Comparison and assessment of large-scale surface temperature in climate model simulations

Raquel Barata[1], Raquel Prado[1], Bruno Sansó[1], Benjamin D. Santer[2], and Giuliana Pallotta[2]

[1]University of California Santa Cruz, Santa Cruz, CA 95064
[2]Lawrence Livermore National Laboratory, Livermore, CA 94550

Correspondence: Raquel Barata (rbarata@ucsc.edu)

Abstract. We assess the behavior of large-scale spatial averages of surface temperature in climate model simulations and in reanalysis products. We rely on univariate and multivariate Dynamic Linear Model (DLM) techniques to estimate both long-term and seasonal changes in the externally-forced temperature. The residuals capture the internal variability of the climate system and exhibit complex temporal autocorrelation structure. To characterize this internal variability, we explore the structure of the residuals using univariate and multivariate autoregressive (AR) models. The climate model data analyzed here consist of three different types of numerical experiments from phase 5 of the Coupled Model Intercomparison Project (CMIP5): preindustrial control runs, simulations of historical climate change, and decadal predictions. Our focus is on results from one particular model (MIROC5), as well as on two different reanalysis-based estimates of observed changes in climate (from NCEP-2 and ERA-Interim). The climate variable of interest is monthly-mean 2 meter surface temperature over the time period from January 1981 to December 2010, spatially averaged over four different domains (global, tropical, Northern Hemisphere, and Southern Hemisphere). Our results illustrate differences in all components of the climate "signal" (the response to changes in external forcings), most notably between the reanalysis products and the model-generated simulations. Despite the differences in underlying factors contributing to variability, the three types of simulation yield very similar spectral estimates of internal temperature variability. This is of particular interest for the decadal simulation runs as influence from initialization might be expected. In general, we find there is no evidence that the MIROC5 model systematically underestimates the amplitude of observed surface temperature variability on multi-decadal timescales – a finding that has considerable relevance for efforts to identify human-caused "fingerprints" in observational surface temperature data.

1 Introduction

Exploring the impacts of human-caused climate change is of great relevance and interest to society. The fifth phase of the Coupled Model Intercomparison Project (CMIP5) generated many different ensembles of climate model simulations (Taylor et al., 2012). These simulations have enhanced our scientific understanding of the ability of current models to represent key features of present-day climate. They have also helped to identify human and natural influences on historical climate and to quantify uncertainties in projections of future climate change. The CMIP5 framework incorporates results from large multi-

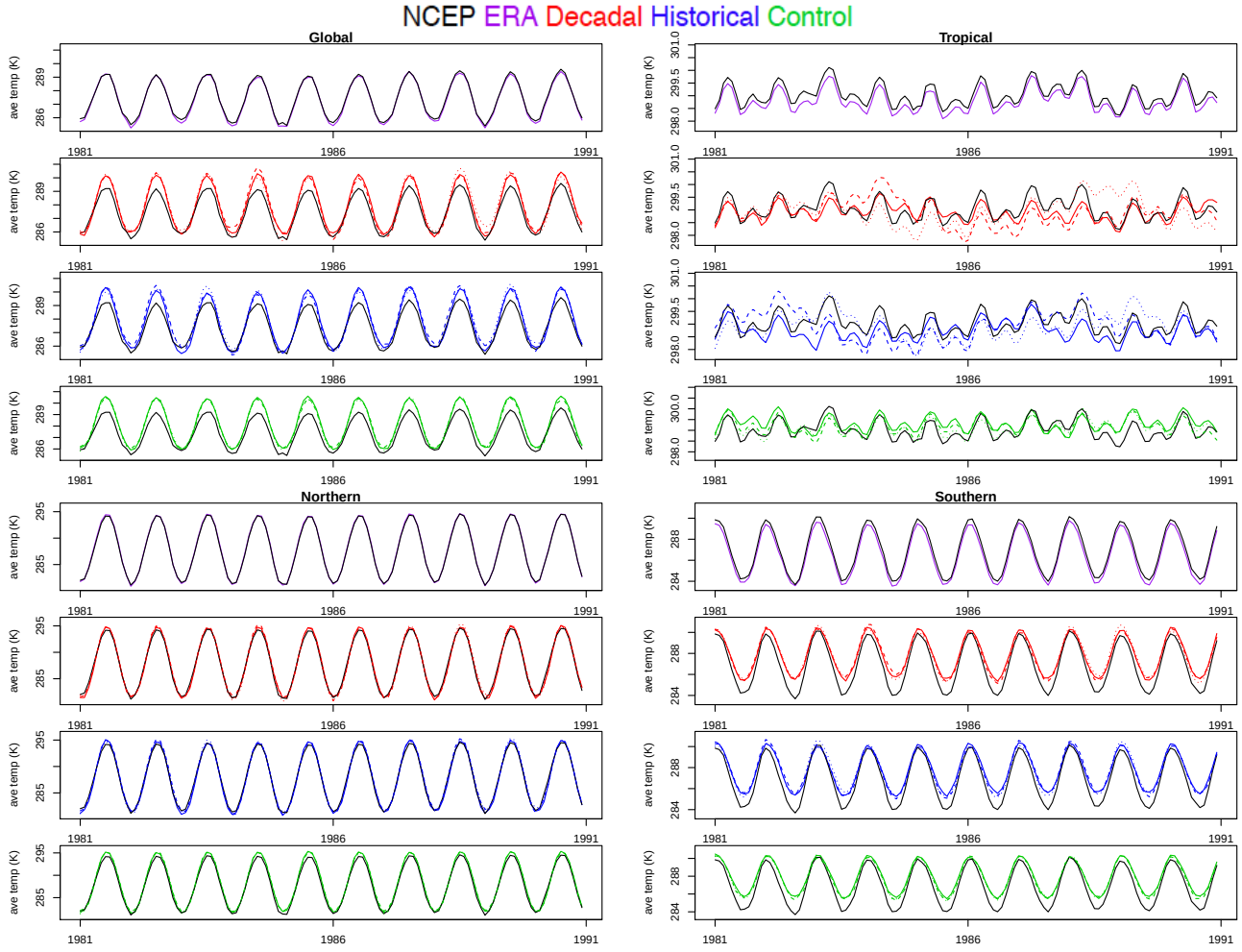

Figure 1. Time series of monthly-mean, spatially averaged 2 meter surface temperature from the MIROC5 model. Results are for the three different types of simulation analyzed here (decadal prediction, historical, and control), as well as for the NCEP-2 and ERA-Interim reanalyses. Three "replicates" of each model simulation are indicated by different line types. Top panels: Global (left) and tropical (right). Bottom panels: Northern (left) and Southern Hemisphere (right).

model ensembles, and frequently includes multiple realizations for each model and type of simulation. More extensive details of the CMIP5 experimental design are found in Taylor (2009).

The goal of this paper is to draw on Bayesian statistical methods to compare the observational record with three different types of CMIP5 simulations: 1) decadal predictions of climate, initialized from a specific observational state; 2) uninitialized simulations driven by estimated historical changes in key anthropogenic and natural forcings; and 3) control integrations with no year-to-year changes in external forcings, which provide estimates of the natural internal variability of the climate system. We seek to determine whether there are statistically significant differences in aspects of the variability, both between the model

simulations and the reanalysis products and within the three different types of simulation. We also examine model errors in mean temperature, and as well as differences in variability between the two reanalysis products. Our analysis focuses on monthly-mean 2 meter surface temperature time series from January 1981 to December 2010. All model and observational data are available in gridded form for a global domain. Results are spatially averaged over four regions: global, tropical, Northern Hemisphere and Southern Hemisphere. The latitudinal boundaries of these regions are: 90°S to 90°N, 20°S to 20°N, 0° to 90°N and 90°S to 0° (respectively). We calculate area-weighted spatial averages for each of the four regions of interest. This allows us to explore the sensitivity of our results to spatial differences in the large-scale structure of the "signal" (the climate response to imposed changes in external forcings) and the "noise" of natural internal variability.

Our study relies on simulations generated using version 5 of the atmosphere–ocean General Circulation Model (A/OGCM) developed jointly by the Atmosphere and Ocean Research Institute at the University of Tokyo, the National Institute for Environmental Studies, and the Japan Agency for Marine-Earth Science and Technology. This is commonly referred to as the Model for Interdisciplinary Research on Climate (MIROC5) (see Watanabe et al., 2010). As an example of the data considered here, we show in Figure 1 the first 10 years of the three different types of simulation analyzed. The globally averaged data exhibit a pronounced annual cycle, which is clearly dominated by the Northern Hemisphere. As expected based on the changes in incoming solar radiation as a function of latitude and season, the phasing of the annual cycle differs in the Northern and Southern Hemispheres. A semi-annual cycle is also visible in the tropics (Santer et al., 2018).

