Adv. Stat. Clim. Meteorol. Oceanogr., 5, 67–85, 2019
https://doi.org/10.5194/ascmo-5-67-2019

ASCMO
Open Access

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

Raquel Barata, Raquel Prado, and Bruno Sansó

Department of Statistics, University of California Santa Cruz, Santa Cruz, CA 95064, USA

Correspondence: Raquel Barata (rbarata@ucsc.edu)

Received: 13 August 2018 – Revised: 29 April 2019 – Accepted: 3 May 2019 – Published: 27 May 2019

Abstract. We present a data-driven approach to assess and compare the behavior of large-scale spatial averages of surface temperature in climate model simulations and in observational products. We rely on univariate and multivariate dynamic linear model (DLM) techniques to estimate both long-term and seasonal changes in temperature. The residuals from the DLM analyses capture the internal variability of the climate system and exhibit complex temporal autocorrelation structure. To characterize this internal variability, we explore the structure of these residuals using univariate and multivariate autoregressive (AR) models. As a proof of concept that can easily be extended to other climate models, we apply our approach to one particular climate model (MIROC5). Our results illustrate model versus data differences in both long-term and seasonal changes in temperature. Despite differences in the underlying factors contributing to variability, the different types of simulation yield very similar spectral estimates of internal temperature variability. In general, we find that 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 regarding efforts to identify anthropogenic "fingerprints" in observational surface temperature data. Our methodology and results present a novel approach to obtaining data-driven estimates of climate variability for purposes of model evaluation.

1 Introduction

Exploring the impacts of anthropogenic climate change is of great relevance and interest to society. Phase 5 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-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).

In this paper we present a statistical model-based approach to compare the observational record with three different types of CMIP5 simulations. We seek to develop a consistently principled way of determining whether there are statistically significant differences in aspects of the variability, both between the model simulations and the observational record, and within three different types of simulation. Our primary goal is to illustrate the utility of Bayesian statistical techniques that are not in widespread use in climate science. We develop a data-driven model diagnostic method and protocol with potential application to future CMIP model evaluation.

Many attempts have been made to characterize the climate response to external and internal forcing simulated in the CMIP5 experiments. Simulated internal variability is of particular interested as there is evidence that it influences trends in regional temperatures making model evaluation a challenge (Kay et al., 2015; Gibson et al., 2017; Perkins-Kirkpatrick et al., 2017). Separating the externally and internally forced components is a non-trivial challenge, and many current studies rely on ad hoc approaches. The methods presented here have two main advantages: the protocol we develop to jointly estimate the components of the tem-

perature time series is statistically principled and consistent across different series, and the results incorporate probabilistic uncertainty as the approach is model-based.

To illustrate our statistical methods, we focus on one specific climate model: version 5 of the atmosphere–ocean general circulation model (AOGCM), which has been jointly developed 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 (see Watanabe et al., 2010). From this model, commonly referred to as the Model for Interdisciplinary Research on Climate (MIROC5), we examine three different types of 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. Our analysis focuses on monthly mean 2 m surface temperature time series over four regions: global, tropical, Northern Hemisphere and Southern Hemisphere. 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 statistical analysis of MIROC5 simulations can be easily extended to other models in the CMIP archive.

Our approach extracts long-term externally forced changes in temperature, seasonality and estimates of internal variability of the climate model simulations and compares them to the corresponding components in observational products. As in Imbers et al. (2014), our focus is on investigating the spectral characteristics of internal variability. 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 Solomon et al., 2011; Schmidt et al., 2014). Additionally, we investigate whether there are identifiable differences between the spectral properties in the decadal prediction, historical, and control simulations that are related to factors such as the inclusion of external forcings and the initialization approach.

The paper is organized as follows. In Sect. 2, we describe the model simulations and the observational products analyzed here. Section 3 presents our statistical modeling approach and introduces the DLM used to estimate the baseline and seasonal components of the time series. Section 3 also describes the AR model that we apply to the residual time series in order to estimate natural internal variability. In Sect. 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 in a short-term analysis (30 years). Section 5 presents results obtained similarly for a long-term analysis (63 years). Section 6 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. As this study is primarily an exposition of methodology rather than a comprehensive analysis of internal variability behavior in CMIP5 models, we focus here on simulations performed with one particular state-of-the-art climate model (MIROC5). 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 anthropogenic changes in well-mixed greenhouse gases and natural changes in volcanic aerosols (Kirtman et al., 2013). For a full description of the characteristics of the different CMIP5 simulations, see Van Vuuren et al. (2011). The forced simulations also reflect the 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 Modeling (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 volcanic eruption or 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 short-term analysis. It also explains the absence of decadal simulations in our long-term analysis, which spans the period from 1950 to 2012. 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 the initialization methods and the organization of the decadal prediction simulations can be found in Meehl et al. (2009).

