Multi-Frequency Analysis of Simulated versus Observed Variability in Tropospheric Temperature

GIULIANA PALLOTTA^a AND BENJAMIN D. SANTER^b

^a *Applied Statistics Group, Computational Engineering Division, Lawrence Livermore National Laboratory, Livermore, California;*

^b *Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, California*

(Manuscript received 13 January 2020, in final form 31 July 2020)

ABSTRACT: Studies seeking to identify a human-caused global warming signal generally rely on climate model estimates of the “noise” of intrinsic natural variability. Assessing the reliability of these noise estimates is of critical importance. We evaluate here the statistical significance of differences between climate model and observational natural variability spectra for global-mean mid- to upper-tropospheric temperature (TMT). We use TMT information from satellites and large multimodel ensembles of forced and unforced simulations. Our main goal is to explore the sensitivity of model-versus-data spectral comparisons to a wide range of subjective decisions. These include the choice of satellite and climate model TMT datasets, the method for separating signal and noise, the frequency range considered, and the statistical model used to represent observed natural variability. Of particular interest is the amplitude of the interdecadal noise against which an anthropogenic tropospheric warming signal must be detected. We find that on time scales of 5–20 years, observed TMT variability is (on average) overestimated by the last two generations of climate models participating in the Coupled Model Intercomparison Project. This result is relatively insensitive to different plausible analyst choices, enhancing confidence in previous claims of detectable anthropogenic warming of the troposphere and indicating that these claims may be conservative. A further key finding is that two commonly used statistical models of short-term and long-term memory have deficiencies in their ability to capture the complex shape of observed TMT spectra.

KEYWORDS: Spectral analysis/models/distribution; Statistics; Climate variability

1. Introduction

The coupled atmosphere–ocean system exhibits intrinsic variability on a wide range of space and time scales. Examples include such phenomena as the Madden–Julian oscillation (MJO), El Niño–Southern Oscillation (ENSO), and the Atlantic multidecadal oscillation (AMO). Natural internal variability occurs in the absence of any human-caused changes in atmospheric composition, and constitutes the background noise against which any slowly evolving human-caused warming signal must be detected. Estimates of the spectrum of internal variability are a critical component of anthropogenic signal detection studies (Hasselmann 1979; Bloomfield and Nychka 1992; Hegerl et al. 2007; Bindoff et al. 2013; Imbers et al. 2014).

Reliable estimation of interdecadal internal variability from short observational records is a challenging task (Frankcombe et al. 2015, 2018; Cheung et al. 2017a,b; Kravtsov 2017; Kravtsov et al. 2018; Kajtar et al. 2019). One challenge is that observations are a complex mixture of internal variability and

externally forced signals. Isolating internal variability requires estimation of the individual or combined signals associated with many different external factors. These factors include purely natural changes in solar irradiance and volcanic activity (Kopp and Lean 2011; Solomon et al. 2011) as well as human-caused changes in well-mixed greenhouse gases (GHGs), stratospheric ozone, and particulate pollution (Myhre et al. 2013). Additional difficulties in estimating intrinsic climate variability arise from nonrandom residual errors in observations (Mears and Wentz 2016; Po-Chedley et al. 2015; Karl et al. 2015) and from possible modulation of internal variability by external influences (Maher et al. 2015; Pausata et al. 2015).

In climate models, however, “pure” internal variability can be estimated from multicentury control runs with no secular changes in external factors. Internal variability can also be inferred by randomly perturbing the initial climate state in a large ensemble (LE) of externally forced simulations performed with a single model (Deser et al. 2014, 2020; Fyfe et al. 2017). Variability information from control runs and LEs is an integral part of most modern studies seeking to identify anthropogenic signals in observed climate data (Hegerl et al. 2007; Bindoff et al. 2013; Fyfe et al. 2017; Swart et al. 2018; Santer et al. 2019).

Our focus here is on assessing how comparisons of simulated and observed natural variability spectra are affected by uncertainties in data, climate models, and the separation of signal and noise. We also explore the impact of using different statistical models to characterize observed natural variability (Bloomfield and Nychka 1992; Imbers et al. 2014). We are particularly interested in determining whether current climate

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JCLI-D-20-0023.s1>.

Corresponding author: Giuliana Pallotta, pallottagold1@llnl.gov

DOI: 10.1175/JCLI-D-20-0023.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

models systematically underestimate the amplitude of observed natural variability of global-mean mid- to upper-tropospheric temperature (TMT) on time scales of 1–2 decades. If such a bias existed, it would imply that signal-to-noise (S/N) ratios had been spuriously inflated in previous anthropogenic signal detection studies with TMT (Santer et al. 2013a, 2019).

We rely on tropospheric temperature for multiple reasons. First, considerable scientific and political attention has been focused on the question of whether satellite TMT datasets show statistically significant warming (Christy 2015; Santer et al. 2017a,b). Answering this question requires information on the credibility of model estimates of natural TMT variability (Santer et al. 2018). Second, unlike surface temperature records obtained from land thermometers, ships of opportunity, and ocean buoys, satellite TMT measurements have time-invariant near-global coverage. This is advantageous for reliable estimation of variability (AchutaRao et al. 2006). Third, structural uncertainties in observed and modeled TMT variability can be well characterized: information on global-scale TMT changes is available from three satellite research groups (Mears and Wentz 2016; Zou et al. 2018; Spencer et al. 2017), and “synthetic” TMT has been calculated from over three dozen climate models (Santer et al. 2018).

Operating in the frequency domain has a number of advantages (Vyushin and Kushner 2009). Spectral analysis facilitates model-versus-data comparisons of the overall spectral “shape,” the interdecadal variance relevant to detection of an anthropogenic TMT signal, and the phase and amplitude of specific spectral peaks. Calculating spectra also allows analysts to determine whether the partitioning of variance as a function of frequency is well described by a simple power law (Hasselmann 1976). Consistent power-law relationships in models and observations would provide support for efforts to use the well-observed amplitude of climate variability on interannual time scales to constrain the more uncertain amplitude of variability on interdecadal time scales (Santer et al. 2013a; Fredriksen and Rypdal 2017).

“Observed” natural variability is generally inferred by removing estimates of forced signals from climate records (Frankcombe et al. 2015, 2018). After signal removal, the behavior of the observed temperature residuals is often represented by one or more statistical models (Bloomfield and Nychka 1992; Imbers et al. 2014). This step is necessary because there is only one realization of observed interdecadal internal variability. Statistical modeling allows analysts to generate thousands of surrogate observations. Each surrogate is consistent with certain statistical properties of the “signal-removed” observational data, such as the short-term and/or longer-term persistence. Sampling distributions of surrogate properties can then be used to evaluate the statistical significance of observed temperature trends (Imbers et al. 2014) or to test the significance of differences between natural variability spectra obtained from observations and a climate model.

A variety of statistical models have been used in such studies, ranging from simple low-order autoregressive (AR) models to more sophisticated autoregressive fractionally integrated moving average (FARIMA) models (Bloomfield and Nychka 1992; Imbers et al. 2014; Franzke 2012b; Franzke et al.

2015). Other common statistical representations of internal variability involve calculation of Hurst exponents (Hurst 1950; Vyushin et al. 2009; Zhu et al. 2010; Mann 2011) or power-law fits to spectra of the signal-removed temperature residuals (Zhu et al. 2019). The challenge is to determine the most appropriate statistical model of interdecadal internal variability given relatively short observational records and uncertainties in estimating and removing externally forced signals from observations.

Most previous comparisons of simulated and observed climate spectra have focused on global-mean surface temperature (GMST) (Bloomfield and Nychka 1992; Imbers et al. 2014), instrumental surface temperature measurements from individual stations (Franzke 2012b; Franzke et al. 2015), and paleoclimate temperature reconstructions (Zhu et al. 2019). Other model–data comparisons examine modes of internal variability that contribute to observed GMST variability (Steinman et al. 2015; Cheung et al. 2017a; Kim et al. 2018). While some of these investigations suggest that climate models do not systematically underestimate the amplitude of observed variability on interdecadal time scales (Hegerl et al. 2007; Bindoff et al. 2013; Imbers et al. 2014; Franzke et al. 2015; Fredriksen and Rypdal 2016, 2017; Zhu et al. 2019), other studies yield opposite findings (Cheung et al. 2017a; Kim et al. 2018; Kravtsov et al. 2018). Differences are related to the spatial scale of the analysis, the method used for partitioning signal and internal variability, the time scales of interest, and the models and observations selected. Given these conflicting results, it is of interest to explore the size and direction of model variability biases with independently monitored TMT data.

For both GMST and TMT, there are few studies that have evaluated the statistical significance of differences between modeled and observed climate spectra (Alvarez-Esteban et al. 2016). To date, only a study by Gillett et al. (2000) explicitly tested whether simulated and observed tropospheric temperature spectra are significantly different. The Gillett et al. investigation evaluated the significance of variance ratios at individual frequencies, relying on spectra calculated from an individual model control run and a single set of weather balloon temperature measurements.

We update and extend the work of Gillett et al. (2000) using multimodel ensembles of simulations from earlier and later phases of the Coupled Model Intercomparison Project (CMIP), multiple satellite TMT datasets, different signal removal strategies, and different statistical models to represent observed natural variability. Our goal is to develop a general framework that is applicable to other climate variables and is suitable for testing the significance of differences between modeled and observed spectra under a wide range of analyst choices.

2. Observations and model simulations

The observed TMT data analyzed here are obtained from the Microwave Sounding Unit (MSU) and Advanced Microwave Sounding Unit (AMSU). These instruments measure the microwave emissions from oxygen molecules, and have been flown on over a dozen polar-orbiting weather satellites (Mears and Wentz 2016). Emissions are proportional to the temperatures of broad atmospheric layers. By measuring at different microwave

frequencies, the temperatures of different layers can be retrieved. MSU and AMSU provide a 41-yr record of tropospheric temperature with near-global coverage (Mears and Wentz 2016; Spencer et al. 2017; Zou et al. 2018).

We use the latest versions of satellite TMT datasets from three research groups: Remote Sensing Systems (RSS), the Center for Satellite Applications and Research (STAR), and the University of Alabama at Huntsville (UAH). At the time this study was performed, data were available from January 1979 to December 2018. Differences between these three TMT products are primarily due to differences in each group's adjustments for the effects of drifts in satellite orbits and radiometer calibrations (Mears and Wentz 2016; Po-Chedley et al. 2015).

Synthetic TMT data are from simulations performed under the older phase 5 and the newer phase 6 of the Coupled Model Intercomparison Project (CMIP5 and CMIP6) (Taylor et al. 2012; Eyring et al. 2016). These model intercomparison efforts have a wide range of scientific goals, including identifying systematic model errors, quantifying uncertainties in projections of future climate change, and improving understanding of key physical processes in individual components of the climate system. While the scientific community has examined CMIP5 output for over a decade, the CMIP6 simulation archive is still being populated. Our primary focus, therefore, is on the existing CMIP5 simulation archive. Model–data comparisons involving CMIP5 forced and unforced simulations are discussed in detail in sections 4–8. CMIP6 results are treated more briefly in section 9. Our analysis of TMT variability in CMIP6 is restricted to forced simulations.