The analysis performed in this paper involves decomposing each temperature time series into what we refer to as a "baseline" and a seasonal component. The baseline captures long-term externally forced changes, as well as short-term cooling responses to volcanic eruptions. The seasonal component is dominated by the externally forced annual and semi-annual cycles. We extract the baseline temperature and seasonality of the climate model simulations and compare them to the corresponding components in the observational products. For each simulation type and region, the baseline and seasonal components are removed from the time series using Dynamic Linear Models (DLMs). This yields residual time series that primarily represent natural internal climate variability. As in Imbers et al. (2014), our focus is on investigating the spectral characteristics of internal variability with autoregressive (AR) models. We compare the properties of these AR spectra in the climate model simulations and the two reanalyses. We seek to determine whether model-versus-observed spectral differences are significant, and can be interpreted in terms of known model deficiencies (such as systematic errors in external forcings; see Schmidt et al., 2014). We find pronounced model-versus-observed dissimilarities in all three components of interest here: the baseline temperature, seasonal amplitudes inferred from the DLMs, and in the AR spectral characteristics of the residuals. A second objective is to investigate whether there are identifiable differences between the spectral properties in the decadal prediction, historical, and control simulations that are related to such factors as the inclusion of external forcings and the initialization approach.

The paper is organized as follows. In Section 2, we describe the model simulations and the reanalysis products analyzed. Section 3 presents our statistical modeling approach and introduces the DLM used for estimating the baseline and seasonal components of the time series. It also describes the AR model that we apply to the residual time series in order to estimate natural internal variability. In Section 4, we show the results obtained from the application of the DLM and AR models to the surface temperature time series for the four regions previously mentioned. Section 5 provides a summary and brief discussion.

2 Data

2.1 Climate model simulations

CMIP5 is a coordinated international modeling activity involving a large suite of simulations performed with several dozen different climate models. We focus here on simulations performed with one particular climate model (MIROC5). As mentioned above, we analyze both forced and unforced climate simulations. The forced decadal prediction and historical runs are used to explore the response of the climate system to specified historical changes in anthropogenic and natural external factors. Examples of such external factors include human-caused changes in well-mixed greenhouse gases and natural changes in volcanic aerosols (Kirtman et al., 2013). The forced simulations also generate natural internal variability of the climate system. In contrast, the MIROC5 control integration yields an estimate of "pure" natural internal variability, uncontaminated by externally forced climate changes. Below, we briefly describe the three types of climate simulation that are of interest here.

Decadal prediction simulations are the newest addition to the CMIP activity, and are therefore the most exploratory. These near-term simulations were organized through a collaboration between the World Climate Research Programme's Working Group on Coupled Modelling (WGCM) and the Working Group on Seasonal to Interannual Prediction (WGSIP). There are two core sets of these near-term experiments. The first is a set of 10-year hindcasts initialized from a number of different observational starting points. Such simulations allow analysts to assess the prediction skill and to investigate the sensitivity of skill to differences in the initial state (e.g., to the presence or absence of a strong El Niño or La Niña). The second set of decadal prediction runs extended the 10-year hindcasts to 30 years. The influence of external forcing is more prominent in these longer simulations (Taylor et al., 2012). The period from 1981 to 2010 is one of the few periods for which 30-year long decadal simulations are available. This dictates the time period and the length of our analysis. The decadal prediction runs include the same time-varying anthropogenic and natural external forcings that are used in the historical simulations.

The modeling groups participating in CMIP5 used different methods and observational data sets for initializing the decadal simulations. Most initialization schemes utilize observed ocean and sea ice conditions. A full discussion of initialization methods and the organization of the decadal prediction simulations can be found in Meehl et al. (2009).

Six individual realizations ("replicates") of the MIROC5 decadal prediction run were available (see Figure 1). Each realization has small differences in the initial state in 1981. These small initial differences amplify with time, eventually yielding different sequences of natural internal variability in each realization (Kirtman et al., 2013).

Historical runs are not initialized from a specific observed three-dimensional ocean state. Such simulations typically commence from estimated atmospheric greenhouse gas levels in 1850 or 1860, and are then run until the early 21st century. Like the decadal simulations, the historical simulations are driven by estimated changes in well-mixed greenhouse gases, particulate pollution, land surface properties, solar irradiance, and volcanic aerosols. The MIROC5 historical integrations span the period from 1850 to 2012; five historical replicates were available. To facilitate comparison with the decadal analysis, our analysis of the historical runs is restricted to the period from January 1981 to December 2010 (see Figure 1).

As noted above, the decadal and historical simulations are performed with exactly the same physical climate model using identical anthropogenic and natural external forcings. Differences between the MIROC5 historical and decadal prediction runs

are related to the initialization of the latter. Initialization forces the model ocean temperature and sea-ice to be consistent with the estimated observational state in 1981. No such consistency with observations is imposed in the historical run. The two types of simulation can therefore produce noticeably different climate states in 1981. This difference is due to two factors. First, any systematic model errors (in either the applied forcings and/or the climate response to these forcings) should begin to manifest within 1-2 years of the start of the historical run in 1850, causing the simulated climate in the historical run to drift away from observed climate. Second, even if there were no model forcing or response errors, the phasing of internal variability is different in the historical and decadal prediction runs – so the mean states of these two types of simulation are unlikely to be exactly the same in 1981 (except by chance).

In observational climate records, internal variability must be statistically separated from other sources of variance: it is occurring at the same time as the climate system is responding to multiple external forcings. Control simulations provide estimates of "pure" internal variability, which is an integral component of climate change detection and attribution studies (Santer et al., 2018). In the MIROC5 pre-industrial control simulation analyzed here, there are no year-to-year changes in atmospheric concentrations of greenhouse gases, particulate pollution, volcanic aerosols, or solar irradiance. Changes in climate arise solely from the behavior of modes of variability intrinsic to the coupled atmosphere-ocean-sea ice system. Examples of such modes of variability include the El Niño/Southern Oscillation (ENSO), the Interdecadal Pacific Oscillation (IPO), and the North Atlantic Oscillation (NAO). Control runs are typically used to simulate many centuries of internal variability, and do not have any direct correspondence with actual time. Here, we extract ten 30-year non-overlapping monthly-mean temperature time series from the 670-year MIROC5 control run. Each 30-year segment contains a different unique manifestation of internal variability, so they are similar to the "replicates" available for the decadal prediction and historical runs (see Figure 1).

Several points should be emphasized prior to discussion of the model results. First, the A/OGCM simulations analyzed here generate their own intrinsic variability – i.e., they produce their own sequences of El Niños, La Niñas, and other quasi-periodic modes. In the historical runs, there is no correspondence between the modeled and observed phasing and amplitude of these modes, except by chance. In the decadal prediction runs, the situation is different. The observational ocean data used in the initialization provide some information about the current state of ENSO and other, longer-timescale modes of variability. This observational information constrains (at least in the first 1-2 years after initialization) the climate trajectory that is followed in the decadal prediction run, and imparts some short-term similarity between the simulation and observations. As the length of time after initialization increases, chaotic variability begins to overwhelm the information that the initialization provided about the likely trajectories of real-world modes of internal variability, and the phasing of internal variability begins to diverge in observations and the decadal prediction runs.

Second, the real world, the historical runs, and the decadal prediction simulations contain common components of temperature variability associated with natural changes in solar irradiance and volcanic activity. For the period of interest here (1981 to December 2010), the main solar forcing of interest is the roughly 11-year solar cycle (Kopp and Lean, 2011). The major volcanic eruptions are those of El Chichón in 1982 and Pinatubo in 1991. Both eruptions produced short-term (1-2 year) cooling of the Earth's surface, followed by gradual recovery to pre-eruption temperature levels (Santer et al., 2001). As noted above, the control simulation does not include any solar or volcanic forcing, so each 30-year control segment should not exhibit any

synchronicity between the simulated and observed temperature variability (except by chance). Further details of the MIROC5 model and the simulations performed with it can be found in Watanabe et al. (2010).

2.2 Reanalysis products

We compare the climate model simulations to two different reanalysis data sets. Reanalyses rely on a state-of-the-art numerical weather prediction (NWP) model to produce internally and physically consistent estimates of changes in real-world climate. The NWP model assimilates raw observational data from satellites, radiosondes, aircraft, land surface measurements, and many other sources, and produces an optimal "blend" of the assimilated data. A key point is that reanalyses are retrospective – the forecast model does not change over time, so the reanalysis output is not contaminated by spurious changes in climate associated with progressive improvement of the forecast model, or by changes over time to the assimilation system. A number of different groups around the world have generated reanalysis-based estimates of historical climate change. Each group uses a different NWP model and assimilation system, and makes different subjective judgments regarding the types of observations that are assimilated, the weights applied to each data type, and the bias correction procedures applied to the ingested observations. This leads to differences in the estimates of "observed" climate change and climate variability generated by different reanalysis products (Kalnay et al., 1996). These differences have generally decreased over time, as NWP models and assimilation methods have improved.

We use two reanalysis products here. The first is version 2 of the reanalysis performed by the National Centers for Environmental Prediction (NCEP), referred to as subsequently as NCEP-2. Although we only consider data for our time period of interest, NCEP-2 spans the longer period from 1979 to 2016. Further details of NCEP-2 are available in Kanamitsu et al. (2002). The second reanalysis was generated by the European Centre for Medium-Range Weather Forecasts (ECMWF) in collaboration with a number of other institutions. We refer to this subsequently as ERA-Interim (ERA-I). It begins in 1979 and is continuously updated. Results from both reanalyses are shown in Figure 1. For a detailed documentation of ERA-I, see Berrisford et al. (2011) and Dee et al. (2011). A more thorough discussion and comparison of these reanalyses is available in Fujiwara et al. (2017).

