

Santer, B. D., Wigley, T. M. L., Mears, C., Wentz, F. J., Klein, S. A., Seidel, D. J., . . . Schmidt, G. A. (2005). Amplification of surface temperature trends and variability in the tropical atmosphere. *Science*, *309*(5740), 1551–1556. doi: 10.1126/science.1114867

Sato, M., Hansen, J. E., McCormick, M. P., & Pollack, J. B. (1993). Stratospheric aerosol optical depths, 1850–1990. *J. Geophys. Res.*, *98*, 22,987–22,994. (Updated data downloaded from `https://data.giss.nasa.gov/modelforce/strataer/tau.line_2012.12.txt`. [Dataset])

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., … Zhang, J. (2014). Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *J. Adv. Model. Earth Syst.*, *6*, 141–184. doi: 10.1002/2013MS000265

Schmidt, G. A., Ruedy, R., Hansen, J. E., Aleinov, I., Bell, N., Bauer, M., … Yao, M.-S. (2006). Present-day atmospheric simulations using GISS ModelE:Comparison to in situ, satellite, and reanalysis data. *Journal of Climate*, *19*, 153–192. doi: 10.1175/jcli3612.1

Seidel, D. J., Li, J., Mears, C., Moradi, I., Nash, J., Randel, W. J., … Zou, C.-Z. (2016). Stratospheric temperature changes during the satellite era. *Journal of Geophysical Research: Atmospheres*, 664—681. doi: 10.1002/2015jd024039

Shah, K. P., & Rind, D. (1995). Use of microwave brightness temperatures with a general circulation model. *J. Geophys. Res.*, *100*, 13,841–13,874.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 683–690. doi: 10.1111/j.2517-6161.1991.tb01857.x

Simmons, A., Hersbach, H., Munoz-Sabater, J., Nicolas, J., Vamborg, F., Berrisford, P., … Woollen, J. (2021). *Low frequency variability and trends in surface air temperature and humidity from ERA5 and other datasets.* ECMWF. doi: 10.21957/LY5VBTBFD

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

Spencer, R. W., Christy, J. R., & Braswell, W. D. (2017). UAH version 6 global satellite temperature products: Methodology and results. *Asia-Pacific Journal of Atmospheric Sciences*, *53*(1), 121–130. doi: 10.1007/s13143-017-0010-y

Steiner, A. K., Ladstädter, F., Randel, W. J., Maycock, A. C., Fu, Q., Claud, C., … Zou, C.-Z. (2020). Observed temperature changes in the troposphere and stratosphere from 1979 to 2018. *Journal of Climate*, *33*(19), 8165–8194. doi: 10.1175/jcli-d-19-0998.1

Taylor, K. E., Williamson, D., & Zwiers, F. (2000). *The sea surface temperature and sea ice concentration boundary conditions for AMIP II simulations.* PCMDI Report 60, Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory. Retrieved from `https://pcmdi.llnl.gov/mips/amip/index.html`

Thompson, D. W. J., Seidel, D. J., Randel, W. J., Zou, C.-Z., Butler, A., Mears, C., … Lin, R. (2012). The mystery of recent stratospheric temperature trends. *Nature*, *491*, 692–697. doi: 10.1038/nature11579

Thorne, P. W., Brohan, P., Titchner, H. A., McCarthy, M. P., Sherwood, S. C., Peterson, T. C., … Kennedy, J. J. (2011). A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *Journal of Geophysical Research*, *116*. doi: 10.1029/2010jd015487

Timmreck, C., Mann, G. W., Aquila, V., Hommel, R., Lee, L. A., Schmidt, A., … Weisenstein, D. (2018). The interactive stratospheric aerosol model intercomparison project (isa-mip): motivation and experimental design. *Geoscientific Model Development*, *11*(7), 2581–2608. doi: 10.5194/gmd-11-2581-2018

Vose, R. S., Huang, B., Yin, X., Arndt, D., Easterling, D. R., Lawrimore, J. H., … Zhang, H. M. (2021). Implementing full spatial coverage in NOAA's global temperature analysis. *Geophysical Research Letters*, *48*. doi: 10.1029/2020gl090873

Wentz, F. J., & Schabel, M. (1998). Effects of orbital decay on satellite-derived lower-tropospheric temperature trends. *Nature*, *394*, 661–664. doi: 10.1038/29267

Wigley, T. M. L. (2006). Appendix A: Statistical issues regarding trends. In T. R. Karl, S. J. Hassol, C. D. Miller, & W. L. Murray (Eds.), *Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences.* U.S. Climate Change Science Program and the Subcommittee on Global Change Research, Washington DC.

Zanchettin, D., Timmreck, C., Khodri, M., Schmidt, A., Toohey, M., Abe, M., . . . Weierbach, H. (2022). Effects of forcing differences and initial conditions on inter-model agreement in the VolMIP volc-pinatubo-full experiment. *Geoscientific Model Development*, *15*(5), 2265–2292. doi: 10.5194/gmd-15-2265-2022

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., . . . Taylor, K. E. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*. doi: 10.1029/2019gl085782

Zou, C.-Z., & Qian, H. (2016). Stratospheric temperature climate data record from merged SSU and AMSU-A observations. *Journal of Atmospheric and Oceanic Technology*, *33*(9), 1967–1984. (Data downloaded from `ftp://ftp.star.nesdis.noaa.gov/pub/smcd/emb/mscat/data/`. [Dataset]) doi: 10.1175/JTECH-D-16-0018.1