Preindustrial control runs (CTL) have no year-to-year changes in anthropogenic or natural external forcings, and provide estimates of “pure” natural internal variability. These estimates were available from 36 different CMIP5 models. CMIP5 control runs have been used to evaluate the significance of observed global-mean TMT trends (Santer et al. 2017b) and to identify anthropogenic fingerprint patterns in satellite TMT data (Santer et al. 2013a, 2018).

The externally forced experiments analyzed here are simulations of historical climate change (HIST). The HIST runs are forced by time-varying anthropogenic changes in well-mixed GHGs, stratospheric ozone, particulate pollution, and land surface properties, as well as by natural changes in solar irradiance and volcanic aerosols. The HIST integrations in CMIP5 and CMIP6 end in December 2005 and in December 2014, respectively. To facilitate comparison with satellite data over 1979–2018, HIST simulations were extended with results from two scenarios of twenty-first-century climate change: the representative concentration pathway 8.5 (RCP8.5) (Meinshausen et al. 2011) for CMIP5 and the shared socioeconomic pathway 5 (SSP5) for CMIP6 (Riahi et al. 2017). We refer to these subsequently as HIST+RCP8.5 and HIST+SSP5, respectively.

Both RCP8.5 and SSP5 assume continued exploitation of fossil fuels for economic development, and are comparable in terms of their twenty-first-century GHG emissions (Riahi et al. 2017). While alternative RCP and SSP scenarios with different emissions assumptions were available for splicing with the HIST integrations, all scenarios have highly similar GHG

emissions over the 2005–18 period (van Vuuren et al. 2011; Riahi et al. 2017). Splicing HIST with different scenarios than those selected here has minimal impact on our model-versus-data variability comparisons.

The HIST+RCP8.5 and HIST+SSP5 integrations were performed with 37 different CMIP5 and 21 different CMIP6 models (respectively). In each ensemble, the spread in tropospheric warming arises from model differences in multiple factors: the applied external forcings (particularly the uncertain forcing associated with anthropogenic aerosols), the responses to these forcings, and the amplitude and phase of multidecadal internal variability (Zelinka et al. 2014; Hawkins et al. 2016; Santer et al. 2017a).

We compare observed natural variability spectra with both CTL and extended HIST results. A brief explanation for this decision is warranted. In most previous efforts to isolate observed internal variability, an estimated anthropogenic signal is removed from observations (Bloomfield and Nychka 1992; Imbers et al. 2014; Frankcombe et al. 2015, 2018). Even if this signal accurately captured the true (but uncertain) anthropogenic component of temperature change, the residuals would not reflect internal variability alone—they would also include the short-term (2–3-yr) cooling caused by major volcanic eruptions and the effects of solar irradiance changes over the roughly 11-yr solar cycle. This total natural variability V_{TOT} is larger than the internally generated variability V_{INT} (Santer et al. 2013b).

Removing a simple statistical representation of the anthropogenic signal (such as a linear or low-order polynomial fit) from both the observations and the extended HIST runs facilitates a direct comparison of their V_{TOT} spectra. If the observed signal-removed TMT residuals are dominated by internal variability, it is reasonable to compare spectra of observed V_{TOT} and CTL V_{INT} . Using both CTL and extended HIST runs allows us to study the sensitivity of model-versus-observed spectral differences to the choice of simulation type.

Further details of the satellite and model TMT data analyzed here are provided in sections 1 and 2 and Tables S1–S3 in the online supplemental material. Appendix A describes how we calculate synthetic MSU temperatures from model simulations. Appendix B summarizes the regression-based method (Fu et al. 2004) we use to adjust observed and synthetic TMT information for the contribution TMT receives from anthropogenic cooling of the lower stratosphere (Solomon et al. 2016). Appendixes C and D cover the calculation of TMT anomalies and the estimation of power spectral density (respectively).

3. Signal removal methods

A variety of different methods have been applied for separating externally forced temperature signals from internal variability [see Frankcombe et al. (2015) for a summary]. The simplest methods involve fitting least squares linear trends (Bloomfield and Nychka 1992; Imbers et al. 2014) or low-order polynomials to observations or extended HIST simulations. Other methods involve signal estimates obtained from individual Earth system models (ESMs), the multimodel average (MMA) of a large ensemble of ESMs, and large initial-condition ensembles (LEs) performed with a single ESM

(Knutson et al. 2013; Kravtsov 2017; Deser et al. 2020). So-called semiempirical approaches rely on scaled versions of the MMA (Steinman et al. 2015; Cheung et al. 2017a; Kajtar et al. 2019). Simpler multibox energy-balance models (EBMs) have also been used for estimating and removing forced signals (Rypdal and Rypdal 2014; Fredriksen and Rypdal 2017).

Our study is not intended to be a comprehensive exploration of all currently available signal removal methods. Instead, our primary interest is in exploring the impact of a small number of commonly used signal removal options on the consistency between simulated and observed natural variability estimates. A more formal intercomparison of the efficacy of different signal removal methods would rely on information from a large initial condition ensemble. Because the true signal and noise properties of an individual model can be well estimated in an LE, large ensembles are ideal test beds for quantifying the error characteristics of different signal removal approaches. The benefit of such “LE as test bed” studies would be enhanced by use of multiple LEs, spanning a range of model differences in equilibrium climate sensitivity (ECS), anthropogenic aerosol forcing, and internal variability properties (Santer et al. 2019).

We apply four different methods to remove externally forced temperature signals from the observations and the extended HIST runs: removal of linear, quadratic, and cubic fits and subtraction of the unscaled multimodel average. In the linear, quadratic, and cubic signal removal cases (henceforth LIN, QUAD, and CUB), we remove a polynomial up to third order. Signals are estimated separately for each satellite and extended HIST time series. This procedure reduces overall differences in tropospheric warming rates between models and observations, between observational datasets, and between CMIP models with different climate sensitivity.

The MMA removal approach requires more detailed explanation. The phasing and amplitude of internal variability is uncorrelated (except by chance) across individual realizations of the extended HIST experiment performed with a single model. It is also uncorrelated across different models. Internal variability is damped by averaging over individual realizations and models, more clearly revealing the underlying TMT signal in response to combined natural and anthropogenic external forcing (see magenta line in Figs. 1a–c).

In the CMIP5 case, application of the unscaled MMA signal removal method to observed TMT data requires at least three assumptions: 1) the average ECS of CMIP5 models, roughly 3.2°C (Andrews et al. 2012), approximates the true but uncertain real world ECS; 2) the MMA TMT time series provides a reliable estimate of the true (but uncertain) time-varying response to combined anthropogenic and natural external forcing; and 3) satellite TMT datasets do not contain large inhomogeneities that affect the overall magnitude and evolution of their tropospheric warming. Underlying assumption 2 are two further assumptions: that the 37 models used to estimate the CMIP5 MMA constitute a reasonably independent sample (Weigel et al. 2010) and that application of an objective model weighting scheme (Eyring et al. 2019) would yield TMT changes similar to those of the unweighted MMA.

If any of these assumptions are unjustified, subtraction of the MMA from observations will not cleanly isolate “observed” internal variability. For example, if the model-average ECS value of 3.2°C were appreciably larger or smaller than the real-world ECS, the observed TMT residuals after MMA subtraction would contain trends. These trends would inflate the power at the lowest frequencies resolved.

Similar problems arise when the unscaled MMA is subtracted from HIST+RCP8.5 simulations performed with models that have substantially higher or lower ECS than 3.2°C. For the CMIP5 models considered here, ECS ranges from 2° to 4.7°C (Andrews et al. 2012; Zelinka et al. 2014). In such high and low ECS cases, removing the unscaled MMA from individual model HIST+RCP8.5 runs can spuriously inflate estimates of simulated interdecadal internal variability.

These well-recognized deficiencies in the unscaled MMA motivated efforts to account for model ECS differences. Scaling based on model ECS alone is problematic, however, in the case of models with both high ECS and large negative anthropogenic aerosol forcing (Santer et al. 2019). Other scaling approaches typically rely on regression between the MMA and one or more modes of internal variability (Steinman et al. 2015; Frankcombe et al. 2015, 2018; Cheung et al. 2017a,b; Kajtar et al. 2019). Removal of this type of scaled version of the MMA from an individual model may produce internal variability estimates that differ from those obtained when the individual model’s true forced signal is removed¹ (Kravtsov 2017). We do not rely on any form of MMA scaling here, and for most CMIP5 models we do not have synthetic satellite temperatures from a sufficient number of HIST+RCP8.5 realizations to obtain a reliable estimate of the forced signal (see Table S2).

Despite the obvious problems with removal of the unscaled MMA, we regard this as a useful sensitivity test. There are two reasons for this. First, subtracting the unscaled MMA from individual models with different ECS values inflates the model range of interdecadal variability in the extended HIST runs. This provides valuable information on the impact of large, known ECS errors on the low-frequency portion of the model TMT spectra. It is of interest to determine whether interdecadal variability in the “MMA-removed” satellite data falls within the model range.

Second, if any of the three above-noted assumptions underlying application of the MMA are unjustified, the MMA-removed observational TMT data will have large low-frequency residuals, and thus (through the long-term memory models that are fit to these residuals) will expand the range of estimated natural interdecadal variability. This implicitly allows for a contribution to the observed interdecadal variability from time scales >20 years, which are poorly sampled in the relatively short single realization of the satellite TMT record (Kravtsov et al. 2018).

¹ Assuming that a sufficiently number of realizations are available to estimate the true forced signal.

FIG. 1. (a)–(c) Raw time series of near-global averages of monthly mean TMT for the three satellite datasets analyzed here (RSS, STAR, and UAH, respectively), the individual HIST+RCP8.5 simulations performed with 37 CMIP5 models, and the HIST+RCP8.5 multimodel average (MMA). (e)–(g) TMT residuals for RSS, STAR, and UAH, respectively, after MMA removal. (i)–(k) TMT residuals for RSS, STAR, and UAH, respectively, after linear detrending. The start dates of the El Chichón and Pinatubo eruptions are denoted by black triangles and dotted vertical lines. (d), (h), (l) The estimated power spectral density (PSD) for the raw, MMA-removed, and linearly detrended time series, respectively. The CMIP5 HIST+RCP8.5 ensemble-mean spectrum is the mean of the spectra across different models. The shaded area in magenta in (d), (h), and (l) corresponds to the 5%–95% PSD range in the HIST+RPC8.5 model ensemble.

4. Comparing satellite and CMIP5 time series and spectra

The satellite and CMIP5 extended HIST time series analyzed here exhibit gradual warming of the global troposphere (Figs. 1a–c). Warming is punctuated by short-term (2–3-yr) cooling associated with the major eruptions of El Chichón in 1982 and Pinatubo in 1991. Because averaging over realizations and models damps internal variability (see above), volcanic cooling is clearer in the MMA than in observational TMT data, where it is partly obscured by warming associated with El Niño events (Santer et al. 2018).