3 Statistical models for model-generated and reanalysis time series

DLMs are a popular Bayesian modeling approach for the analysis of non-stationary time series. We follow approaches detailed in Harrison and West (1999) and Prado and West (2010) in order to estimate time-varying baseline and seasonality components. In Section 3.1, we present the DLM used here to extract baseline and seasonal components from the spatially averaged model and reanalysis surface temperature time series. Section 3.2 details the multivariate extensions of the univariate analysis that are required to deal with the availability of multiple replicates of the model simulations. Section 3.3 presents our DLM discount factor selection strategy, which relies on evolution variance specification as a method for extracting comparable components of the climate signal. Section 3.4 describes a Bayesian approach to fitting autoregressive models to the DLM residuals; the goal here is to capture the internal variability of the climate model simulations. Finally, Section 3.5 presents the results from a

comparison of the internal variability in the three different types of model simulation and the two reanalyses. This comparison relies on inferred spectral densities via the total variation distance.

3.1 Baseline and seasonal temperature estimation

Consider first a single reanalysis time series for one of the four domains considered. Let y_t denote the univariate domain-average temperature at time t, for $t = 1, \ldots, T$ where $T = 360$ months (30 years). We aim to decompose each time series into a baseline temperature $\eta_{1,t}$ and seasonal components $\alpha^k_{1,t}$ for harmonics $k = 1, \ldots, K$. Let $N_d(\mathbf{m}, \mathbf{S})$ denote a d-dimensional normal with mean $\mathbf{m}$ and variance $\mathbf{S}$. We specify our model used to emulate the baseline and seasonality of the data as a second-order polynomial DLM with Fourier form seasonality, i.e.,

$$y_t = \eta_{1,t} + \sum_{j=1}^{K} \alpha^k_{1,t} + \nu_t, \quad \nu_t \sim N(0, V) \tag{1}$$

where V is the unknown observational variance (Harrison and West, 1999). It is also assumed that the observational errors ν_t are independent over time. We further assume that the baseline component has a structure described by:

$$\begin{pmatrix} \eta_{1,t} \\ \eta_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{1,t-1} \\ \eta_{2,t-1} \end{pmatrix} + \boldsymbol{\omega}^{\eta}_t, \quad \boldsymbol{\omega}^{\eta}_t \sim N_2(\mathbf{0}, V\mathbf{W}^{\eta}_t) \, . \tag{2}$$

Here the system evolution error vectors $\boldsymbol{\omega}^{\eta}_t$ are assumed to be independent over time. We denote the baseline evolution matrix as $\mathbf{G}^{\eta} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. A maximum of $\lfloor p/2 \rfloor$ harmonics can be included in the model where p is the fundamental period. Here, $p = 12$ months, which is the annual cycle. We choose to include harmonics $1, .., K$ with $K = 4$ in the seasonal component of the DLM to capture the annual, semi-annual, trimestral and quarterly cycles. Statistical assessment based on the calculation of the highest posterior density regions indicated that higher order harmonics were not significant. Each harmonic k included in the model is described with a Fourier form representation of cyclical functions, given as

$$\begin{pmatrix} \alpha^k_{1,t} \\ \alpha^k_{2,t} \end{pmatrix} = \begin{pmatrix} \cos(\frac{2\pi}{p}k) & \sin(\frac{2\pi}{p}k) \\ -\sin(\frac{2\pi}{p}k) & \cos(\frac{2\pi}{p}k) \end{pmatrix} \begin{pmatrix} \alpha^k_{1,t-1} \\ \alpha^k_{2,t-1} \end{pmatrix} + \boldsymbol{\omega}^{\alpha,k}_t, \quad \boldsymbol{\omega}^{\alpha,k}_t \sim N_2(\mathbf{0}, V\mathbf{W}^{\alpha,k}_t) \, . \tag{3}$$

We denote the kth seasonal evolution matrix $\mathbf{G}^{\alpha,k} = \begin{pmatrix} \cos(\frac{2\pi}{p}k) & \sin(\frac{2\pi}{p}k) \\ -\sin(\frac{2\pi}{p}k) & \cos(\frac{2\pi}{p}k) \end{pmatrix}$. It is assumed that $\boldsymbol{\omega}^{\alpha,k}_t$ are independent over time, as well as independent of $\boldsymbol{\omega}^{\eta}_t$ for $t = 1, \ldots, T$.

Using the superposition principle (Harrison and West, 1999), we write the model as a hierarchy with an observation equation and a system equation, as

$$y_t = \mathbf{F}'\boldsymbol{\theta}_t + \nu_t, \quad \nu_t \sim N(0, V) \tag{4}$$

$$\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_n(\mathbf{0}, V\mathbf{W}_t) \, . \tag{5}$$

Here $n = 2 + 2K$ is the length of state vector $\boldsymbol{\theta}_t$. $\mathbf{G}$ and $\mathbf{W}_t$ are defined, respectively, as $\mathbf{G} = \text{blockdiag}(\mathbf{G}^{\eta}, \mathbf{G}^{\alpha,1}, \ldots, \mathbf{G}^{\alpha,K})$ and $\mathbf{W}_t = \text{blockdiag}(\mathbf{W}^{\eta}_t, \mathbf{W}^{\alpha,1}_t, \ldots, \mathbf{W}^{\alpha,K}_t)$. The state vector $\boldsymbol{\theta}_t$ takes the form $\boldsymbol{\theta}_t = (\eta_{1,t}, \eta_{2,t}, \alpha^1_{1,t}, \alpha^1_{2,t}, \ldots, \alpha^K_{1,t}, \alpha^K_{2,t})$ where $\mathbf{F}' = (\mathbf{F}^{\eta'}, \mathbf{F}^{\alpha,1'}, \ldots, \mathbf{F}^{\alpha,K'})$ with $\mathbf{F}^{\cdot,\cdot'} = (1, 0)$ for all components.

3.2 Multivariate extension for simulation data

For an ensemble of the R replicates of model simulations from a specified region, we consider a multivariate DLM that is an immediate extension of the univariate case. For the decadal, historical and control experiments, R is 6, 5 and 10 respectively. Let $y_{t,r} = \mathbf{F}'\boldsymbol{\theta} + \nu_{t,r}$ denote the univariate average regional temperature at time t, for $t = 1, \ldots, T$, of replicate $r \in \{1, \ldots, R\}$. Each $\nu_{t,r}$ is independent and identically distributed from $N(0, V)$. Replacing y_t in (4) with a vector of R replicate values $\mathbf{Y}_t = (y_{t,1}, \ldots, y_{t,R})'$ and $\boldsymbol{\nu}_t = (\nu_{t,1}, \ldots, \nu_{t,R})'$ with a vector of R i.i.d error terms, only the observation equation changes, i.e.,

$$\mathbf{Y}_t = \mathbf{F}'\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N_R(0, V\mathbf{I}_R) \tag{6}$$

$$\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_n(0, V\mathbf{W}_t) \tag{7}$$

$\mathbf{F}'$ is now a $R \times n$ dynamic regression matrix with identical rows, $\mathbf{F}'_r = (\mathbf{F}^{\eta'}, \mathbf{F}^{\alpha,1'}, \ldots, \mathbf{F}^{\alpha,K'})$, with components defined in the previous section. As in the univariate case, the multivariate DLM still yields a single estimate for baseline and seasonal components, but this estimate now reflects the overall behavior of the replicates. The internal variability of each individual replicate is captured by the components of $\boldsymbol{\nu}_t$.

Assuming the system evolution covariance matrices $\mathbf{W}_t$ at each time t are known, the posterior distributions for $\boldsymbol{\theta}_t$ at each time can be sequentially updated using the Kalman filtering and backward smoothing methods for unknown constant observational variance (Harrison and West, 1999). Following this approach, conjugate priors are chosen as follows: a Normal distribution for the initial state vector $\boldsymbol{\theta}_0 \sim N_n(\mathbf{m}_0, V\mathbf{C}_0)$ and an Inverse Gamma for the unknown constant $V \sim IG(n_0/2, n_0S_0/2)$ with values $\mathbf{m}_0 = (285, 0, ..., 0)'$, $\mathbf{C}_0 = \text{diag}(5, 2 \times 10^{-6}, 5, 1, ..., 1)$, $n_0 = 1$ and $S_0 = 0.01$.