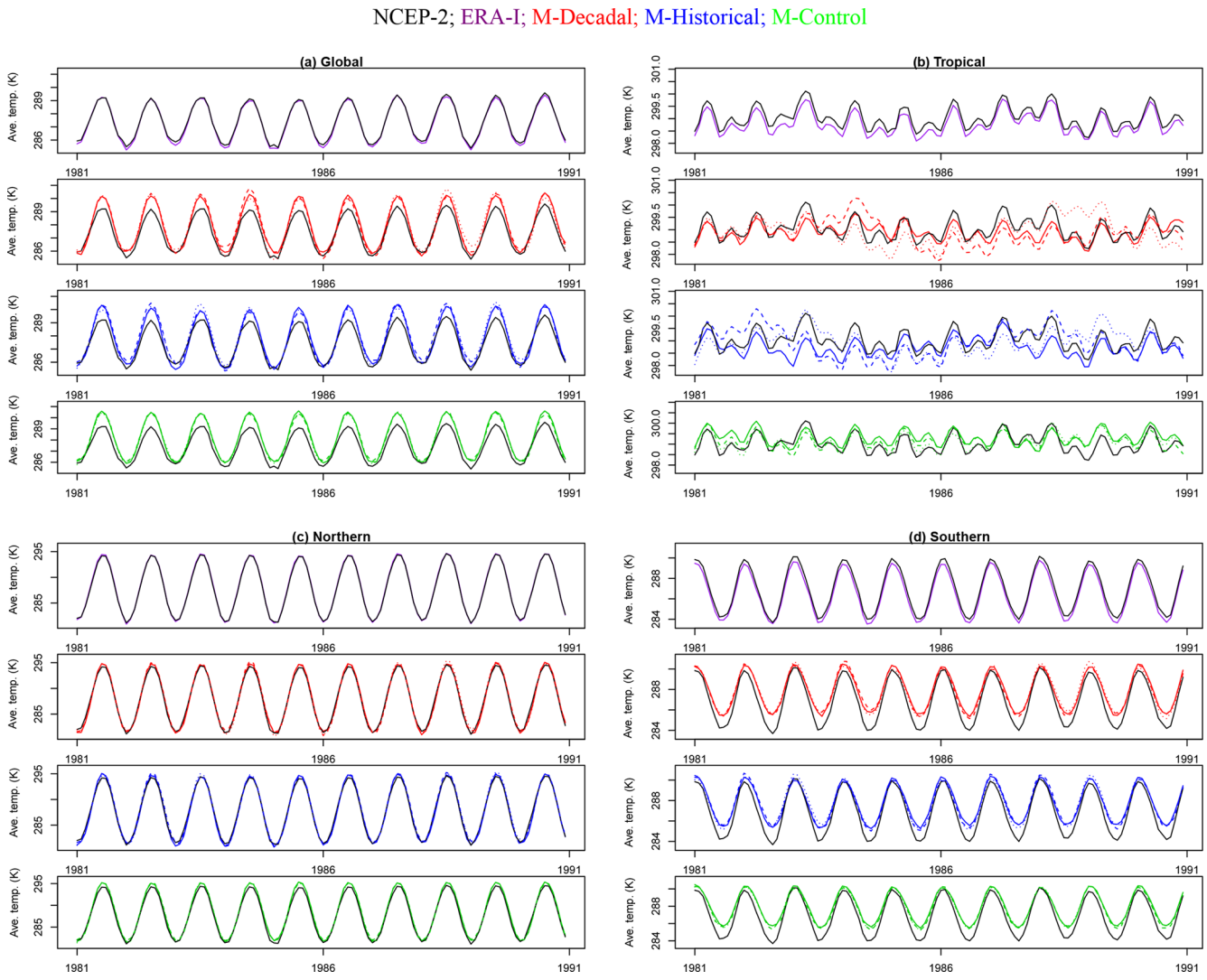

Figure 1. The 10-year series of monthly mean, spatially averaged 2 m surface temperature from the MIROC5 model. The top plots in each panel correspond to NCEP-2 and ERA-Interim reanalyses. The three remaining plots in each panel correspond to the three different types of simulations, in vertical descending order, decadal, historical, and control (labeled "M-" for model in the header). Three realizations of each model simulation are indicated by different line types. The top panels represent global **(a)** and tropical **(b)** regions; the bottom panels represent the Northern Hemisphere **(c)** and Southern Hemisphere **(d)**.

Six individual realizations of the MIROC5 decadal prediction run were available (see Fig. 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 realizations were available.

The longest period of overlap between the MIROC5 decadal and historical runs is from January 1981 to December 2010 (see Fig. 1). This is the period of our short-term analysis. Our long-term analysis over January 1950 to December 2012 allows us to explore the sensitivity of model versus data comparisons to the use of a longer record (and hence provides a stronger observational constraint on decadal variability).

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 pre-

diction 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. Therefore, the two types of simulation can 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).

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 the 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. To create compatibility with the record length of the data available for other simulation runs, we extract 10 nonoverlapping monthly mean temperature time series from the 670-year MIROC5 control run. In the short-term analysis, we extract ten 30-year time series. For the long-term analysis, ten 63-year time series are extracted from the control run. Each 30-year (or 63-year) segment contains a different unique manifestation of internal variability, so they are similar to the realizations available for the decadal prediction and historical runs and we regard them as such (see Fig. 1).

Several points should be emphasized prior to the discussion of the model results. First, the AOGCM 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, imparting 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 observational record, 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 30-year period of interest (January 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 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 Observational records