The average TMT trend over 1979 to 2018 is $0.29^{\circ}\text{C decade}^{-1}$ in the 37 CMIP5 models analyzed here; the smallest and largest trends in individual realizations of the multimodel ensemble are 0.13° and $0.45^{\circ}\text{C decade}^{-1}$, respectively. Satellite tropospheric warming trends are consistently smaller than the average of the model results, ranging from $0.14^{\circ}\text{C decade}^{-1}$ in UAH to $0.22^{\circ}\text{C decade}^{-1}$ in STAR. Simulated tropospheric warming that generally exceeds observed warming rates also characterizes CMIP6 extended HIST simulations (see section 9) and two large initial condition ensembles (Santer et al. 2019).

Differences between observed TMT trends and simulated trends in CMIP5 are due to at least four different factors: 1) known systematic errors in the early twenty-first-century volcanic and solar forcing used in CMIP5 (Solomon et al. 2011; Kopp and Lean 2011; Santer et al. 2017a); 2) model errors in the response to forcing (Trenberth and Fasullo 2010); 3) differences in the phasing of simulated and observed decadal variability (Meehl et al. 2014; England et al. 2014; Fyfe et al. 2016); and 4) remaining inhomogeneities in observations (Pochedley et al. 2015; Mears and Wentz 2016). These factors are not mutually exclusive. A continuing scientific challenge is reliable quantification of each factor's contribution to model-versus-observed warming rate differences.

The different impacts of applying the MMA and LIN signal removal methods are visually obvious in Figs. 1e–g and Figs. 1i–k (respectively). In the MMA case, each observed residual TMT time series has complex low-frequency behavior: the average value of the residuals is close to zero in the last several decades of the twentieth century and is negative in the early twenty-first century. These secular changes in the residuals are unlikely to be due to internal variability alone

(Santer et al. 2017a). In contrast, removal of a linear signal from each dataset does not yield noticeable differences between the behavior of the residuals in the late twentieth and early twenty-first centuries.

Figures 1d, 1h, and 1i illustrate how spectra are affected by operating on raw TMT time series, removing the MMA, and applying LIN signal removal (respectively). Consider the observational results first. In the raw data, tropospheric warming is larger in RSS and STAR than in UAH. These warming rate differences are aliased at low frequencies. This is why RSS and STAR have higher power than UAH in the 5–20-yr frequency band (Fig. 1d).

Removal of the unscaled MMA produces the opposite result. The difference between tropospheric warming trends in UAH and the multimodel average is about $-0.15^{\circ}\text{C decade}^{-1}$. This is larger than the observed-minus-MMA trend difference of roughly -0.10° and $-0.07^{\circ}\text{C decade}^{-1}$ for RSS and STAR, respectively. With MMA removal, therefore, low-frequency variability is amplified in UAH relative to RSS and STAR. In the LIN case, the spectra for RSS, STAR and UAH are highly similar because overall differences in observed warming rates are removed. The same result holds for QUAD and CUB signal removal.

Comparisons between the observed and HIST+RCP8.5 spectra reveal interesting similarities and differences (Figs. 1d,h,i). There is close agreement between simulated and observed power spectral density (PSD) on periods ranging from several months to several years. This agreement is relatively insensitive to whether spectra are calculated from raw data or are compared after MMA or LIN signal removal.

On time scales between 5 and 20 years, however, the average power in the 37 HIST+RCP8.5 simulations is almost always larger than in the satellite data, consistent with time domain analyses (Santer et al. 2011, 2013a, 2018). The sole exception is for UAH data and MMA signal removal. In this particular case, the UAH variance on 20-yr time scales is above the upper end of the 5%–95% confidence interval on the HIST+RCP8.5 results (Fig. 1h). This highlights the unusual nature of the UAH TMT time series: even large climate sensitivity errors (see above) are unlikely to produce the UAH-minus-MMA low-frequency behavior.

5. Comparing CMIP5 control and HIST+RCP8.5 spectra for individual models

Figures 1d, 1h, and 1i show spectra averaged across the CMIP5 multimodel ensemble. In this section, we examine individual model spectra for preindustrial control runs and extended HIST simulations (Fig. 2). The HIST+RCP8.5 spectra are for temperature residuals obtained by applying the four signal removal methods described in section 3. Comparison of the CTL spectra and signal-removed HIST+RCP8.5 spectra provides information on the impact of different signal removal strategies on variability estimates, and on the consistency between CTL V_{INT} and HIST+RCP8.5 V_{TOT} results. Of particular interest here is whether the interdecadal TMT variability in the signal-removed forced runs lies within the range of CTL interdecadal variability.

Consider first the sensitivity of the HIST+RCP8.5 spectra to the signal removal method. The LIN, QUAD, and CUB

approaches yield very similar spectral density results, except on the longest time scales in certain models (e.g., in CESM-BGC, GFDL CM3, and the GISS models). The most pronounced spectral differences are between the unscaled MMA and the three other methods. These differences manifest most clearly on time scales longer than several years (Fig. 2) and are not of uniform direction. The MMA-removed variance on 5–20-yr time scales can be either consistently higher (INM-CM4) or consistently lower (ACCESS1.0) than in the LIN, QUAD, and CUB cases. The latter condition is more common.

Next, we compare the CTL V_{INT} and extended HIST V_{TOT} spectra. Both CTL and HIST+RCP8.5 simulations were available for 35 different CMIP5 models.² For MMA removal, there are 7 of 35 cases in which the HIST+RCP8.5 spectral density on time scales of 5 to 20 years is consistently *above* the upper end of the 95% uncertainty range on the CTL spectral density.³ Not all of these seven models have high or low ECS values. This illustrates that the efficacy of the unscaled MMA as a signal removal method is not solely driven by differences between the average ECS of CMIP5 models and an individual model's ECS; the efficacy must also depend on MMA-versus-single model differences in the magnitude and time history of anthropogenic aerosol forcing (Zelinka et al. 2014; Santer et al. 2019).

If the LIN, QUAD, and CUB signal removal methods are considered jointly, there are 11 cases where the spectral density on 5–20-yr time scales consistently exceeds the CTL 95% uncertainty range.⁴ There are no cases in which any of the four signal removal methods yield low-frequency spectral density consistently *below* the control uncertainty range. We conclude from this result that our signal removal methods are not significantly reducing “pure” interdecadal variability by mistakenly attributing a significant portion of this variability to the forced signal.

For most CMIP5 models, the interannual variability in the HIST+RCP8.5 simulation is generally within the 95% confidence interval of their CTL V_{INT} spectrum (Fig. 2). There are exceptions in both directions. For example, IPSL-CM5A-LR, MIROC-ESM-CHEM, and NorESM1-ME have individual spectral peaks in the 6-month to 2-yr band that are above the upper bound of the 95% confidence interval on their CTL V_{INT} results, while CESM-BGC and CMCC-CM have decreases in PSD that lie below the lower bound of the 95% CTL confidence interval. These results suggest that there are grounds for further inquiry into forced changes in the variability of TMT on semiannual, annual, and interannual time scales (Santer et al. 2018).

² Synthetic satellite temperatures could not be calculated for 2 of the 37 CMIP5 models for which extended HIST simulations were available [the EC-EARTH and GISS-E2-R (p3) models].

³ The seven models in question are CMCC-CMS, FGOALS-g2, FIO-ESM, GFDL CM3, INM-CM4, MIROC-ESM-CHEM, and MRI-CGCM3.

⁴ For BCC-CSM1.1, CanESM2, CCSM4, CESM1-BGC, GFDL-ESM2M, GISS-E2-H (p1), GISS-E2-H (p3), GISS-E2-R (p1), GISS-E2-R (p2), NorESM1-M, and NorESM1-ME.

FIG. 2. Power spectral density (PSD) for near-global averages of monthly mean tropospheric temperature in CMIP5 preindustrial control runs and HIST+RCP8.5 simulations. For each model control run, the average PSD over all overlapping 40-yr chunks (bold cyan line) is plotted against the period (in years) on a log-log scale. The shaded area delimits the chi-squared 95% confidence intervals for the PSD estimate. The HIST+RCP8.5 spectra were calculated after application of the four different signal removal methods investigated here (LIN, QUAD, CUB, and unscaled MMA) to each model's 480-month HIST+RCP8.5 time series. The analysis period for the HIST+RCP8.5 runs is from January 1979 to December 2018. The background shading in each panel is based on a simple proxy for equilibrium climate sensitivity (Santer et al. 2018). There are 35 models with both CTL and HIST+RCP8.5 simulations.

One noticeable aspect of the CTL spectra is the wide range of their individual shapes (Fig. 2 and Fig. S1). For example, the HadGEM2-CC and MIROC-ESM models lack pronounced spectral peaks and show gradual increases in power with increasing time scale. In contrast, models such as CMCC-CESM, FIO-ESM, and IPSL-CM5B-LR exhibit greater complexity of shape, with spectral peaks on the 3–5-yr time scales characteristic of internal variability associated with ENSO.

The presence or absence of discrete spectral peaks tends to be correlated with the goodness of fit of a power law (Vyushin et al. 2012), represented here by the coefficient of determination R^2 . The value of R^2 varies from 0.2 in a model with a large ENSO peak (CMCC-CESM) to 0.91 in a model without discrete spectral peaks (HadGEM2-CC) (Fig. 3a). In models with smaller R^2 values, the spectral density at the longest time scale that can be usefully resolved here (20 years) is generally lower than the spectral density on 3–5-yr ENSO time scales. There are also pronounced intermodel differences in β , the slope of the power-law fit (Fig. 3b). Values of β vary by up to a factor of 1.7. Estimates of β are meaningful only for climate

models with a high goodness of fit to a power law (i.e., for $R^2 > 0.7$).

Based on the R^2 results in Figs. 2 and 3, we would conclude that less than one-third of the CMIP5 model control runs (11 out of 36) show a consistent scaling relationship between the amplitude of interannual- and decadal-time scale TMT variability (Santer et al. 2013a). If this finding were applicable to the real world, it would imply that well-observed interannual TMT variability may not provide a strong constraint on the size of the more uncertain observed interdecadal TMT variability.

6. Comparison of simulated and observed bandpower

Our focus in this section is on comparing V_{TOT} spectra in observations and in HIST+RCP8.5 simulations. We perform these comparisons by averaging spectral density over three different frequency ranges: all frequencies resolved here, high frequencies, and low frequencies. We refer to these frequency ranges subsequently as ALL, HIGH, and LOW. They correspond, respectively, to periods spanning the ranges from 2 months to 20 years, 1 to 5 years, and 5 to 20 years.

FIG. 3. (a) Goodness-of-fit R^2 and (b) slope β of the power-law representation of the spectral density of TMT in CMIP5 control runs. Values of R^2 and β were estimated from the average spectral densities of preindustrial control runs performed with 36 different CMIP5 models. For details of averaging, refer to Fig. 2. The numerical values of R^2 and β are color-coded according to their size; white values indicate $R^2 > 0.7$.

The HIGH frequency range captures variability on inter-annual time scales and within the 2–5-yr time scales characteristic of ENSO-induced surface and tropospheric temperature variability (AchutaRao and Sperber 2006; Bindoff et al. 2013). The LOW frequency range samples interdecadal variability that is of interest in anthropogenic signal detection (Santer et al. 2018). The statistical significance results presented in section 8 are relatively insensitive to different plausible choices of the boundary between the HIGH and LOW frequency ranges.