3.3 Specification of the evolution variance

To complete the model specification, we require the sequence of state evolution variance matrices, $\mathbf{W}_t$. The structure and magnitude of $\mathbf{W}_t$ control stochastic variation and stability of the evolution of the model over time. More precisely, if the posterior variance of the state vector $\boldsymbol{\theta}_{t-1}$ at time $t-1$ is denoted as $Var(\boldsymbol{\theta}_{t-1}|\mathbf{Y}_{1:(t-1)}) = \mathbf{C}_{t-1}$, the sequential updating equations produce the prior variance of $\boldsymbol{\theta}_t$, $\mathbf{R}_t = Var(\boldsymbol{\theta}_t|\mathbf{Y}_{1:(t-1)}) = \mathbf{G}\mathbf{C}_{t-1}\mathbf{G}' + \mathbf{W}_t$. Between observations, the addition of the error term $\boldsymbol{\omega}_t$ leads to an additive increase in the initial uncertainty $\mathbf{G}\mathbf{C}_{t-1}\mathbf{G}'$ of the system variance. Thus it is natural to write $\mathbf{W}_t$ as a fixed proportion of $\mathbf{G}\mathbf{C}_{t-1}\mathbf{G}'$ such that $\mathbf{R}_t = \mathbf{G}\mathbf{C}_{t-1}\mathbf{G}'/\delta \geq \mathbf{G}\mathbf{C}_{t-1}\mathbf{G}'$. Here δ is defined to be a discount factor such that $0 < \delta \leq 1$. This suggests an evolution variance matrix of the form $\mathbf{W}_t = \frac{1-\delta}{\delta}\mathbf{G}\mathbf{C}_{t-1}\mathbf{G}'$, where the $\delta = 1$ results in the static model (Harrison and West, 1999).

Our method utilizes component discounting to specify $\mathbf{W}_t$. In other words, we use one discount factor for the baseline, δ_{base}, and one for the seasonal components, δ_{seas}. The seasonal discount factor is set to be 1, which ensures that the smoothed harmonic estimates do not change over time. This choice makes the DLM seasonal component analogous to calculating a constant climatology, a common practice in climate science. If the seasonal amplitudes are changing over time, the changes are aliased in the DLM residuals. As a sensitivity test, we considered lower seasonal discount factors, which allow the amplitudes to vary over time. This test indicated that high discount factors were generally optimal. For the replicate simulation data, a high

baseline discount value is chosen to limit the amount of replicate variability shown in the posterior estimates of the baseline parameters. The chosen value is the largest discount factor which maximizes the observed predictive density for all 10 control segments individually, as the control "replicates" exhibit the least absolute variance of the three types of model simulation.

The true (but unknown) long-term baseline temperature change can be more reliably estimated from the simulations (which have multiple realizations of signal and noise) than from the single realization of the reanalysis datasets. For each region of interest, therefore, it is necessary to select a baseline discount factor which accounts for this difference in the inherent "noisiness" of the simulated and reanalysis baselines. The selected discount factors ensure that the amplitude of the long-term externally forced variability is comparable in the baselines of the model simulations and the reanalysis products.

A few additional words of explanation are necessary. Our method for extracting a baseline estimate for the climate model ensembles averages the uncorrelated internal variability of the individual members of the ensemble. The result is a much smoother baseline estimate than if we were to consider each realization individually. If the same discount factor was selected for a univariate reanalysis time series, δ_{base}^{obs}, as that chosen for the average of the replicate data, δ_{base}^{mod}, the resulting observational baseline estimates would be more "wiggly" in comparison to those of the replicate data. Mathematically, we require the variance of the evolution error to be of the same magnitude in both cases. To quantify the overlap between the two baseline evolution error distributions $N_2(0, \mathbf{W}_t^{\eta,mod})$ and $N_2(0, \mathbf{W}_t^{\eta,obs})$, we use the Bhattacharyya distance (Derpanis, 2008). More specifically, we select the value of δ_{base}^{obs} that minimizes the cumulative value of the Bhattacharyya distance over time. The value is computed for the comparison between the NCEP reanalysis and the historical ensemble-mean data. It is important to note that the choice of discount factor ensures comparable baseline temperature variability between the reanalyses, the decadal prediction and historical runs within any one region, but not between regions.

3.4 Internal variability assessment method

In addition to estimating the overall temperature baseline and seasonal effects in the three classes of simulations and the reanalyses, we are also interested in describing the internal variability of the MIROC5 model and the two reanalyses, and in assessing whether the model- and reanalysis-based estimates of internal variability are consistent. Because the structure of the DLM proposed above does not account for internal variability, the residuals (after removal of baseline and seasonal components from the time series) will exhibit autocorrelation.

Let z_t denote the residuals obtained by subtracting, for the current spatial domain of interest, the posterior mean of the univariate DLM at time t from a reanalysis time series. That is, $z_t = y_t - \mathbf{F}'\hat{\boldsymbol{\theta}}_t$, where $\hat{\boldsymbol{\theta}}_t$ denotes the posterior mean of $\boldsymbol{\theta}_t$ at time t. We use an autoregressive model of order q, denoted by AR(q), to capture the temporal structure of z_t, i.e.,

$$z_t = \sum_{j=1}^{q} \phi_j z_{t-j} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2), \tag{8}$$

where ϵ_t are independent over time and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_q)$ is the vector of AR coefficients. We initially explored the possibility of using a time-varying model, but found that a standard static AR model was reasonable. For the MIROC5 simulations with R replicates, this univariate model is easily extended to a multivariate autoregressive model. Let $z_{t,r}$ denote the residual time series for replicate $r \in \{1, \ldots, R\}$ for $t = 1, \ldots, T$. Thus, $z_{t,r} = \sum_{j=1}^{q} \phi_j z_{t-j,r} + \epsilon_{t,r}$ with each $\epsilon_{t,r}$ independent and

distributed $N(0,\sigma^2)$. Replace z_t and ϵ_t in (8) with vectors of length R, $\mathbf{Z}_t = (z_{1,t},\ldots,z_{t,R})'$ and $\boldsymbol{\epsilon}_t = (\epsilon_{t,1},\ldots,\epsilon_{t,R})'$ where $\boldsymbol{\epsilon}_t \sim N(\mathbf{0},\sigma^2 I_R)$. This hierarchical AR is chosen, instead of a general vector AR model, to estimate a single vector of autoregression coefficients per climate ensemble. With conjugate priors $\boldsymbol{\phi} \sim N_q(0,\mathbf{I}_q)$ and $\sigma^2 \sim IG(1,0.01)$, it is easy to sample the posterior distributions directly using standard Bayesian linear regression techniques.

In fitting AR models, we make the simplifying assumption that q may vary between the four spatial domains considered here, but that in any one domain, all residual time series for the three types of simulation and the two reanalyses have the same order q. This ensures that "within domain" spectral differences are unrelated to differences in q. We select the order q using the univariate time series of residuals for each simulation type, each domain, and each individual replicate. The order of the fit maximizes the log-predictive likelihood; further fitting details are available in Prado and West (2010). The highest order of distinctly non-zero coefficients, over all types of simulation, all replicates, and all reanalyses, is then used as the order for all univariate and multivariate autoregressive models in that spatial domain. In other words, we assume that within each region, the order q is the same for the replicates of each of the three model simulation types and the two reanalyses. The resulting posterior samples are used for the AR spectral analysis. Our tests of the statistical significance of model-versus-reanalysis spectral differences (see below) are robust with respect to the choice of model order q.

For coefficients $\boldsymbol{\phi}$ of an AR(q) process, the characteristic polynomial is given by $\Phi(u) = 1-\phi_1 u-\phi_2 u^2-\cdots-\phi_q u^q$. The polynomial can have r real-valued and c pairs of complex reciprocal roots such that $q = r+2c$. Although we do not necessarily expect complex roots, when present, they appear in pairs of complex conjugates and are interpretable as quasi-periodicities in the data. For each pair of complex roots written in terms of the modulus and frequency (ρ_j,ω_j), or equivalently the modulus and wavelength (ρ_j,λ_j) where $\lambda_j = 2\pi/\omega_j$ (months), for $j = 1,\ldots,c$. A modulus close to 1 indicates a slow decay rate in the correlation patterns, suggesting a persistent cyclical pattern occurring every wavelength λ_j months. More importantly, the autoregressive model allows for closed form calculation of the spectral density given estimates of the coefficients $\boldsymbol{\phi}$:

$$f(\omega) = \frac{\sigma^2}{2\pi|1-\phi_1 e^{-i\omega}-\cdots-\phi_q e^{-iq\omega}|^2}. \tag{9}$$

Here, $i = \sqrt{-1}$. Using the posterior samples of $\boldsymbol{\phi}$ for a given type of model simulation, the Bayesian approach provides a simple way to sample the corresponding spectral density. Normalizing the equation with respect to the white-noise spectrum, $\sigma^2/2\pi$, allows for the comparison of spectra solely with respect to differences in the AR coefficients. Further details of the autoregressive model, the quasi-periodicities, and the spectral densities can be found in Prado and West (2010).