We compare the climate model simulations to reanalysis observational products and to one in situ observational record. 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 observational products for the short-term analysis. The first is version 2 of the reanalysis performed by the National Centers for Environmental Prediction (NCEP), subsequently referred to 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 subsequently refer to this reanalysis as ERA-Interim (ERA-I). It begins in 1979 and is continuously updated. Results from both reanalyses are shown in Fig. 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).

We rely on two observational records for our long-term analysis. The first is the ensemble mean of the 20th Century Reanalysis Version 2 (20CRV2) observational product. This reanalysis was jointly performed by the National Oceanic and Atmospheric Administration (NOAA) and the Cooperative Institute for Research in Environmental Sciences (CIRES) at the University of Colorado. The 20CRV2 reanalysis does not assimilate any upper-air information from radiosondes and satellites – it only incorporates surface observations of synoptic pressure, monthly sea surface temperature, and sea ice distribution (Compo et al., 2011). Inclusion of a reanalysis that has fewer sources of input data in our analysis potentially provides a more homogeneous surface temperature record. Further details on the 20CRV2 reanalysis can be found in Compo et al. (2011). The second observational record in our long-term analysis is the in situ observational record from the Berkeley Earth Surface Temperature project (BEST). Gridded monthly mean temperature fields were generated using an averaging process described in Rohde et al. (2013). Results are for land and ocean temperature and for air temperature over sea ice. BEST data are available from 1850 to the present, although we only consider data from January 1950 until December 2012. Note that observational products before the satellite era, which began around 1979, are subject to regions with limited in situ data. The choice of time period for the long-term analysis is motivated by surface temperature coverage being degraded in the first half of the century and problems with sea surface temperature (SST) measurements at the time of the Second World War. Further details on the data and BEST averaging process can be found in Rohde et al. (2013).

All model and observational surface temperature data are available in gridded form for a global domain. We calculate area-weighted spatial averages over four regions: the globe (90° S to 90° N), the tropics (20° S to 20° N), the Northern Hemisphere (0 to 90° N), and the Southern Hemisphere (90° S to 0°). As an example of the data considered here, we show the first 10 years of the three different types of simulation analyzed in Fig. 1. Globally averaged temperature exhibits 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 semiannual cycle is apparent in the tropics (Santer et al., 2018). Our statistical analyses focus on these area-averaged time series.