A brief explanation is required for our choice of 20 years as the longest period that can be usefully resolved from 40-yr TMT records. One simple practical rule in spectral analysis is that L years of data can only provide useful variability information on time scales smaller than $L/3$ (Bloomfield and Nychka 1992). Application of this rule to satellite TMT data would yield a maximum period of ~ 13 years. In our case, however, the availability of multicentury control runs and multiple realizations of historical climate change (along with our use of results from three dozen climate models) allows us to obtain more reliable estimates of the amplitude of simulated tropospheric temperature variability on interdecadal time scales. For the satellite observations, it is the statistical modeling of “observed” natural variability that provides quantitative estimates of uncertainty in TMT spectra on time scales >13 and ≤ 20 years (see section 7).

We focus here on discussion of absolute band power, which is simply the average power over the frequency range of interest, calculated via rectangular integration.⁵ Model-versus-observed differences in band power are not typically employed for formal significance testing. A more commonly used alternative is the total variational distance (TVD), which is defined as the sum of absolute differences between modeled and observed normalized spectral densities over all the frequencies in the spectrum (Alvarez-Esteban et al. 2016; Euan et al. 2015). By definition, therefore, TVD does not provide directional information, and normalization by the temporal standard deviation of each time series reduces any overall differences in variability amplitude. Both normalization and use of absolute differences are disadvantages here, since we are interested in evaluating whether CMIP5 and CMIP6 models systematically underestimate the amplitude of “observed” interdecadal natural variability.

Band power values for the three satellite datasets and the 37 CMIP5 HIST+RCP8.5 simulations are shown in Fig. 4. Consider first the LOW frequency band. With the LIN, QUAD, and CUB signal removal methods, there is no evidence that CMIP5 models systematically underestimate observed V_{TOT} . The converse is the case: the mean and median of the model band power in the LOW frequency range exceeds that of all three satellite datasets (Figs. 4b–d). For MMA signal removal, RSS and STAR have band power on 5- to 20-yr time scales that is close to the mean of the CMIP5 results (Fig. 4a). As noted in the discussion of Fig. 1, MMA removal yields low-frequency band power for UAH that is noticeably larger than in the two other satellite datasets, and is also larger than in the mean and median of the CMIP5 results.

Comparisons of modeled and observed band power for the HIGH frequency range also reveal pronounced differences between MMA signal removal and the LIN, QUAD, and CUB cases. The latter three methods yield close agreement between band power in the observations and in the mean and median of the CMIP5 HIST+RCP8.5 runs. For MMA removal, however, observed band power in the HIGH range is larger than in the average and median of the model results (Fig. 4a). This may be partly due to leakage of the MMA-versus-observed trend differences into the upper end of the 1–5-yr band (Fig. 1h).

The ALL results reflect aspects of both the HIGH and LOW comparisons. When band power is averaged over all resolved frequencies, the MMA case remains the only case where the mean band power for the 37 CMIP5 models is smaller than observed. Note also that there are small differences between the ALL results for LIN, QUAD, and CUB signal removal, with the latter method producing average band power that is closest in simulations and observations.

⁵ The rectangle method (also known as the midpoint rule) is a numerical integration technique that estimates the definite integral of a function by using rectangles to approximate the area under the curve.

FIG. 4. Average band power over the ALL, HIGH, and LOW frequency ranges in the CMIP5 HIST+RCP8.5 simulations and in RSS, STAR, and UAH satellite datasets. Band power was calculated after applying the (a) MMA, (b) LIN, (c) QUAD, and (d) CUB signal removal methods, respectively. The solid and open black squares denote the multimodel average and median, respectively, of the CMIP5 band power results.

7. Statistical models of observed variability

To test the significance of the model-versus-observed band power differences in Fig. 4, we require one or more statistical models of the “observed” natural variability of TMT. Early studies represented internal variability with simple low-order autoregressive (AR) statistical models, such as an AR(1) or AR(2) process (Hasselmann 1976; Imbers et al. 2014). Subsequent work used more complex statistical models to characterize the long-term memory of climate time series (Bloomfield and Nychka 1992; Franzke 2012a,b; Imbers et al. 2014). Examples include autoregressive integrated moving average (ARIMA) models and fractional versions of an ARIMA model. In the latter class of FARIMA(p, d, q) models, the order of differencing d is allowed to take fractional values and produces long-term memory for $0 < d < 1/2$ (Taqq et al. 1995). The indices p and q are the orders of the autoregressive and moving-average components of the model, respectively. The simplest FARIMA($0, d, 0$) model has been used to describe a slowly decaying temporal autocorrelation structure in studies of long-term climate variability and trend significance (Bloomfield and Nychka 1992; Koscielny-Bunde et al.

1998; Franzke 2012a,b; Imbers et al. 2014; Franzke et al. 2015; Bowers and Tung 2018).

We consider eight different statistical models of the natural variability of TMT. The first three are AR(1), AR(2), and AR(4) models. The fourth is an ARMA(p, q) model with $p = 1$ and $q = 1$ (i.e., with order 1 for both the autoregressive and moving-average components). The remaining four FARIMA models have various combinations of p , d , and q . The frequently used FARIMA($0, d, 0$) model (Franzke et al. 2015) has no autoregressive or moving average terms, while FARIMA($1, d, 1$) has order 1 for both of these terms. FARIMA($0, d, 2$) and FARIMA($2, d, 0$) have (respectively) 0 and 2 for the AR order and 2 and 0 for the moving-average order. Together, these eight models capture a wide range of different combinations of short- and long-term memory behavior.

Despite their widespread use in the literature, we found that AR(1) and FARIMA($0, d, 0$) were not the best-fit models to the observed TMT residuals for the four signal removal strategies and the entire 2-month to 20-yr frequency range considered here. Based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC) (Schwarz 1978), the

best-fit statistical models in the AR, ARMA, and FARIMA model classes were judged to be AR(4), ARMA(1, 1), and FARIMA(2, d , 0) (respectively). The full list of statistical models evaluated in each class is given in the supplemental material.

Figure 5 displays spectral density for the FARIMA(2, d , 0) best-fit statistical model, the commonly used AR(1) and FARIMA(0, d , 0) models, the observations, and the HIST+RCP8.5 and CTL runs. For each statistical model, the spectrum is the average of the individual spectra estimated from 10 000 time series of length 40 years. Each of these 10 000 time series is a random sample generated by the stochastic process whose parameters p , d , and q were estimated from the observed signal-removed TMT data. The stochastically generated spectra are good approximations to the theoretical spectra (see supplemental material).

In terms of the overall shape of the spectrum, the observed TMT results in Fig. 5 are in closer agreement with the best-fit statistical models than with the AR(1) and FARIMA(0, d , 0) models. The AR(1) model consistently overestimates observed PSD for periods between 6 months and 3 years. In contrast, the FARIMA(0, d , 0) model systematically underestimates observed PSD for periods of 1–7 years and overestimates observed PSD for periods longer than 7 years (except in the UAH/MMA signal removal case). Clearly, the choice of statistical model is important in terms of the fidelity with which the “observed” natural variability spectrum can be represented.

It is also of interest to briefly compare the ensemble-mean spectra for the HIST+RCP8.5 and CTL runs (Fig. 5).⁶ For periods up to roughly 3 years, the two types of simulation yield similar average PSD. On periods longer than roughly 3 years, average PSD is consistently higher in the HIST+RCP8.5 simulations than in the CTL runs (i.e., V_{TOT} is larger than V_{INT}). For LIN, QUAD, and CUB, higher V_{TOT} is primarily due to the fact that the HIST+RCP8.5 runs include both natural external forcing and any nonlinear anthropogenic signal components that are not well captured by the removal of low-order polynomial fits. In the MMA signal removal case, the higher average 3–20-yr spectral density in the HIST+RCP8.5 runs is due to differences between the model average forced response and the forced response in individual models.

8. Significance of model–observed spectral differences

Observed signal-removed TMT data provide one specific realization of natural internal variability, yielding a single estimate of band power. Different sequences of internal variability would produce different estimates of observed band power. We account for this uncertainty by applying the statistical modeling framework described in section 7. This allows us to compare the bandpower of individual CMIP5 models against null distributions of “observed” band power rather than against a single observational value (see Fig. 6). Null distributions are based on the statistical models that we fit to the satellite datasets. For each statistical model, signal removal

method, and satellite TMT dataset, a null distribution is constructed by randomly generating $N_{\text{TOT}} = 10\,000$ surrogate time series, each of length 40 years. The statistical models cover a wide range of stochastic representations of the true (but uncertain) observed internal variability. This procedure yields more reliable inferences regarding the statistical significance of model-versus-observed band power differences.

For each CMIP5 HIST+RCP8.5 simulation, we calculate the probability $p_i = N_i / N_{\text{TOT}}$, where N_i is the number of null distribution values less than the band power value for the i th CMIP5 model. Averaging the individual p_i values for the 37 CMIP5 models yields $\bar{p}$, which is shown in Fig. 7. There is one $\bar{p}$ value for each of three observational datasets, four signal removal methods, eight statistical models, and three frequency ranges. Values of $\bar{p}$ between 0 and 0.5 indicate that on average, CMIP5 models underestimate the observationally derived band power. Values of $\bar{p} \leq 0.05$ point toward a significant underestimate of observed natural variability at a stipulated 5% level. Conversely, $\bar{p} > 0.5$ and ≤ 1 denotes a model-average overestimate of observationally derived band power.

The final row of Fig. 7 displays $\bar{p}$, the average $\bar{p}$ value across observational datasets and statistical models. There is one $\bar{p}$ value for each signal removal strategy and frequency range. For example, $\bar{p}_{\text{LOW}} = 0.38, 0.79, 0.77$, and 0.73 for MMA, LIN, QUAD, and CUB signal removal strategies, respectively. The overall average of $\bar{p}_{\text{LOW}}$ across signal removal methods is 0.61. This indicates that most CMIP5 models overestimate the band power of the signal-removed observational TMT data for periods between 5 and 20 years.

The markedly smaller $\bar{p}_{\text{LOW}}$ value for MMA removal arises because the observed TMT residuals generated with this signal removal method have pronounced low-frequency structure (see Figs. 1e–g). As noted in section 4, at least four factors contribute to this low-frequency structure. These factors include differences between the true (but uncertain) real-world ECS and the average ECS of CMIP5 models, systematic forcing errors in the CMIP5 HIST+RCP8.5 simulations, different phasing of decadal internal variability in models and observations, and remaining errors in the satellite data. The impact of the first of these four factors is reduced by applying the LIN, QUAD, and CUB signal removal methods.

As a result of the more pronounced low-frequency variability in the observed MMA-removed residuals, statistical models developed from these residuals tend to produce larger long-term memory, higher median values for null distributions of low-frequency band power, and smaller values of $\bar{p}_{\text{LOW}}$ relative to the other signal removal cases (see Fig. 7, and compare left and right columns of Fig. 6). Values of $\bar{p}_{\text{ALL}}$ and $\bar{p}_{\text{HIGH}}$ are also systematically lower for MMA removal than for subtraction of LIN, QUAD, and CUB signals. For $\bar{p}_{\text{ALL}}$, this occurs because the ALL results include the low-frequency portion of the spectrum. In the case of $\bar{p}_{\text{HIGH}}$, the large low-frequency residuals obtained with MMA removal “leak” into the upper end of the 1–5-yr band of the spectrum.