3.5 Total variation distance for comparing internal variability

We use the total variation distance (TVD) for normalized spectral densities to quantify the differences between the spectral densities of the three types of climate model simulation and the reanalyses. This metric allows us to test whether the model and reanalysis time series have significantly different spectra. TVD was originally employed for comparing probability distributions, and has also been used to measure the similarity of normalized spectra in Euan et al. (2015) and Alvarez-Esteban et al. (2016). These applications rely on observational oceanographic data; they focus on classifying time series in the frequency domain and on detecting transitions and periods of stationary behavior. In order for TVD to be applicable to power spectra, normalization of

NCEP ERA Decadal Historical Control

	MIROC5 global	MIROC5 tropical	MIROC5 northern	MIROC5 southern
$(\delta_{base}^{mod}, \delta_{base}^{obs})$	(0.94,0.96)	(0.91,0.94)	(0.99,0.99)	(0.99, 0.99)
DLM MAP V	0.03, 0.02, 0.22, 0.18, 0.21	0.13, 0.12, 0.90, 0.65, 0.62	0.11, 0.11, 0.44, 0.39, 0.46	0.05, 0.04, 0.25, 0.20, 0.25
AR order q	4	7	5	5
AR MAP σ^2	0.011, 0.010, 0.046, 0.038, 0.066	0.006, 0.006, 0.032, 0.04, 0.032	0.030, 0.027, 0.126, 0.098, 0.179	0.018, 0.013, 0.053, 0.044, 0.079
maximum modulus	0.79*, 0.80*, 0.94*, 0.94*, 0.93*	0.87, 0.84, 0.92, 0.93, 0.90	0.67, 0.78*, 0.93*, 0.94*, 0.92*	0.72, 0.81*, 0.93*, 0.94*, 0.94*
complex root, max modulus, ρ	0.32, 0.40, 0.37, 0.44, 0.48	0.87, 0.84, 0.92, 0.93, 0.90	0.67, 0.42, 0.52, 0.53, 0.55	0.72, 0.47, 0.47, 0.52, 0.55
corresponding wavelength, λ	4.25, 4.38, 4.10, 4.24, 4.09	28.61, 28.63, 57.21, 56.03, 73.93	26.47, 3.47, 5.21, 4.98, 4.86	17.51, 4.82, 3.42, 3.66, 5.16

Table 1. Model baseline discount factor δ_{base}^{mod} and observation baseline discount factor δ_{base}^{obs}. DLM smoothed estimates of observational variance, V. AR order q and MAP of AR variance σ^2. Overall maximum modulus with * indicating correspondence to real roots, maximum complex modulus and corresponding wavelength (months) calculated from the AR MAP characteristic polynomial.

the spectral densities is first required; the integral of the normalized density must be equal to one. This is equivalent to normalizing the time series by dividing by its overall variance. The TVD of two normalized spectral densities $f^*(\omega) = f(\omega)/\int_\Omega f(\omega)d\omega$ and $g^*(\omega) = g(\omega)/\int_\Omega g(\omega)d\omega$ is defined as $\text{TVD}(f^*, g^*) = 1 - \int_\Omega \min\{f^*(\omega), g^*(\omega)\}d\omega$. For discrete normalized spectra, the TVD can equivalently be written in terms of the L_1 distance, $\text{TVD}(f^*, g^*) = ||f^* - g^*||_1/2 = \sum_{\omega\in\Omega} |f^*(\omega) - g^*(\omega)|/2$. The distance measure takes on values $0 \leq TVD \leq 1$, with 0 being the smallest possible discrepancy and 1 the largest.

Using the posterior spectra samples from the AR model, we can compute posterior distributions for the TVDs compared to a reference spectrum. In the first step of our analysis, we use a white-noise spectrum as a reference, and examine whether the residuals for the actual model temperature time time series are statistically distinguishable from this reference spectrum. Next, using the maximum *a posteriori* (MAP) NCEP-2 spectrum, we employ TVD to assess the significance of the discrepancies between the internal variability spectra of NCEP-2 and the three types of MIROC5 simulation. We also show TVD for the comparison between the NCEP-2 and ERA-I spectra, which provides a measure of the degree of difference we might expect due to uncertainties in the reanalysis-based estimates of temperature change. As noted above, these uncertainties arise from differences in reanalysis models, assimilation approaches, bias correction procedures for the assimilated data, *etc.*

4 Result from assessment of large-scale temperature

In this section, we first apply the previously described methodology to the time series of monthly-mean, spatially averaged near-surface temperature from the three sets of MIROC5 simulations. We then compare the model results to results obtained for the NCEP-2 and ERA-I reanalysis products. Note that the results are not meant to be compared directly between each region: the scales for our comparison metrics can vary markedly for the four spatial domains. Table 1 provides summaries of DLM and AR statistical model parameters and posterior inferences for each spatial domain. The table includes the baseline discount factors δ_{base}^{mod} and δ_{base}^{obs}, MAP DLM observation equation variance V, AR model order q, MAP AR variance σ^2, maximum moduli of all reciprocal roots from the AR characteristic polynomial based on the posterior means of the AR coefficients, maximum moduli of the reciprocal complex roots and corresponding wavelengths (months).

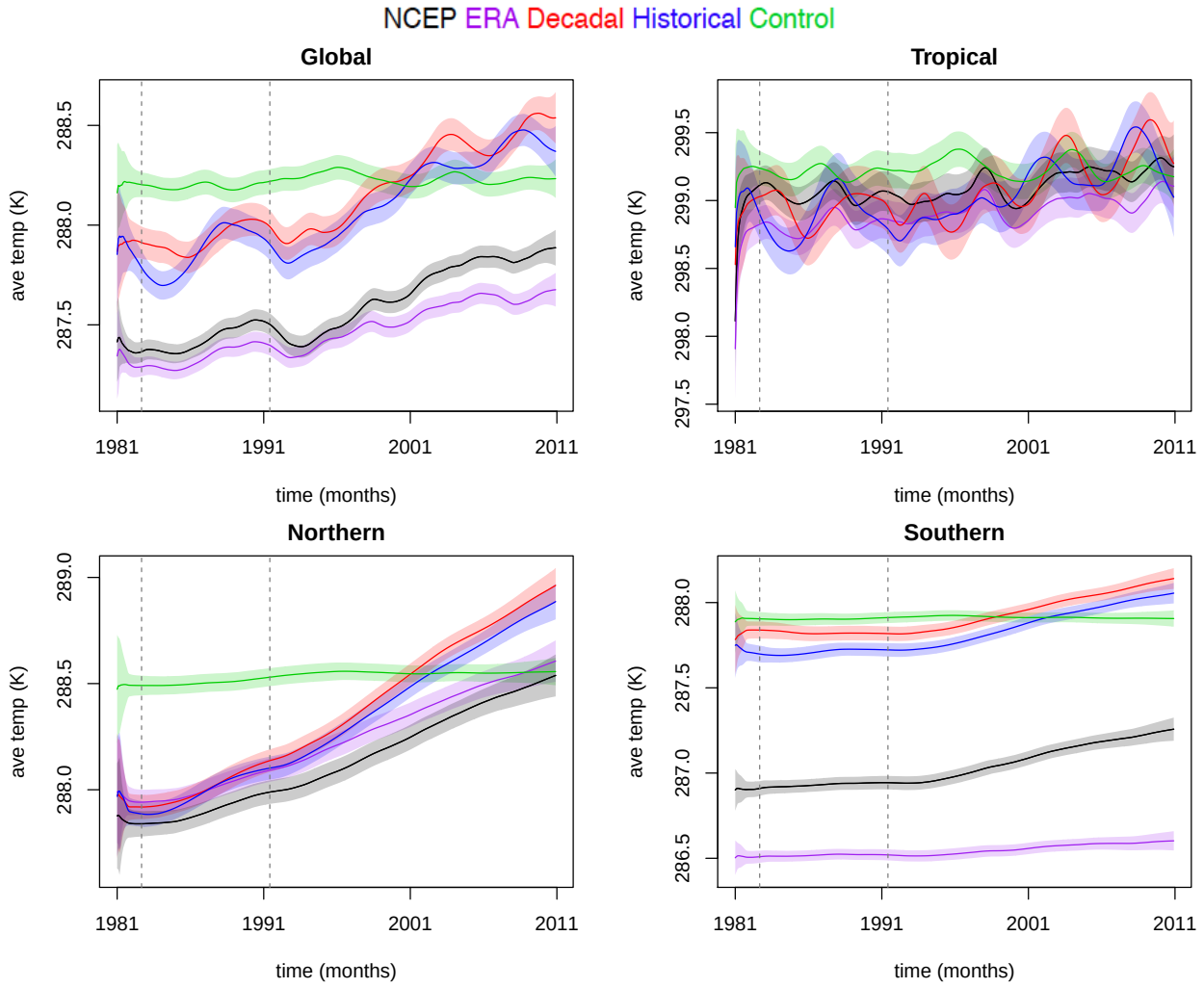

Figure 2. Baseline temperature estimates. Different line colors denote the type of simulation and the reanalysis product. Top: Global (left), tropical (right). Bottom: Northern (left), Southern Hemispheres (right). Vertical lines indicate the volcanic eruption of El Chichón in 1982 and Pinatubo in 1991.

Figure 2 displays the 95% posterior intervals for baselines $\eta_{1,t}$, estimated using the DLM model introduced in Section 3.1. The control run baseline estimates are noticeably flat relative to the baselines inferred for the other types of simulation and for the reanalysis products. This difference is expected – the control run lacks year-to-year changes in external forcings, while the reanalysis products and the historical and decadal prediction runs are affected by time-varying anthropogenic and natural forcings. This explains why, in each of the four spatial domains we considered, the baselines in the externally forced runs and the reanalyses show secular temperature increases over 1981 to 2010, consistent with warming of the Earth's surface in response to time-increasing net anthropogenic forcing.