3 Statistical models for model-generated and reanalysis time series

The protocol presented in this section involves decomposing each temperature time series into what we refer to as a "baseline" and a seasonal component using dynamic linear models (DLMs). The baseline component aims to capture long-term externally forced changes in temperature, whereas the seasonal component is dominated by the externally forced annual and semiannual cycles. We extract the baseline temperature and seasonality of the climate model simulations and seek to compare them to the corresponding components in the observational products. The DLM residuals, which have the baseline and seasonal components of the temperature removed, are time series that primarily represent natural internal climate variability. Using autoregressive (AR) models, we investigate the spectral characteristics of internal variability both between the simulations and observational products, as well as within the simulated experiments.

DLMs are a popular Bayesian modeling approach for the analysis of nonstationary time series. We follow approaches detailed in West and Harrison (1999) and Prado and West (2010) in order to estimate time-varying baseline and seasonality components. One of the main advantages of using DLMs is that they naturally deal with nonstationary data and allow us to extract the baseline and seasonality components jointly, while quantifying the uncertainty associated with each component. In Sect. 3.1, we present the DLM used here to extract the baseline and seasonal components from the spatially averaged model and observational 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 realizations of the model simulations. Section 3.3 presents our DLM discount factor selection strategy. Section 3.4 describes a Bayesian approach to fitting autoregressive models to investigate the DLM residuals and their spectral properties. Section 3.5 presents the results of using the total variation distance (TVD) as a method to compare the spectra of the DLM residuals. Section 3.6 concludes with a summary of our complete data analysis protocol.

3.1 Baseline and seasonal temperature estimation

Consider first a single observational product 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 is 360 months for the short-term analysis (30 years) and 756 months for the long-term analysis (63 years). We decompose each time series into a baseline temperature $\eta_{1,t}$ and seasonal components $\alpha^k_{1,t}$ for harmonics $k = 1, \ldots, K$ of a fundamental period p. Let $N_d(\mathbf{m}, \mathbf{S})$ de-

note 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.,

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

where V is the unknown level one variance (West and Harrison, 1999). For convenience it is assumed that the level one errors ν_t are independent in time, although subsequently we model the dependence of the residuals. We further assume that the baseline component has a structure described by

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

Here the system evolution error vectors $\boldsymbol{\omega}_t^{\eta}$, or level two errors, 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 include harmonics $1, .., K$ with $K = 4$ in the seasonal component of the DLM to capture the annual, semiannual, triannual, and quarterly cycles. Our statistical assessment based on the calculation of the highest posterior density regions (see West and Harrison, 1999) 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

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

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 $\boldsymbol{\omega}_t^{\alpha,k}$ are independent over time, as well as independent of $\boldsymbol{\omega}_t^{\eta}$ for $t = 1, \ldots, T$.

Using the superposition principle (West and Harrison, 1999), we write the model as a hierarchy with a level one equation (commonly referred to as the observation equation in DLM literature) and a level two (or 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 we denote $n = 2 + 2K$ as the length of state vector $\boldsymbol{\theta}_t$. The matrices $\mathbf{G}$ and $\mathbf{W}_t$ are defined as $\mathbf{G} = \text{blockdiag}(\mathbf{G}^{\eta}, \mathbf{G}^{\alpha,1}, \ldots, \mathbf{G}^{\alpha,K})$ and $\mathbf{W}_t = \text{blockdiag}(\mathbf{W}_t^{\eta}, \mathbf{W}_t^{\alpha,1}, \ldots, \mathbf{W}_t^{\alpha,K})$, respectively. The state vector denoted as $\boldsymbol{\theta}_t$ takes the form $\boldsymbol{\theta}_t = (\eta_{1,t}, \eta_{2,t}, \alpha_{1,t}^1, \alpha_{2,t}^1, \ldots, \alpha_{1,t}^K, \alpha_{2,t}^K)$, 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 realizations of model simulations from a specified region, we consider a multivariate DLM that is an immediate extension of the univariate case. For the short-term analysis of the decadal, historical, and control experiments, R is 6, 5, and 10 respectively. Let $y_{t,r} = \mathbf{F}'\boldsymbol{\theta}_t + \nu_{t,r}$ denote the univariate spatially averaged temperature at time t, for $t = 1, \ldots, T$, of realization $r \in \{1, \ldots, R\}$. Each $\nu_{t,r}$ is independent and identically distributed from $N(0, V)$, where V now denotes the level one variance of the simulation data. Replacing y_t in Eq. (4) with a vector of R realization values $\mathbf{Y}_t = (y_{t,1}, \ldots, y_{t,R})'$ and $\nu_{t,r}$ with $\boldsymbol{\nu}_t = (\nu_{t,1}, \ldots, \nu_{t,R})'$, a vector of R independent and identically distributed error terms, only the level one equation changes, i.e.,

$$\mathbf{Y}_t = \mathbf{F}'\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \quad \boldsymbol{\nu}_t \sim N_R(0, V\mathbf{I}_R) \tag{6}$$
$$\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_n(0, V\mathbf{W}_t). \tag{7}$$