Another noteworthy feature of Fig. 7 is that MMA signal removal produces consistently smaller $\bar{p}$ values for UAH than for RSS or STAR. This occurs because the difference between overall tropospheric warming trends in the MMA and in

⁶ Figure 2 provided this comparison for individual models only.

FIG. 5. Power spectral density of simulated and observed TMT variability. PSD was estimated after removing an externally forced signal from observations and from HIST+RCP8.5 simulations. Four different signal removal methods were applied: (a)–(c) MMA, (d)–(f) LIN, (g)–(i) QUAD, and (j)–(l) CUB. PSD is shown for the (left) RSS, (center) STAR, and (right) UAH satellite datasets, the HIST+RCP8.5 simulations performed with 37 different CMIP5 models, 36 different CMIP5 preindustrial control runs, the best-fit FARIMA(2, d , 0) statistical model, and the commonly used AR(1) and FARIMA(0, d , 0) statistical models. The best-fit AR(4) and ARMA(1, 1) models (not shown) yield spectra that are very similar to the FARIMA(2, d , 0) case. The HIST+RCP8.5 and CTL spectra do not depend on the choice of observational dataset.

satellite data is largest in UAH, enhancing the long-term memory calculated with the UAH residuals (cf. observational spectra in Fig. 1h). Even in the UAH MMA removal case, however, the average $\bar{p}_{\text{LOW}}$ value across the eight statistical models is 0.26, and there is no statistically significant difference in low-frequency band power between UAH and the CMIP5 HIST+RCP8.5 simulations.

Recall that AR(4), ARMA(1, 1), and FARIMA(2, d , 0) were determined to be the best-fit statistical models in the AR, ARMA, and FARIMA classes. For a given signal removal method, satellite dataset, and frequency band, these three models yield similar $\bar{p}$ values (Fig. 7). The $\bar{p}$ values for the commonly used AR(1) and FARIMA(0, d , 0) statistical models are noticeably different from the best-fit model results. For UAH satellite data and MMA signal removal, for example, the null distribution of low-frequency band power generated

with the best-fit FARIMA(2, d , 0) model has higher median band power and a larger interquartile range relative to the two FARIMA models with no autoregressive component [FARIMA(0, d , 0) and FARIMA(0, d , 2); see Fig. 6e]. Likewise, the “best-fit” AR(4) model has larger median low-frequency band power than the AR(1) model. The key point here is that if we consider only the best-fit statistical models (rather than all eight statistical models), there is still no evidence that CMIP5 climate models significantly underestimate the size of observed natural variability on periods ranging from 5 to 20 years.

As noted above and as shown in Figs. 4 and 5, variability in the 5–20-yr band has reduced amplitude in the CMIP5 CTL runs relative to the signal-removed HIST+RCP8.5 simulations. Since the observationally derived null distributions of band power are identical in Figs. 7 and 8 (for the HIST+RCP8.5

FIG. 6. Comparisons of the simulated and observed band power of TMT integrated over (a),(b) the entire spectrum and over (c),(d) high-frequency (1–5 years) and (e),(f) low-frequency bands (5–20 years) for (left) MMA and (right) LIN signal removal. Symbols denote results from UAH satellite data and from HIST+RCP8.5 simulations performed with 37 different CMIP5 models. The CMIP5 CTL multimodel average and median band power are also shown (and do not involve signal removal). To enable more robust statistical inferences regarding band power differences between CMIP5 models and satellite data, eight different statistical models were fit to the UAH signal-removed TMT residuals. Each statistical model was used to generate 10 000 realizations of surrogate “observed” natural variability; spectra and band power were computed from these surrogates. Each box plot corresponds to the null distribution from a statistical model. The box edges span the 25th and 75th percentiles; the median is marked in each box. The whiskers represent $\sim 99.3\%$ of the distribution. MMA and MMM denote the multimodel average and median, respectively, of the CMIP5 HIST+RCP8.5 band power results. The vertical dashed line is centered on the band power in the UAH signal-removed data.

and CTL simulations, respectively), it is the smaller low-frequency band power in the CTL runs that yields systematically smaller $\bar{p}_{\text{LOW}}$ values relative to the HIST+RCP8.5 $\bar{p}_{\text{LOW}}$ results. While most $\bar{p}_{\text{LOW}}$ values in Fig. 7 were greater than 0.5, all $\bar{p}_{\text{LOW}}$ values in Fig. 8 are less than 0.5. The same systematic difference between CTL and HIST+RCP8.5 results holds for $\bar{p}_{\text{ALL}}$ and $\bar{p}_{\text{HIGH}}$.

The results in Figs. 7 and 8 clearly illustrate that the choice of model simulation can have noticeable impact on the estimated consistency between simulated and observed natural variability. Since we know a priori that the observed signal-removed TMT data do not constitute pure internally generated variability, the observed V_{TOT} versus HIST+RCP8.5 V_{TOT} comparison in Fig. 7 is more appropriate and more meaningful than the observed V_{TOT} versus CTL V_{INT} comparison in Fig. 8. Even in the latter case, however, values of $\bar{p}_{\text{LOW}}$ in the bottom row do not indicate a statistically significant underestimate of observed low-frequency TMT variability (at a stipulated 5% level) in CMIP5 control runs.

9. Comparison of CMIP5 and CMIP6 simulations

Simulations performed under the latest phase of the Coupled Model Intercomparison Project (CMIP6) are currently being assessed for the upcoming 2021 Sixth Assessment Report of the IPCC (Eyring et al. 2016). Relative to CMIP5, there are a larger number of CMIP6 Earth system models, some of which incorporate aspects of the physical climate system that were not well studied in CMIP5 (such as the interactions between climate change and major ice sheets). Many CMIP6 models have higher horizontal and vertical resolution than their predecessors, improvements to their physics, parameterizations, and external forcings, larger ensemble sizes of historical simulations, and a wider range of numerical experiments seeking to better quantify the responses to individual forcings (Gillett et al. 2016; Eyring et al. 2019).

The archive of CMIP6 simulation output was still being populated at the time our study was performed. We relied on synthetic satellite temperatures from a total of 98 HIST realizations calculated with 21 different models (see Table S3). As

FIG. 7. Probability that CMIP5 HIST+RCP8.5 simulations have larger band power than in observationally derived null distributions of band power. Results are for three different frequency ranges (ALL, HIGH, and LOW). Four different functional forms of externally forced TMT signal (MMA, LIN, QUAD, and CUB) were removed from the 37 different HIST+RCP8.5 simulations and from the (top) RSS, (middle) STAR, and (right) UAH satellite data. Eight different statistical models were then fit to the signal-removed satellite TMT residuals. The null distributions generated with these statistical models were used to calculate a probability p_i for each CMIP5 HIST+RCP8.5 run. The average probability across CMIP5 models, $\bar{p}$, is displayed here. Blue (red) shading indicates that $\bar{p}$ is greater than (less than) 0.5. The final row provides $\bar{\bar{p}}$, the average of $\bar{p}$ across satellite datasets and statistical models. The final column is the average $\bar{p}$ value (across signal removal methods and frequency ranges) for a given statistical model and satellite temperature dataset.

noted in section 2, CMIP6 HIST runs were spliced together with results from the SSP5 scenario, which commenced in January 2015. We focus here on comparing V_{TOT} spectra in CMIP6 HIST+SSP5 runs and in CMIP5 HIST+RCP8.5 simulations. We seek to determine whether there are fundamental differences in TMT variability between these two generations of CMIP models.

A comparison of the CMIP5 V_{TOT} spectra for the CMIP5 and CMIP6 extended HIST runs is shown in Fig. 9. Despite differences in the average ECS of CMIP5 and CMIP6 models (Zelinka et al. 2020), the shape of the V_{TOT} spectrum is broadly similar in CMIP5 and CMIP6. This holds for each of the four signal removal strategies. There are also some noticeable differences between the CMIP5 and CMIP6 V_{TOT} spectra. On average, spectral density in CMIP6 models is consistently higher in the 2–5-yr ENSO band, irrespective of signal removal strategy. CMIP6 variability is also larger on time scales of 10–20 years, but only for the LIN, QUAD, and CUB signal removal methods. This suggests that the low-frequency departures of individual CMIP5 and CMIP6 models from their respective MMAs are relatively similar. The fact that the same is not true for low-order polynomial fits may point toward a temporally more complex component of forced TMT change in CMIP6.

Another noticeable difference in the CMIP5 and CMIP6 V_{TOT} spectra occurs for the annual cycle of TMT, which is

larger in CMIP6. This difference in annual cycle amplitude is damped by subtraction of the respective CMIP5 and CMIP6 MMAs, which is why it is only visible in the LIN, QUAD, and CUB signal removal methods. Forced changes in the amplitude of the annual cycle of TMT have been identified elsewhere (Santer et al. 2018). Further work is required to determine why they are larger in the CMIP6 models analyzed here.

As expected based on the results in Fig. 9, the statistical significance of model-versus-observed variability differences is broadly similar in CMIP5 and CMIP6. Recall that for the CMIP5 HIST+RCP8.5 runs, averaged across all three observational datasets and all 8 statistical models, the $\bar{p}_{\text{LOW}}$ values were 0.38, 0.79, 0.77, and 0.73 for the MMA, LIN, QUAD, and LOW signal removal methods (respectively; see bottom row of Fig. 7). The corresponding $\bar{p}_{\text{LOW}}$ values for the CMIP6 HIST+SSP5 runs are 0.37, 0.90, 0.88, and 0.86 (see Fig. S2). The higher $\bar{p}_{\text{LOW}}$ values for the CMIP6 LIN, QUAD, and CUB cases are consistent with the higher interdecadal spectral density in CMIP6 (see Figs. 9b–d). The primary result here is that on average, CMIP5 and CMIP6 models both overestimate observed interdecadal variability. There are, however, relatively few cases for which this overestimate is statistically significant at the 5% level or better.

Note that $\bar{p}_{\text{HIGH}}$ is generally closer to 1 for CMIP6 than for CMIP5 (cf. Fig. S2 and Fig. 7). This suggests that as for interdecadal variability, the overestimate of observed TMT

FIG. 8. As in Fig. 7, but for the probability that CMIP5 CTL simulations have larger band power than in observationally derived null distributions of band power.

variability on time scales of 1 to 5 years is larger in CMIP6 than CMIP5. Higher CMIP6 variance in both the annual cycle and in the ENSO band contribute to this result (Fig. 9).

10. Conclusions

We have provided a rigorous framework for assessing the statistical significance of differences between the simulated and observed natural variability of tropospheric temperature. Our focus has been on the sensitivity to different analyst choices. While elements of such sensitivity studies are distributed throughout the literature (Bloomfield and Nychka 1992; Imbers et al. 2014), it is difficult to compare results from assessments that have relied on different individual statistical models, types of climate model simulation (CTL or historical), signal removal strategies, geographical domains, and climate variables. It is of scientific value to explore such sensitivities in a systematic way. This was our goal here.