Note that the surface temperature baseline has a larger overall trend in the historical and decadal prediction runs than in NCEP-2 and ERA-I. This discrepancy in simulated and observed warming rates is at least partly related to the omission of the observed early 21st century increase in stratospheric volcanic aerosols in the model historical and decadal prediction simulations (Solomon et al., 2011). In the real world, the cooling caused by this post-2000 increase in stratospheric volcanic aerosols offset part of the anthropogenic warming signal (Schmidt et al., 2014; Santer et al., 2014). Superimposed on these long-term warming trends in the reanalyses and the decadal prediction and historical runs are short-term (1-2 year) surface cooling signals associated with the major eruptions of El Chichón in 1982 and Pinatubo in 1991 (Santer et al., 2001). Because averaging over larger domains damps spatial noise, volcanic cooling signals are most pronounced in the global-spatial average, and are noisiest in the smaller-scale tropical averages.

The surface cooling signals caused by El Chichón and Pinatubo are markedly smaller in the Northern and Southern Hemisphere averages than for the global domain. This is the result of the Northern and Southern Hemispheres baselines being estimated with discount factors close to 1 (see Table 1). It is not unexpected for the hemisphere-specific externally-forced components to be less variable than the spatial domains which contain interaction between the distinct hemispheric seasonal cycles. Selection of high discount factors suggests that the externally-forced longer-term variability in both hemispheres was very close to linear. Any shorter-term forced variability not captured by the baselines will be reflected in the residuals. For both the Northern Hemisphere and Southern Hemisphere, the residuals do not exhibit substantial cooling after the volcanic eruptions of El Chichón in 1982 and Pinatubo in 1991. Further investigation of the differences in the amplitude of the global-average and hemispheric-average volcanic signals may be of interest. Alternately, the baseline temperatures for the tropical region are estimated with much lower discount factors (see Table 1), indicating the externally-forced longer-term variability in the tropics was more variable.

Figure 2 also yields many other features of interest, such as differences in mean temperature in 1981. Because the decadal prediction runs are initialized from observed ocean temperature and sea ice data, it is not unreasonable to expect that at the time of initialization in 1981, the mean surface temperature in these simulations should be close to the mean temperature of the two reanalysis products. This is the case for the Northern Hemisphere and tropical averages, but not for the averages over the other three regions. The largest mean state differences in 1981 are in the Southern Hemisphere, where the decadal prediction runs are noticeably warmer than either reanalysis. Because this Southern Hemisphere bias is large, it also influences the global temperature average.

One possible interpretation of this large SH bias is that it may arise because of differences between the observed sea surface temperature (SST) data sets used as boundary conditions for the two reanalyses and the surface temperature data selected for initialization of the MIROC5 decadal prediction runs. Observational SST uncertainties are likely greater in the more poorly sampled Southern Hemisphere than in the Northern Hemisphere – which may explain why the 1981 warm bias in the decadal prediction runs is largest in the SH.

Use of different observational SST data sets may also explain why the two reanalyses show the largest mean state differences in the Southern Hemisphere. An alternative (and not mutually exclusive) interpretation is that the "between-reanalysis" mean state differences reflect the sparser observational coverage in the Southern Hemisphere, and a larger SH imprint of structural

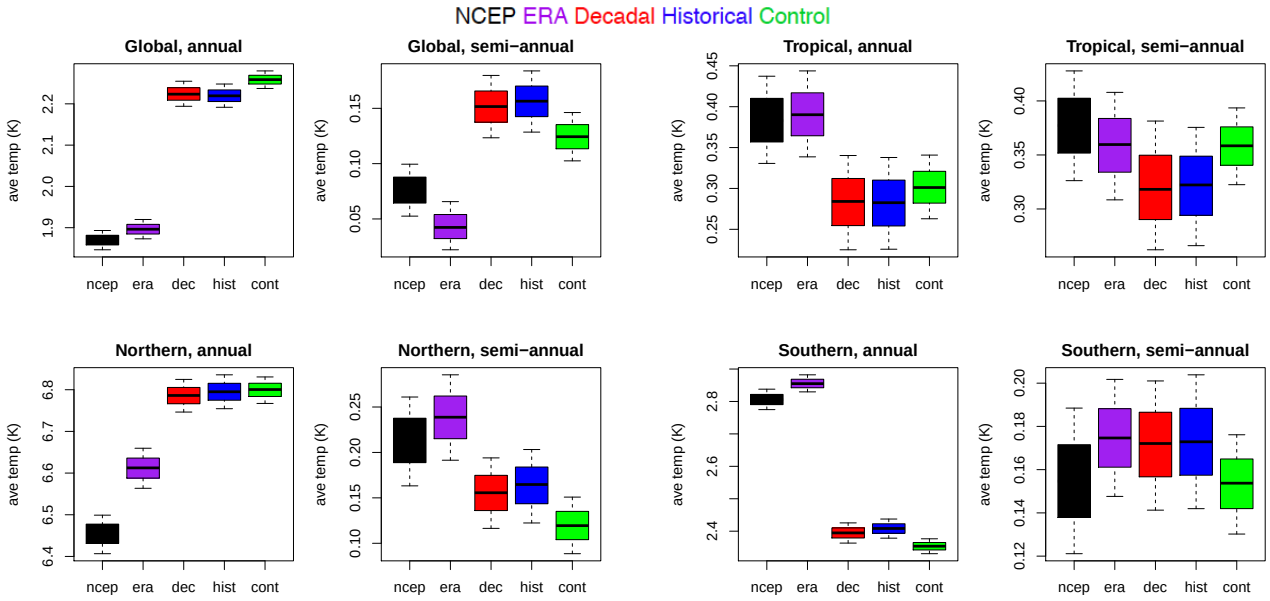

Figure 3. Posterior amplitude samples for harmonics $k = 1, 2$. Whiskers indicate the maximum and minimum values, boxes indicate 95% posterior intervals. Different colors denote the type of simulation and the reanalysis product. Top: Global, tropical. Bottom: Northern, Southern hemispheres.

differences between the NCEP-2 and ERA-Interim forecast models (e.g., in terms of physics, parameterizations, resolution, and data assimilation systems).

Note that the model-versus-reanalysis warm biases mentioned above do not only pertain to the decadal prediction runs – they also affect the historical and control simulations. In all four spatial domains considered, the model-generated baseline temperatures are consistently warmer than in either reanalysis product. The baseline temperatures in the decadal prediction integrations do not appear to exhibit appreciable post-initialization secular drift, and are similar to the baseline temperatures in the historical runs. This implies that our DLM model is primarily capturing the externally forced component of surface temperature changes in MIROC5, and that the amplitude and structure of this forced response is relatively insensitive to whether the simulation is "free running" or initialized from observations.

Figure 3 illustrates the 95% posterior intervals of the seasonal amplitudes $\alpha_{1,t}^{k}$ for $k = 1, 2$ (i.e., for the amplitudes of the annual and semi-annual cycles, respectively). Amplitudes were estimated using the DLM model in Section 3.1. For all four spatial domains, the harmonics $k = 3$ and $k = 4$ (corresponding to the trimestral and quarterly cycles, respectively) are very close to 0 and indistinguishable from one another, and are therefore not shown. Results for the annual and semi-annual cycles are more interesting. Consider the reanalyses first. For all four spatial domains, and for both for $k = 1$ and $k = 2$, NCEP-2 and ERA-I yield very similar amplitudes. The only significant difference between the two reanalyses is in the Northern Hemisphere, where the ERA-I annual cycle amplitude is markedly higher than in NCEP-2.