Note that $\mathbf{F}'$ is now a $R \times n$ dynamic regression matrix with identical rows, $\mathbf{F}_r' = (\mathbf{F}^{\eta'}, \mathbf{F}^{\alpha,1'}, \ldots, \mathbf{F}^{\alpha,K'})$ for $r = 1, \ldots, R$, with components defined in the previous section. As in the univariate case, the multivariate DLM still yields a single estimate for the baseline and seasonal components; however, this estimate now reflects the overall behavior of the realizations. The internal variability of each individual realization (as well as any other variability not included in the baseline and seasonal components) 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 filtering and backward smoothing methods for unknown constant level one variance (West and Harrison, 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 $\mathbf{m}_0 = (285, 0, \ldots, 0)'$, $\mathbf{C}_0 = \text{diag}(5, 2 \times 10^{-6}, 5, 1, \ldots, 1)$, $n_0 = 1$ and $S_0 = 0.01$.

3.3 Specification of the evolution variance

To complete the model specification, we require the sequence of the 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 $\text{Var}(\boldsymbol{\theta}_{t-1}|\mathbf{Y}_{1:(t-1)}) = \mathbf{C}_{t-1}$, the sequential updating equations produce the prior variance of

$\boldsymbol{\theta}_t$, $\mathbf{R}_t = \mathrm{Var}(\boldsymbol{\theta}_t|\mathbf{Y}_{1:(t-1)}) = \mathbf{G}\mathbf{C}_{t-1}\mathbf{G}' + \mathbf{W}_t$. 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 the static model with parameters that do not change over time (West and Harrison, 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}. To set the optimal seasonal discount factor values, we consider a maximum likelihood approach based on computing the one-step-ahead forecast distributions over a grid set of values of $(\delta_{\mathrm{base}}, \delta_{\mathrm{seas}})$ in $(0.9, 1) \times (0.9, 1)$ (see West and Harrison, 1999; Prado and West, 2010, for further details). We found high discount factors were generally optimal, suggesting the amplitudes do not vary significantly over time in the large-scale regions considered. In particular, the optimal seasonal discount factor δ_{seas} was found to be one, 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, which is a common practice in climate science. If small changes in the seasonal amplitudes exist (Santer et al., 2018) the changes are aliased in the DLM residuals, although we found no evidence of this in the data.

Selection of the baseline discount factor presents more of a challenge. We are empirically separating the long-term and seasonal temperatures from a correlated noise (which primarily consists of internal variability). There are many possibilities to do this, and intuitively, increasing the variability of one component will decrease that of another component. We seek to develop a modeling approach which provides a consistent framework for the analysis of all series. Within this constraint, it is also necessary to account for the theoretical differences in the three types of simulation. Recall from Sect. 2.1, the control run lacks external forcing, whereas the historical and decadal prediction runs are affected by time-varying anthropogenic and natural forcings. If the baseline is to capture long-term externally forced changes in temperature, intuitively this suggests the control baselines should be flat and exhibit little variability aside from random noise. To ensure this feature in our baseline, within each region, we optimize the discount factor (again using the maximum likelihood approach based on the one-step-ahead forecast distributions mentioned above) for the control runs and use that same discount factor, $\delta_{\mathrm{base}}^{\mathrm{mod}}$, for all model simulation types in that region. This data-driven modeling choice allows for a very clear comparison of the effects of time-varying anthropogenic and natural forcings present in the historical, decadal and observational series within each region, with respect to the control baseline.