Our model–data variability comparisons were conducted in the frequency domain using the band power averaged over three different frequency ranges (ALL, HIGH, and LOW). The LOW range (5–20 years) is of critical importance in anthropogenic signal detection studies. Variability in this range constitutes background noise against which analysts attempt to detect gradually evolving warming. If current climate models significantly underestimated observed variability in the LOW range, it would call into question claims that an anthropogenic warming signal had been identified in tropospheric temperature (Santer et al. 2013a, 2018).

In the most relevant comparisons, which involve “signal-removed” satellite data and HIST+RCP8.5 simulations, CMIP5 models generally *overestimate* observed low-frequency variability (see Fig. 7). The only case in which LOW band power is noticeably (but not significantly) smaller in CMIP5

models involves UAH tropospheric temperature data and MMA signal removal. The reason for this result is that overall tropospheric warming in UAH is less than half of the warming in the MMA (0.14° vs $0.29^{\circ}\text{C decade}^{-1}$, respectively). After MMA removal from UAH data, this large trend difference is aliased in the low-frequency portion of the UAH spectrum. This explains why statistical models fit to the UAH residuals have larger long-term memory, higher median values for null distributions of LOW band power (Figs. 6e,f), and consistently smaller values of $\bar{p}_{\text{LOW}}$ for MMA removal (relative to RSS and STAR satellite datasets with larger tropospheric warming; see Fig. 7).

Extended HIST simulations performed with newer-generation CMIP6 models yield qualitatively similar statistical significance results (Fig. S2). Relative to CMIP5, the primary difference is that CMIP6 models have an even larger overestimate of “observed” interdecadal TMT variability for the LIN, QUAD, and CUB signal removal cases. As for CMIP5, the only case where CMIP6 models underestimate “observed” interdecadal variability involves UAH satellite data and MMA signal removal.

There are valid reasons for scientific concern about the credibility of the muted tropospheric warming in UAH. Over tropical oceans, UAH tropospheric temperature trends are smaller than surface trends (Po-Chedley et al. 2015). Such behavior is inconsistent with basic moist thermodynamics (Stone and Carlson 1979; Santer et al. 2005). Additionally, the UAH scaling ratios between trends in tropospheric temperature and trends in independently monitored atmospheric moisture are not in accord with physical expectations (Mears and Wentz 2016). A discontinuity in the amplitude of the UAH annual cycle of TMT provides further evidence of residual nonclimatic inhomogeneity in the UAH data (Santer et al.

FIG. 9. Comparison of the power spectral density of TMT in CMIP5 and CMIP6 models. Results are a function of signal removal strategy and are the average spectra for the HIST+RCP8.5 simulations performed with 37 different CMIP5 models and for the HIST+SSP5 simulations performed with 21 different CMIP6 models. Four different strategies were used to remove an estimated externally forced TMT signal. (a) In the first strategy, the unscaled CMIP5 MMA and the unscaled CMIP6 MMA are removed from the HIST+RCP8.5 and HIST+SSP5 simulations, respectively, performed with individual models. The remaining three signal removal strategies involve subtraction of the individual (b) linear, (c) quadratic, and (d) cubic fits to each HIST+RCP8.5 and HIST+SSP5 simulation. All spectra were calculated for TMT data spatially averaged over 82.5°N–82.5°S. The analysis period is from January 1979 to December 2018.

2018). As we show in Fig. 1h, the UAH-estimated interdecadal variability for MMA signal removal is unusually large even relative to the LOW band power that arises from large known climate sensitivity errors (see section 3).

Because of uncertainties in ECS, anthropogenic aerosol forcing, and the observations themselves, reliable estimation of “pure” internal variability V_{INT} from observations remains challenging, particularly on interdecadal time scales. We focused here on comparing observed and modeled estimates of “total” natural variability V_{TOT} . For low-frequency band power, our results indicate that signal-removed observations are in better accord with V_{TOT} results from extended HIST simulations than with V_{INT} results from unforced CTL runs. This was the case for both CMIP5 and CMIP6 models.

In terms of statistical model selection, we found that the traditionally used AR(1) and FARIMA(0, d , 0) models do not provide the best representations of the observed TMT spectra, irrespective of which satellite dataset and signal removal method we used (see Fig. 5). Higher-order AR models and FARIMA models with a nonzero autoregressive component

are more successful in capturing the complex observed distribution of spectral density as a function of frequency.

A key issue in our study and in trend significance assessments is whether observed low-frequency natural variability should be estimated after signal removal or from raw data.⁷ One scientific perspective is that the latter strategy is preferable (Franzke 2010, 2012a; Bowers and Tung 2018). This is not an unreasonable strategy when dealing with climate data at individual locations, where the true signal is poorly known and S/N ratios are lower than for global-scale spatial averages.

An alternative perspective is that anthropogenic signals have been statistically identified in many different climate variables over a wide range of spatial scales (Hegerl et al. 2007; Bindoff et al. 2013). Under this school of thought, the physical reality of ubiquitous human-caused climate signals justifies removal of an anthropogenic signal prior to estimation of

⁷ A different strategy involves simultaneous estimation of signal and noise from observational time series (Gil-Alana 2005).

natural variability. We follow this strategy here and explicitly consider the impact of signal uncertainties on model-versus-data variability comparisons. As we show in Fig. 1 (cf. Figs. 1d, 1h, and 1i), the decision to operate on raw or signal-removed TMT data has substantial impact on the estimated observed spectrum of natural variability. Our future work will explore the implications of these two choices for trend significance assessments.

We also intend to expand our suite of signal removal methods. Although the four methods used here have been widely applied in many previous climate studies (Bloomfield and Nychka 1992; Knutson et al. 2013; Santer et al. 2011; Imbers et al. 2014), all have well-documented deficiencies (Steinman et al. 2015; Frankcombe et al. 2015, 2018; Kravtsov 2017; Kravtsov et al. 2018; Cheung et al. 2017a,b; Kajtar et al. 2019). Complex low-frequency signals—such as those associated with secular changes in anthropogenic aerosol forcing—are unlikely to be reliably captured by linear, quadratic, and cubic fits. While subtraction of the unscaled MMA can capture such low-frequency signals, it preserves any overall warming rate differences between satellite TMT data and the average of a multimodel ensemble of extended HIST simulations. These differences are then aliased into (and spuriously inflate) statistical model estimates of natural variability derived from MMA-removed observations.

Alternative approaches exist involving subtraction of scaled versions of the MMA (Steinman et al. 2015; Frankcombe et al. 2015, 2018). MMA scaling is likely to improve estimates of internal variability inferred from individual model extended HIST simulations (see section 3).⁸ Whether removal of the scaled MMA from satellite TMT data necessarily improves estimates of the true (but uncertain) observed interdecadal internal variability is still unclear.

Energy balance models (EBMs) provide another means for estimating “noise-free” anthropogenic signals from observations in the presence of uncertainties in ECS and anthropogenic aerosol forcing (Wigley and Raper 1990; Rypdal and Rypdal 2014; Fredriksen and Rypdal 2017). Since climate sensitivity is a free parameter in an EBM, such models could be employed to systematically explore the impact of physically plausible ECS and aerosol forcing uncertainties on the observed signal-removed temperature residuals, and on the null distributions and statistical inferences derived from these residuals.

In comparing the efficacy of signal removal approaches involving EBMs, MMA scaling, and the simpler approaches used here, it will be helpful to make greater use of large initial condition ensembles (LEs) (Deser et al. 2014, 2020; Fyfe et al. 2017; Swart et al. 2018). In LEs, an individual model’s true forced signal and internal variability can be reliably quantified (at least at global and continental scales). This allows analysts to make rigorous quantitative comparisons of the performance

of different signal removal strategies. Multiple LEs should be used for this purpose (Santer et al. 2019; Deser et al. 2020).

The results obtained here and earlier (Santer et al. 2011, 2013a, 2018) suggest that on average, the last three generations of CMIP models overestimated the “total” interdecadal variability of global-mean tropospheric temperature inferred from satellite data. This finding is in accord with some but not all of the independent investigations relying on surface temperature (Imbers et al. 2014; Vyushin et al. 2012; Cheung et al. 2017a; Kim et al. 2018; Zhu et al. 2019; Lee et al. 2019). Further work is necessary to better understand—and hopefully resolve—differences in observed estimates of natural internal variability. Ideally, such work will involve rigorous intercomparison of different signal removal methods using the type of statistical framework applied here.

Acknowledgments. We acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy’s Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Work at LLNL was performed under the auspices of the U.S. Department of Energy under Contract DE-AC52-07 NA27344 through the Regional and Global Model Analysis Program. All primary TMT temperature satellite datasets and CMIP5 model output used here are publicly available. We thank Jeff Painter and Stephen Po-Chedley for calculating synthetic satellite temperatures from CMIP5 and CMIP6 simulation output.

APPENDIX A

Calculation of Synthetic Satellite Temperatures from Model Data

We used a local weighting function method developed at RSS to calculate synthetic satellite temperatures from model output (Santer et al. 2013a). At each model grid point, simulated temperature profiles were convolved with local weighting functions. The weights depend on the grid point surface pressure, the surface type (land, ocean, or sea ice), and the selected layer-average temperature. We calculated both TMT and the temperature of the lower stratosphere (TLS). The latter was used for correcting TMT for the contribution it receives from lower stratospheric cooling (Fu et al. 2004). The local weighting function method provides more accurate estimates of synthetic satellite temperatures than use of a global-mean weighting function, particularly over high-elevation regions.

APPENDIX B

Method for Correcting TMT Values

Estimates of TMT trends obtained from microwave sounders are influenced by the cooling of the lower stratosphere (Fu et al. 2004; Fu and Johanson 2004). For removing this stratospheric cooling component of TMT we used the same

⁸ If an adequate number of extended HIST realizations are available, the use of an individual model’s ensemble-average signal estimate is preferable to the use of the scaled MMA.

regression-based method applied in Santer et al. (2017b) and originally developed in Fu et al. (2004). These datasets are referred to as “corrected TMT.” As in Santer et al. (2017b), correction was performed locally at each model and observational grid point. Corrected grid point data were then spatially averaged over 82.5°N–82.5°S to obtain near-global averages (see above). Model and observational temperature data were processed in the same way, thus ensuring that model-versus-observed differences in the spectra of corrected TMT are not due to differences in the applied regression method. The correction involves latitudinally invariant regression coefficients.

APPENDIX C

Calculation of Tropospheric Temperature Anomalies

Anomalies for the CMIP5 CTL runs were defined as follows. (i) For each climate model for which a control integration is available, extract all the maximally overlapping time series of length $L = 480$ months (i.e., 40 years). If N_m is the length in months of the CMIP5 CTL run used (see Table S1), the number of maximally overlapping chunks N of length L is simply $N = N_m - L + 1$. (ii) For each 480-month chunk, monthly anomalies were defined relative to climatological monthly means for that 480-month chunk.

For the satellite data, the CMIP5 HIST+RCP8.5 runs, and the CMIP6 HIST+SSP5 simulations, monthly-mean TMT anomalies were defined relative to climatological monthly means calculated over the 40-yr period from January 1979 to December 2018.