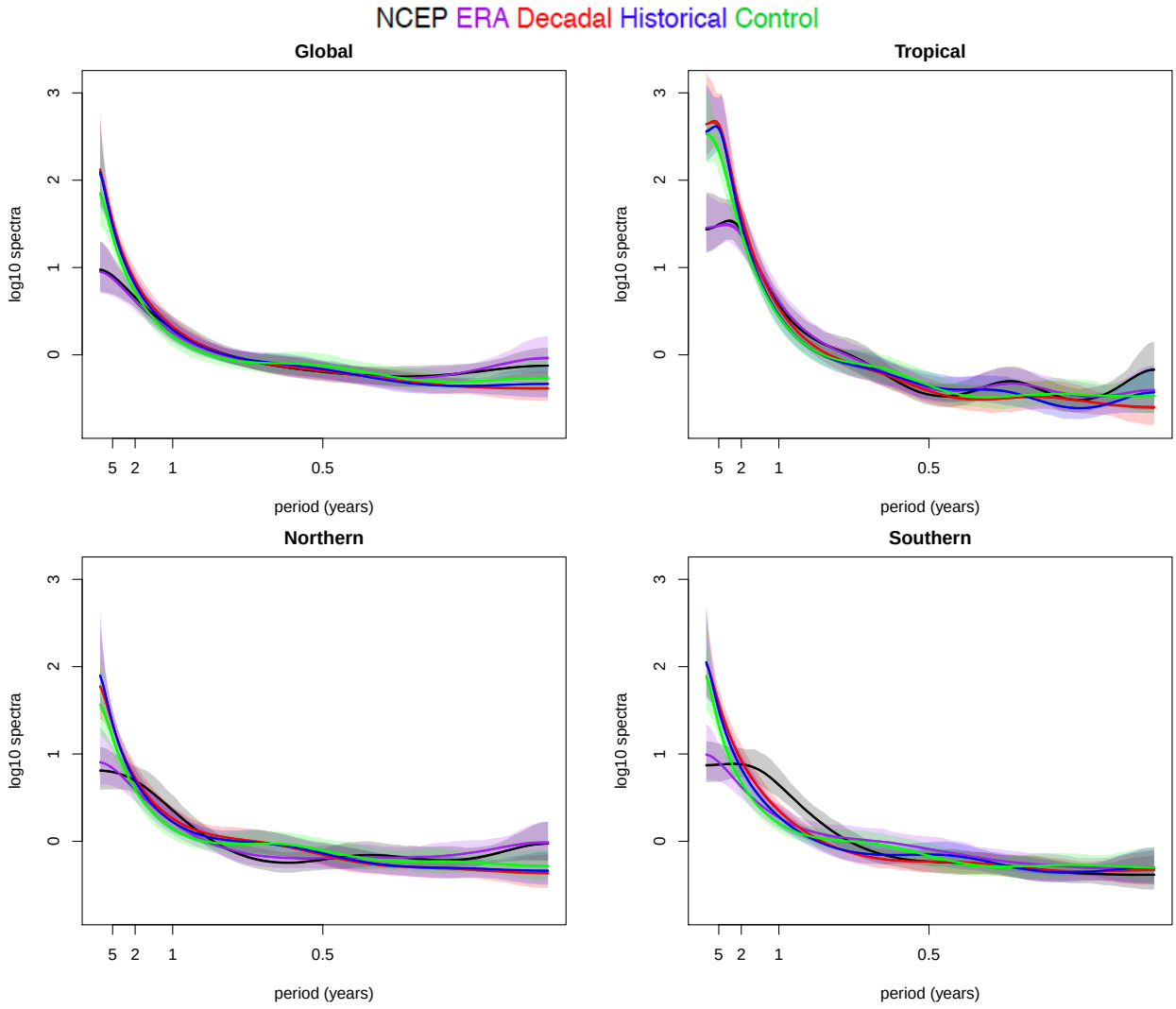

Figure 4. MAP AR $\log_{10}$ spectra normalized with respect to white-noise for each climate simulation by region with 95% posterior intervals shaded. Instead of the frequency ω, the x-axis is labeled at select years ($2\pi/12\omega$). Different line colors denote the type of simulation and the reanalysis product. Top: Global, tropical. Bottom: Northern, Southern hemispheres.

For all spatial domains except the tropics, and for all three types of simulation, the MIROC5 annual cycle amplitudes differ significantly from those in either reanalysis product. Model-versus-reanalysis differences in annual cycle amplitude are most pronounced in the Southern Hemisphere. The sign of the model annual cycle biases is not consistent across domains. In the tropics and SH, the annual cycle amplitude is smaller in the simulations than in the reanalyses. In the other two domains, however, the annual cycle amplitude is larger in the simulations than in NCEP-2 and ERA-I. We do not find any cases in which there are significant amplitude differences between the three types of model simulation.

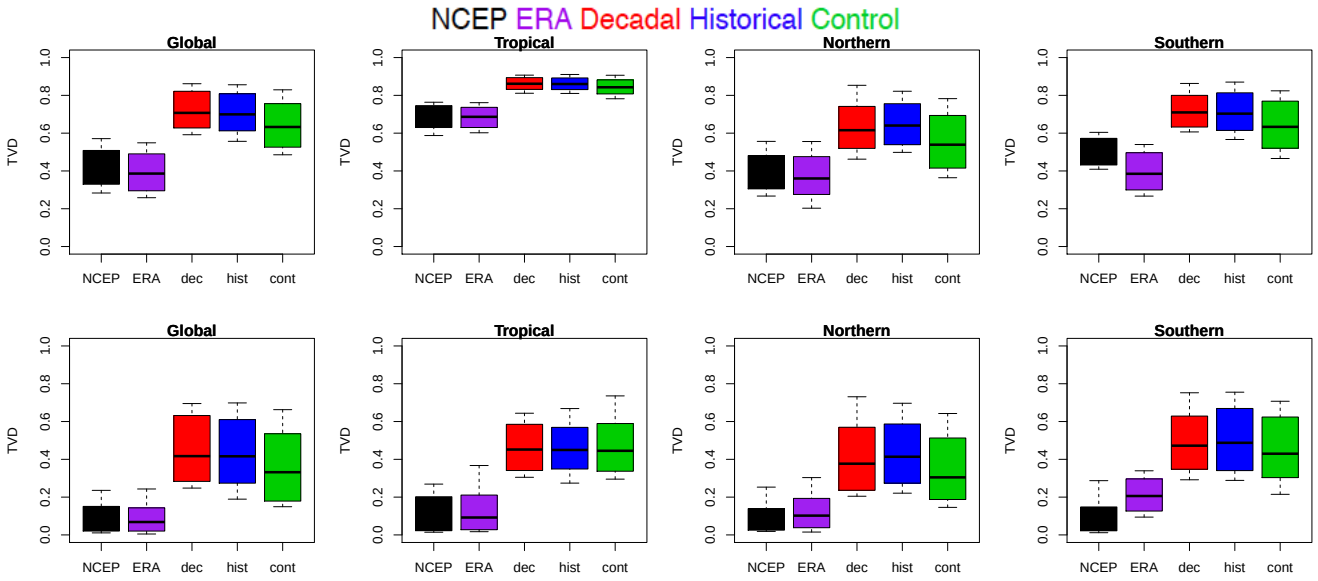

Figure 5. Top: TVD calculated from ϕ samples with white-noise as the reference. Bottom: TVD calculated from ϕ samples with the MAP NCEP spectrum as the reference. Whiskers indicate the maximum and minimum values, boxes indicate 95% posterior intervals. Scenario or observational product indicated by colors. Left to right: Global, tropical, northern, and southern hemispheres.

Figure 4 illustrates the 95% confidence intervals on the posterior spectra, normalized with respect to white noise on the log-scale. Spectra were estimated using the methods presented in Section 3.4. The spectral densities are relatively smoothly varying as a function of frequency, particularly for spectra generated with lower-order AR models (e.g., the $q = 4$ case for the global region; see Table 1). The least-smooth spectra are obtained for temperatures spatially averaged over the tropics, where a higher-order AR model ($q = 7$) provides the best fit to the residuals remaining after removal of the baseline and seasonal temperature components. This result is physically reasonable: the tropical domain is the smallest and "noisiest" of the four domains considered here, and is strongly influenced by modes of internal variability acting on a range of different timescales, such as the Madden-Julian Oscillation, ENSO, and the Interdecadal Pacific Oscillation.

Other features of Figure 4 are also noteworthy. First, the spectra for the three different types of MIROC5 simulation are very similar. This suggests that the DLM method applied here has successfully partitioned the "pure" internally generated component of surface temperature from: 1) the externally forced components of temperature changes in the historical runs; and 2) the combined effects of external forcing and any post-initialization drift in the decadal prediction simulations. Second, at the lowest frequencies, model spectral densities are higher than in NCEP-2 and ERA-I, and the 95% posterior intervals of nearly all of the simulated spectra do not overlap with the reanalysis spectra. This difference in the amplitude of simulated and observed variability (which is most pronounced in the tropics) is consistent with findings obtained elsewhere for multi-model analyses of tropospheric temperature (Santer et al., 2013, 2018; Pallotta and Santer, in preparation). A model bias in the opposite direction to that found here (i.e., a systematic underestimate of the amplitude of observed internal variability on multi-

decadal timescales) would be more concerning – such an error would spuriously inflate signal-to-noise ratios for anthropogenic signal detection (Santer et al., 2013, 2018). We caution, however, that the reanalysis data analyzed here are relatively short (30 years), and do not provide a strong constraint on "observed" estimates of internal variability on multi-decadal timescales.

Recall from Section 3.4 that the presence of complex roots points towards the existence of quasi-cyclical temperature variations. The results in the fourth row of Table 1 indicate that complex roots are only obtained consistently for the tropical domain. For all other domains, the characteristic polynomials from the AR models are dominated by real roots. This suggests that the tropics – which are strongly affected by the El Niño/Southern Oscillation – are capturing some quasi-periodic temperature variability associated with the occurrence of El Niños and La Niñas. Confirmation of this quasi-periodicity comes from the fact that the tropics are also the only domain where the maximum moduli of the reciprocal complex roots of the polynomials exceed 0.8 for both reanalyses and for all three types of simulation (see results in fifth row of Table 1). The wavelengths for the tropical quasi-periodic variability are approximately 28.6 months (2.38 years) for the reanalysis products, 57.2 months (4.77 years) for the decadal prediction run, 56 months (4.67 years) for the historical simulation, and 73.9 months (6.16 years) for the control run. The apparent absence of quasi-periodic behavior on longer timescales is probably (at least in part) a reflection of the relatively short record lengths considered here (see above).