APPENDIX D

Estimation of the Power Spectral Densities

We used Welch’s method (Welch 1967) to estimate power spectral density (PSD) for model simulations and satellite datasets. Welch’s method divides the time series of length L into overlapping windowed segments of length M . The variance in the estimated PSD is reduced by averaging periodograms over the segments. In our analysis, we obtain estimates of one-sided PSD (i.e., the TMT time series are real numbers) by applying a Hamming window to each segment and overlapping individual segments by 50% of the window length.

We used the Mathworks MATLAB’s function “pwelch” to compute PSD. By default, the time series length L is divided into the longest possible segments to obtain close to (but not exceeding) eight segments with 50% overlap. The modified periodograms are averaged to obtain the PSD estimate. Since we are operating on monthly-mean data for the period January 1979 to December 2018, $L = 480$, $M = 106$, and the sampling rate (the number of samples per unit time) is 12 yr^{-1} .

Welch’s method has been shown to work well when dealing with noisy time series. It provides robust estimates of PSD when the shape of the spectrum is not known a priori. For the control runs, the spectrum for each climate model was computed as the average of the individual spectra from all maximally overlapping 480-month chunks of that model’s TMT time series (Pelletier 1998). Use of 480-month chunks ensures

that the climate model spectra and observed spectra are compared at the same frequencies.

For the HIST runs, the spectrum for each climate model was calculated using the same 480-month analysis period (January 1979–December 2018) for which satellite data were available. If more than one HIST realization was available per climate model, we first calculated the individual spectra for each realization and then averaged over realizations.

Power-law representation of spectra

A power-law (PL) fit to PSD is a commonly used method for representing the behavior of spectral densities as a function of frequency (Hasselmann 1976; Pelletier 1998; Vyushin and Kushner 2009; Fredriksen and Rypdal 2016, 2017; Zhu et al. 2019). Under a PL model, the spectral density $S(f)$ depends on the frequency f as follows: $S(f) \propto f^{-\beta}$. By log-transforming both $S(f)$ and f , the parameter β can be easily estimated as the negative slope of the least squares fit to PSD (Paige 1979). Simple measures of goodness of fit can then be computed. Here we used the coefficient of determination R^2 , which ranges from 0 to 1 and is the fraction of variation in $S(f)$ that is explained by the linear fit. The popularity of the power-law model in climate-related work is partly related to interest in the long-term memory properties of the time series, and in the question of how spectral density at low frequencies scales with spectral density at higher frequencies (Fredriksen and Rypdal 2016).

REFERENCES

- AchutaRao, K., and K. R. Sperber, 2006: ENSO simulation in coupled ocean–atmosphere models: Are the current models better? *Climate Dyn.*, **27** (1), 1–15, <https://doi.org/10.1007/s00382-006-0119-7>.
- , B. D. Santer, P. J. Gleckler, K. E. Taylor, D. W. Pierce, T. P. Barnett, and T. M. L. Wigley, 2006: Variability of ocean heat uptake: Reconciling observations and models. *J. Geophys. Res.*, **111**, C05019, <https://doi.org/10.1029/2005JC003136>.
- Alvarez-Esteban, P. C., C. Euán, and J. Ortega, 2016: Time series clustering using the total variation distance with applications in oceanography. *Environmetrics*, **27**, 355–369, <https://doi.org/10.1002/env.2398>.
- Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor, 2012: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere–ocean climate models. *Geophys. Res. Lett.*, **39**, L09712, <https://doi.org/10.1029/2012GL051607>.
- Bindoff, N. L., and Coauthors, 2013: Detection and attribution of climate change: From global to regional. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 867–952.
- Bloomfield, P., and D. Nychka, 1992: Climate spectra and detecting climate change. *Climatic Change*, **21**, 275–287, <https://doi.org/10.1007/BF00139727>.
- Bowers, M. C., and W. Tung, 2018: Variability and confidence intervals for the mean of climate data with short- and long-range dependence. *J. Climate*, **31**, 6135–6156, <https://doi.org/10.1175/JCLI-D-17-0090.1>.
- Cheung, A. H., M. E. Mann, B. A. Steinman, L. M. Frankcombe, M. H. England, and S. K. Miller, 2017a: Comparison of low-frequency internal climate variability in CMIP5 models and observations. *J. Climate*, **30**, 4763–4776, <https://doi.org/10.1175/JCLI-D-16-0712.1>.

- , —, —, —, —, and —, 2017b: Reply to “Comment on comparison of low-frequency internal climate variability in CMIP5 models and observations.” *J. Climate*, **30**, 9773–9782, <https://doi.org/10.1175/JCLI-D-17-0531.1>.
- Christy, J. R., 2015: Data or dogma? Promoting open inquiry in the debate over the magnitude of human impact on Earth’s climate. Hearing in front of the U.S. Senate Committee on Commerce, Science, and Transportation, Subcommittee on Space, Science, and Competitiveness, Testimony, https://www.commerce.senate.gov/public/_cache/files/fcbf4cb6-3128-4fdc-b524-7f2ad4944c1d/80931BD995AF75BA7B819A51ADA9CE99.dr.-john-christy-testimony.pdf.
- Deser, C., A. S. Phillips, M. A. Alexander, and B. V. Smoliak, 2014: Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *J. Climate*, **27**, 2271–2296, <https://doi.org/10.1175/JCLI-D-13-00451.1>.
- , and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277–286, <https://doi.org/10.1038/s41558-020-0731-2>.
- England, M. H., and Coauthors, 2014: Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nat. Climate Change*, **4**, 222–227, <https://doi.org/10.1038/nclimate2106>.
- Euan, C., H. Ombao, and J. Ortega, 2015: Spectral synchronicity in brain signals. <https://arxiv.org/abs/1507.05018>, 39 pp.
- Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, <https://doi.org/10.5194/gmd-9-1937-2016>.
- , and Coauthors, 2019: Taking climate model evaluation to the next level. *Nat. Climate Change*, **9**, 102–110, <https://doi.org/10.1038/s41558-018-0355-y>.
- Frankcombe, L. M., M. H. England, M. E. Mann, and B. A. Steinman, 2015: Separating internal variability from the externally forced climate response. *J. Climate*, **28**, 8184–8202, <https://doi.org/10.1175/JCLI-D-15-0069.1>.
- , —, J. B. Kajtar, M. E. Mann, and B. A. Steinman, 2018: On the choice of ensemble mean for estimating the forced signal in the presence of internal variability. *J. Climate*, **31**, 5681–5693, <https://doi.org/10.1175/JCLI-D-17-0662.1>.
- Franzke, C., 2010: Long-range dependence and climate noise characteristics of Antarctic temperature data. *J. Climate*, **23**, 6074–6081, <https://doi.org/10.1175/2010JCLI3654.1>.
- , 2012a: On the statistical significance of surface air temperature trends in the Eurasian Arctic region. *Geophys. Res. Lett.*, **39**, L23705, <https://doi.org/10.1029/2012GL054244>.
- , 2012b: Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *J. Climate*, **25**, 4172–4183, <https://doi.org/10.1175/JCLI-D-11-00293.1>.
- , S. M. Osprey, P. Davini, and N. W. Watkins, 2015: A dynamical systems explanation of the Hurst effect and atmospheric low-frequency variability. *Sci. Rep.*, **5**, 9068, <https://doi.org/10.1038/srep09068>.
- Fredriksen, H.-B., and K. Rypdal, 2016: Spectral characteristics of instrumental and climate model surface temperatures. *J. Climate*, **29**, 1253–1268, <https://doi.org/10.1175/JCLI-D-15-0457.1>.
- , and M. Rypdal, 2017: Long-range persistence in global surface temperatures explained by linear multibox energy balance models. *J. Climate*, **30**, 7157–7168, <https://doi.org/10.1175/JCLI-D-16-0877.1>.
- Fu, Q., and C. M. Johanson, 2004: Stratospheric influences on MSU-derived tropospheric temperature trends: A direct error analysis. *J. Climate*, **17**, 4636–4640, <https://doi.org/10.1175/JCLI-3267.1>.
- , —, S. G. Warren, and D. J. Seidel, 2004: Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature*, **429**, 55–58, <https://doi.org/10.1038/nature02524>.
- Fyfe, J. C., and Coauthors, 2016: Making sense of the early-2000s warming slowdown. *Nat. Climate Change*, **6**, 224–228, <https://doi.org/10.1038/nclimate2938>.
- , and Coauthors, 2017: Large near-term projected snowpack loss over the western United States. *Nat. Commun.*, **8**, 14996, <https://doi.org/10.1038/ncomms14996>.
- Gil-Alana, L. A., 2005: Statistical modeling of the temperatures in the Northern Hemisphere using fractional integration techniques. *J. Climate*, **18**, 5357–5369, <https://doi.org/10.1175/JCLI3543.1>.
- Gillett, N. P., M. R. Allen, and S. F. B. Tett, 2000: Modelled and observed variability in atmospheric vertical temperature structure. *Climate Dyn.*, **16**, 49–61, <https://doi.org/10.1007/PL00007921>.
- , and Coauthors, 2016: The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6. *Geosci. Model Dev.*, **9**, 3685–3697, <https://doi.org/10.5194/gmd-9-3685-2016>.
- Hasselmann, K., 1976: Stochastic climate models, Part I. Theory. *Tellus*, **28**, 473–485, <https://doi.org/10.3402/tellusa.v28i6.11316>.
- , 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology of Tropical Oceans*, D. B. Shaw, Ed., Royal Meteorological Society, 251–259, <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1976.tb00696.x>.
- Hawkins, E., R. S. Smith, J. M. Gregory, and D. A. Stainforth, 2016: Irreducible uncertainty in near-term climate projections. *Climate Dyn.*, **46**, 3807–3819, <https://doi.org/10.1007/s00382-015-2806-8>.
- Hegerl, G. C., and Coauthors, 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 663–745.
- Hurst, H. E., 1950: Long-term storage capacity of reservoirs. *Proc. Amer. Soc. Civ. Eng.*, **76** (4), 1–30.
- Imbers, J., A. Lopez, C. Huntingford, and M. Allen, 2014: Sensitivity of climate change detection and attribution to the characterization of internal climate variability. *J. Climate*, **27**, 3477–3491, <https://doi.org/10.1175/JCLI-D-12-00622.1>.
- Kajtar, J. B., M. Collins, L. M. Frankcombe, M. H. England, T. J. Osborn, and M. Juniper, 2019: Global mean surface temperature response to large-scale patterns of variability in observations and CMIP5. *Geophys. Res. Lett.*, **46**, 2232–2241, <https://doi.org/10.1029/2018GL081462>.
- Karl, T. R., and Coauthors, 2015: Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, **348**, 1469–1472, <https://doi.org/10.1126/science.aaa5632>.
- Kim, W. M., S. Yeager, P. Chang, and G. Danabasoglu, 2018: Low-frequency North Atlantic climate variability in the Community Earth System Model large ensemble. *J. Climate*, **31**, 787–813, <https://doi.org/10.1175/JCLI-D-17-0193.1>.
- Knutson, T. R., F. Zeng, and A. T. Wittenberg, 2013: Multimodel assessment of regional surface temperature trends: CMIP3 and CMIP5 twentieth-century simulations. *J. Climate*, **26**, 8709–8743, <https://doi.org/10.1175/JCLI-D-12-00567.1>.
- Kopp, G., and J. L. Lean, 2011: A new, lower value of total solar irradiance: Evidence and climate significance. *Geophys. Res. Lett.*, **38**, L01706, <https://doi.org/10.1029/2010GL045777>.
- Koscielny-Bunde, E., A. Bunde, S. Havlin, H. E. Roman, Y. Goldreich, and H.-J. Schellnhuber, 1998: Indication of a

- universal persistence law governing atmospheric variability. *Phys. Rev. Lett.*, **81**, 729–732, <https://doi.org/10.1103/PhysRevLett.81.729>.
- Kravtsov, S., 2017: Comment on “Comparison of low-frequency internal climate variability in CMIP5 models and observations.” *J. Climate*, **30**, 9763–9772, <https://doi.org/10.1175/JCLI-D-17-0438.1>.
- , C. Grimm, and S. Gu, 2018: Global-scale multidecadal variability missing in state-of-the-art climate models. *npj Climate Atmos. Sci.*, **1**, 34, <https://doi.org/10.1038/s41612-018-0044-6>.
- Lee, J., K. R. Sperber, P. J. Gleckler, C. J. W. Bonfils, and K. E. Taylor, 2019: Quantifying the agreement between observed and simulated extratropical modes of interannual variability. *Climate Dyn.*, **52**, 4057–4089, <https://doi.org/10.1007/s00382-018-4355-4>.
- Maher, N., S. McGregor, M. H. England, and A. Sen Gupta, 2015: Effects of volcanism on tropical variability. *Geophys. Res. Lett.*, **42**, 6024–6033, <https://doi.org/10.1002/2015GL064751>.
- Mann, M. E., 2011: On long range dependence in global surface temperature series. *Climatic Change*, **107**, 267–276, <https://doi.org/10.1007/s10584-010-9998-z>.
- Mears, C. A., and F. J. Wentz, 2016: Sensitivity of satellite-derived tropospheric temperature trends to the diurnal cycle adjustment. *J. Climate*, **29**, 3629–3646, <https://doi.org/10.1175/JCLI-D-15-0744.1>.
- Meehl, G. A., H. Teng, and J. M. Arblaster, 2014: Climate model simulations of the observed early-2000s hiatus of global warming. *Nat. Climate Change*, **4**, 898–902, <https://doi.org/10.1038/nclimate2357>.
- Meinshausen, M., and Coauthors, 2011: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, **109**, 213–241, <https://doi.org/10.1007/s10584-011-0156-z>.
- Myhre, G., and Coauthors, 2013: Anthropogenic and natural radiative forcing. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 659–740.
- Paige, C. C., 1979: Computer solution and perturbation analysis of generalized linear least squares problems. *Math. Comput.*, **33**, 171–183, <https://doi.org/10.1090/S0025-5718-1979-0514817-3>.
- Pausata, F. S. R., L. Chafik, R. Caballero, and D. S. Battisti, 2015: Impacts of high-latitude volcanic eruptions on ENSO and AMOC. *Proc. Natl. Acad. Sci. USA*, **112**, 13 784–13 788, <https://doi.org/10.1073/pnas.1509153112>.
- Pelletier, J. D., 1998: The power spectral density of atmospheric temperature from time scales of 10^{-2} to 10^6 yr. *Earth Planet. Sci. Lett.*, **158**, 157–164, [https://doi.org/10.1016/S0012-821X\(98\)00051-X](https://doi.org/10.1016/S0012-821X(98)00051-X).
- Po-Chedley, S., T. J. Thorsen, and Q. Fu, 2015: Removing diurnal cycle contamination in satellite-derived tropospheric temperatures: Understanding tropical tropospheric trend discrepancies. *J. Climate*, **28**, 2274–2290, <https://doi.org/10.1175/JCLI-D-13-00767.1>.
- Riahi, K., and Coauthors, 2017: The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environ. Change*, **42**, 153–168, <https://doi.org/10.1016/j.gloenvcha.2016.05.009>.
- Rypdal, M., and K. Rypdal, 2014: Long-memory effects in linear response models of Earth's temperature and implications for future global warming. *J. Climate*, **27**, 5240–5258, <https://doi.org/10.1175/JCLI-D-13-00296.1>.
- Santer, B. D., and Coauthors, 2005: Amplification of surface temperature trends and variability in the tropical atmosphere. *Science*, **309**, 1551–1556, <https://doi.org/10.1126/science.1114867>.
- , and Coauthors, 2011: Separating signal and noise in atmospheric temperature changes: The importance of timescale. *J. Geophys. Res.*, **116**, D22105, <https://doi.org/10.1029/2011JD016263>.
- , and Coauthors, 2013a: Human and natural influences on the changing thermal structure of the atmosphere. *Proc. Natl. Acad. Sci. USA*, **110**, 17 235–17 240, <https://doi.org/10.1073/pnas.1305332110>.
- , and Coauthors, 2013b: Identifying human influences on atmospheric temperature. *Proc. Natl. Acad. Sci. USA*, **110**, 26–33, <https://doi.org/10.1073/pnas.1210514109>.
- , and Coauthors, 2017a: Causes of differences in model and satellite tropospheric warming rates. *Nat. Geosci.*, **10**, 478–485, <https://doi.org/10.1038/ngeo2973>.
- , and Coauthors, 2017b: Comparing tropospheric warming in climate models and satellite data. *J. Climate*, **30**, 373–392, <https://doi.org/10.1175/JCLI-D-16-0333.1>.
- , and Coauthors, 2018: Human influence on the seasonal cycle of tropospheric temperature. *Science*, **361**, eaas8806, <https://doi.org/10.1126/science.aas8806>.
- , J. Fyfe, S. Solomon, J. Painter, C. Bonfils, G. Pallotta, and M. Zelinka, 2019: Quantifying stochastic uncertainty in detection time of human-caused climate signals. *Proc. Natl. Acad. Sci. USA*, **116**, 19 821–19 827, <https://doi.org/10.1073/pnas.1904586116>.
- Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464, <https://doi.org/10.1214/aos/1176344136>.
- Solomon, S., J. S. Daniel, R. R. Neely, J.-P. Vernier, E. G. Dutton, and L. W. Thomason, 2011: The persistently variable “background” stratospheric aerosol layer and global climate change. *Science*, **333**, 866–870, <https://doi.org/10.1126/science.1206027>.
- , D. J. Ivy, D. Kinnison, M. J. Mills, R. R. Neely III, and A. Schmidt, 2016: Emergence of healing in the Antarctic ozone layer. *Science*, **353**, 269–274, <https://doi.org/10.1126/science.aae0061>.
- Spencer, R. W., J. R. Christy, and W. D. Braswell, 2017: UAH version 6 global satellite temperature products: Methodology and results. *Asia-Pac. J. Atmos. Sci.*, **53**, 121–130, <https://doi.org/10.1007/s13143-017-0010-y>.
- Steinman, B. A., M. E. Mann, and S. K. Miller, 2015: Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures. *Science*, **347**, 988–991, <https://doi.org/10.1126/science.1257856>.
- Stone, P. H., and J. H. Carlson, 1979: Atmospheric lapse rate regimes and their parameterization. *J. Atmos. Sci.*, **36**, 415–423, [https://doi.org/10.1175/1520-0469\(1979\)036<0415:ALRRAT>2.0.CO;2](https://doi.org/10.1175/1520-0469(1979)036<0415:ALRRAT>2.0.CO;2).
- Swart, N. C., S. T. Gille, J. C. Fyfe, and N. P. Gillett, 2018: Recent Southern Ocean warming and freshening driven by greenhouse gas emissions and ozone depletion. *Nat. Geosci.*, **11**, 836–841, <https://doi.org/10.1038/s41561-018-0226-1>.
- Taqqu, M. S., V. Teverovsky, and W. Willinger, 1995: Estimators for long-range dependence: An empirical study. *Fractals*, **03**, 785–798, <https://doi.org/10.1142/S0218348X95000692>.
- Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, <https://doi.org/10.1175/BAMS-D-11-00094.1>.
- Trenberth, K. E., and J. T. Fasullo, 2010: Simulation of present-day and twenty-first-century energy budgets of the Southern Oceans. *J. Climate*, **23**, 440–454, <https://doi.org/10.1175/2009JCLI3152.1>.
- van Vuuren, D. P., and Coauthors, 2011: The representative concentration pathways: An overview. *Climatic Change*, **109**, 5–31, <https://doi.org/10.1007/s10584-011-0148-z>.
- Vyushin, D. I., and P. J. Kushner, 2009: Power-law and long-memory characteristics of the atmospheric general circulation. *J. Climate*, **22**, 2890–2904, <https://doi.org/10.1175/2008JCLI2528.1>.

- , —, and J. Mayer, 2009: On the origins of temporal power-law behavior in the global atmospheric circulation. *Geophys. Res. Lett.*, **36**, L14706, <https://doi.org/10.1029/2009GL038771>.
- , —, and F. Zwiers, 2012: Modeling and understanding persistence of climate variability. *J. Geophys. Res.*, **117**, D21106, <https://doi.org/10.1029/2012JD018240>.
- Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *J. Climate*, **23**, 4175–4191, <https://doi.org/10.1175/2010JCLI3594.1>.
- Welch, P., 1967: The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.*, **15**, 70–73, <https://doi.org/10.1109/TAU.1967.1161901>.
- Wigley, T. M. L., and S. C. B. Raper, 1990: Natural variability of the climate system and detection of the greenhouse effect. *Nature*, **344**, 324–327, <https://doi.org/10.1038/344324a0>.
- Zelinka, M. D., T. Andrews, P. M. Forster, and K. E. Taylor, 2014: Quantifying components of aerosol–cloud–radiation interactions in climate models. *J. Geophys. Res. Atmos.*, **119**, 7599–7615, <https://doi.org/10.1002/2014JD021710>.
- , T. A. Myers, D. T. McCoy, S. Po-Chedley, P. M. Caldwell, P. Ceppi, S. A. Kleini, and K. E. Taylor, 2020: Causes of higher climate sensitivity in CMIP6 models. *Geophys. Res. Lett.*, **47**, e2019GL085782, <https://doi.org/10.1029/2019GL085782>.
- Zhu, F., and Coauthors, 2019: Climate models can correctly simulate the continuum of global-average temperature variability. *Proc. Natl. Acad. Sci. USA*, **116**, 8728–8733, <https://doi.org/10.1073/pnas.1809959116>.
- Zhu, X., K. Fraedrich, Z. Liu, and R. Blender, 2010: A demonstration of long-term memory and climate predictability. *J. Climate*, **23**, 5021–5029, <https://doi.org/10.1175/2010JCLI3370.1>.
- Zou, C.-Z., M. D. Goldberg, and X. Hao, 2018: New generation of U.S. satellite microwave sounder achieves high radiometric stability performance for reliable climate change detection. *Sci. Adv.*, **4**, eaau0049, <https://doi.org/10.1126/sciadv.aau0049>.