Finally, we present results for the total variation distance (TVD), which allows us to make a quantitative evaluation of the differences between the various spectra. The posterior distributions of the TVD are given in Figure 5. The top row shows results for the comparison against a white noise reference spectrum. All reanalysis and model data sets are statistically separable from white noise. For each of the four domains, the reanalysis data sets have smaller TVD values, and are closest to the white noise case; the three sets of simulations are further removed from the white noise reference spectrum. The systematically lower TVD values for NCEP-2 and ERA-I may partly reflect the fact that both reanalyses exhibited decadal temperature variability that was consistently smaller than in the MIROC5 simulations (see above). The largest TVD values for the reanalyses and the model simulations are in the tropics, indicating that tropical temperature variability is most clearly distinguishable from white noise. This is consistent with the above-mentioned finding that the discrepancy between low-frequency temperature variability in the reanalyses and the MIROC5 simulations is largest in the tropics.

The bottom row of Figure 5 displays results for the comparison between the model spectra and the NCEP MAP spectrum. The range of TVD values for the NCEP spectrum versus itself is simply a reflection of posterior sampling variability. The global and tropical regions show distinct differences between the reanalysis products and the three sets of simulations, with little or no overlap between the 95th percentiles of the reanalyses and the 5th percentiles of the simulations. The tropical region exhibits the most significant difference between the NCEP spectrum and the simulated spectra; this is likely due to the above-mentioned discrepancies in low-frequency variance. It may also reflect the fact that the identified quasi-periodic component of tropical temperature variability had a longer timescale in the three sets of simulations than in the reanalysis products.

5 Conclusions

We applied univariate and multivariate Dynamic Linear Modeling (DLM) techniques to estimate two externally forced components of surface temperature time series. These components contain: 1) seasonal information, which is invariant from year-to-year; and 2) the time-varying nonlinear response to combined external forcing by human factors (such as greenhouse gases and particulate pollution) and natural influences (changes in solar irradiance and volcanic activity). Estimation of these two temperature components – which we refer to as "seasonal" and "baseline", respectively – was performed for two reanalysis data sets and for three different types of experiment performed with one selected climate model (MIROC5). The three sets of numerical experiments were initialized decadal predictions, control runs, and uninitialized simulations of historical climate change. Removal of the seasonal and baseline components from the raw temperature data yielded residuals that provided information on unforced natural internal climate variability. We characterized this internal variability by fitting univariate and multivariate autoregressive (AR) models to the residuals. Since estimates of externally forced climate signals and internal variability depend on the particular domain of interest, we explored the efficacy of our DLM and AR signal and noise identification methods for four different spatial domains, ranging in scale from the entire globe to the tropics.

We found significant differences between the reanalysis data and the model-generated simulations in all three temperature components (seasonal, baseline, and internal variability). From a climate perspective, two results were particularly intriguing. First, we note that the three sets of simulations analyzed here are very different. While temperature variability in the control run arises from internal variability alone, variability in the historical and decadal prediction runs is a mixture of internal variability and response to external forcing. Additionally, the decadal prediction runs may also be influenced by post-initialization "drift" in the model climate. Despite these differences in the mix of underlying factors contributing to variability, the three types of simulation yielded very similar spectral estimates of internal temperature variability – as might be expected given that the same physical climate model is being used for each of the three sets of simulations. This similarity of the three sets of model spectra is reassuring, and implies that our statistical analysis methods are reliably extracting the common underlying component of internal variability.

The second intriguing result emerged from the comparison of the model and reanalysis temperature variability on multi-decadal timescales. These timescales are important components of the background "noise" against which a gradually evolving anthropogenic warming signal must be detected. If models systematically underestimated natural internal variability on multi-decadal timescales, it would imply that previously obtained anthropogenic signal detection results were spuriously inflated by low model noise levels. Consistent with related work involving tropospheric temperature ((Santer et al., 2013, 2018; Pallotta and Santer, in preparation)), we find that the MIROC5 model overestimates the amplitude of low-frequency internal variability inferred from reanalysis data. This suggests that results from previous anthropogenic signal identification studies may have been too conservative. It would be of interest to apply our statistical techniques to observational estimates of surface temperature that are longer than the 30-year reanalysis records available here. The ultimate goal is to obtain stronger constraints on the amplitude of observed multi-decadal temperature variability, thereby providing a more solid observational "target" for model evaluation purposes.

Data availability. The data used in Section 4 for MIROC5, NCEP Reanalysis 2, and ERA-Interim are available respectively from: https://esgf-node.llnl.gov/search/cmip5/, https://www.esrl.noaa.gov/psd/, and https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim.

Acknowledgements. The authors wish to thank Ana Kupresanin and Francisco Beltran at LLNL for helpful conversation. Sansó was partially funded by the National Science Foundation grant DMS-1513076. Prado was partially funded by the National Science Foundation grant SES-1461497.

References

Alvarez-Esteban, P. C., Euán, C., and Ortega, J.: Time series clustering using the total variation distance with applications in oceanography, Environmetrics, 27, 355–369, https://doi.org/10.1002/env.2398, 2016.

Berrisford, P., Kållberg, P., Kobayashi, S., Dee, D., Uppala, S., Simmons, A., Poli, P., and Sato, H.: Atmospheric conservation properties in ERA-Interim, Quarterly Journal of the Royal Meteorological Society, 137, 1381–1399, 2011.

Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distribution, Bull. Calcutta Math. Soc, 1943.

Boer, G. J.: Decadal potential predictability of twenty-first century climate, Climate dynamics, 36, 1119–1133, 2011.

Collins, M., Botzet, M., Carril, A., Drange, H., Jouzeau, A., Latif, M., Masina, S., Otteraa, O., Pohlmann, H., Sorteberg, A., et al.: Interannual to decadal climate predictability in the North Atlantic: a multimodel-ensemble study, Journal of climate, 19, 1195–1203, 2006.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., et al.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the royal meteorological society, 137, 553–597, 2011.

Derpanis, K. G.: The bhattacharyya measure, Mendeley Computer, 1, 1990–1992, 2008.

Euan, C., Ombao, H., and Ortega, J.: Spectral synchronicity in brain signals, arXiv preprint arXiv:1507.05018, 2015.

Fujiwara, M., Wright, J. S., Manney, G. L., Gray, L. J., Anstey, J., Birner, T., Davis, S., Gerber, E. P., Harvey, V. L., Hegglin, M. I., et al.: Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems, Atmospheric Chemistry and Physics, 17, 1417–1452, 2017.

Harrison, J. and West, M.: Bayesian forecasting & dynamic models, vol. 1030, Springer New York City, 1999.

Imbers, J., Lopez, A., Huntingford, C., and Allen, M.: Sensitivity of climate change detection and attribution to the characterization of internal climate variability, Journal of Climate, 27, 3477–3491, 2014.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al.: The NCEP/NCAR 40-year reanalysis project, Bulletin of the American meteorological Society, 77, 437–471, 1996.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J., Fiorino, M., and Potter, G.: Ncep–doe amip-ii reanalysis (r-2), Bulletin of the American Meteorological Society, 83, 1631–1643, 2002.

Kirtman, B., Power, S., Adedoyin, A., Boer, G., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A., Kimoto, M., Meehl, G., Prather, M., Sarr, A., Schar, C., Sutton, R., Oldenborgh, G., Vecchi, G., and Wang, H.: Near-term climate change: projections and predictability, In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 2013.

Kopp, G. and Lean, J. L.: A new, lower value of total solar irradiance: Evidence and climate significance, Geophysical Research Letters, 38, 2011.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J.: The coupled model intercomparison project (CMIP), Bulletin of the American Meteorological Society, 81, 313–318, 2000.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., et al.: Decadal prediction: can it be skillful?, Bulletin of the American Meteorological Society, 90, 1467–1485, 2009.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti, S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., et al.: Decadal climate prediction: an update from the trenches, Bulletin of the American Meteorological Society, 95, 243–267, 2014.

Pallotta, G. and Santer, B. D.: Estimating the spectral features of modeled and observed tropospheric temperature, in preparation.

Prado, R. and West, M.: Time series: modeling, computation, and inference, CRC Press, 2010.

Santer, B. D., Wigley, T., Doutriaux, C., Boyle, J., Hansen, J., Jones, P., Meehl, G., Roeckner, E., Sengupta, S., and Taylor, K.: Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends, Journal of Geophysical Research: Atmospheres, 106, 28 033–28 059, 2001.

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

Santer, B. D., Po-Chedley, S., Zelinka, M. D., Cvijanovic, I., Bonfils, C., Durack, P. J., Fu, Q., Kiehl, J., Mears, C., Painter, J., Pallotta, G., Solomon, S., Wentz, F. J., and Zou, C.-Z.: Human influence on the seasonal cycle of tropospheric temperature, Science, 361, 2018.

Schmidt, G., Shindell, D., and Tsigaridis, K.: Reconciling warming trends, Nat. Geosci., 7, 158–160, 2014.

Taylor, K. E.: A summary of the CMIP5 experiment design, http://cmip-pcmdi. llnl. gov/cmip5/docs/Taylor_CMIP5_design. pdf, 2009.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American Meteorological Society, 93, 485–498, 2012.

Watanabe, M., Suzuki, T., Oishi, R., Komuro, Y., Watanabe, S., Emori, S., Takemura, T., Chikira, M., Ogura, T., Sekiguchi, M., et al.: Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity, Journal of Climate, 23, 6312–6335, 2010.