It is important to note that because this is a data-driven modeling scheme and the baseline discount factors can vary from region to region, the amount of externally forced variability accounted for in the baselines will not necessarily be comparable between regions, only within a given region. For example, in the tropics, which is the smallest spatial region considered, we expect to see more variability resulting in lower discount factors. Lower baseline discount factors allow for flexibility, resulting in baselines which may reflect shorter-term externally forced changes such as 1–2 year cooling caused by volcanic eruptions. Alternatively, in the baselines which do not incorporate interaction between the hemispheres (such as the Northern Hemisphere and Southern Hemisphere regions), we can expect less variability which will result in higher discount factors. High discount factors will result in smooth baselines which primarily illustrate trends and omit shorter-term externally forced changes. As a sensitively study, we did consider using the optimized discount factor from the historical data instead. Although this did produce baselines which captured shorter-term changes in the externally forced temperature, the nonconstant baselines for the control incorporated more noise and diluted the utility of the control as a reference to discern the effects of long-term changes due to anthropogenic and natural forcings.

The presence of multiple realizations in the simulated data also presents a set of challenges. The multivariate DLM inherently produces smoother baselines than the univariate DLM as it is statistically averaging multiple simulation realizations. That is, if the same discount factor was selected for a univariate observational time series, $\delta_{\mathrm{base}}^{\mathrm{obs}}$, as that chosen for the average of the realizations, $\delta_{\mathrm{base}}^{\mathrm{mod}}$, the resulting observational baseline estimates would be more "wiggly" than those of the realizations. Obtaining observational baselines (from the univariate DLM) which exhibit the same amount of variability as the simulation baselines (from the multivariate DLM) requires 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 $N_2(0, \mathbf{W}_t^{\eta,\mathrm{mod}})$ and $N_2(0, \mathbf{W}_t^{\eta,\mathrm{obs}})$, we use the Bhattacharyya distance (Derpanis, 2008). More specifically, we select the value of $\delta_{\mathrm{base}}^{\mathrm{obs}}$ 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 in the short-term analysis, and between the 20CRV2 and the historical ensemble-mean for the long-term analysis. It is important to note again that this specification of the observational baseline discount factor ensures comparable baseline estimates of temperature variability between the observational products and the simulations 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, we are also interested in quantifying and assessing whether the model- and observation-based estimates of internal variability are consistent. The estimated DLM residuals (with baseline and seasonality removed) are time series that primarily represent the natural internal climate variability, which is not accounted for by the proposed DLM. We capture the structure of the residuals utilizing AR models, as described in the following paragraphs.

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 or observational 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. In order to explore if the autocovariances from month to month were dependent on the time of year, we initially considered a more general time-varying model, i.e., we considered an autoregressive model with time-varying coefficients and variance. A model such as this can be also written in DLM form with a single discount factor to control the variability of the AR coefficients over time, and another discount factor to control the variability of the variance over time. However, we found that the optimal discount factor values were equal to one, indicating that the standard static AR model was the optimal choice in all the cases. For the MIROC5 simulations with R realizations, this univariate model is easily extended to a multivariate autoregressive model. Let $z_{t,r}$ denote the residual time series for realization $r \in \{1, \ldots, R\}$ 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 Eq. (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 $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma^2 I_R)$. We chose this hierarchical AR model (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 straightforward to sample the posterior distributions directly using standard Bayesian linear regression techniques (Gelman et al., 2013).

In fitting the AR models, we also make the assumption that q may vary between the four spatial domains considered here, but that in any one domain, all residual time series for the simulations and the observational data sets have the same order q. We select the order q using the univariate time series of residuals for each simulation type, each domain, and each individual realization. The order of the fit is the order that maximizes the log-predictive likelihood (further details are available in Prado and West, 2010). The highest order of distinctly nonzero coefficients, over all types of simulations, all realizations, and all reanalyses, is then used as the order for all univariate and multivariate autoregressive models in that spatial domain. A sensitivity study in which we allowed unique orders for all data types indicated that the spectral estimates and model versus observational-record spectral differences are robust with respect to the choice of model order q. We note that constraining the order to be the same for all data types within a region ensures that any spectral differences "within domain" are unrelated to differences in 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. Each pair of complex roots can be 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 one 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 $\boldsymbol{\phi}$:

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

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 climate model simulation and the observational data sets. TVD was originally employed to compare probability distributions, and has also been used to measure the similarity of normalized spectra in Euan et al. (2018) and Alvarez et al. (2016). In order for TVD to be applicable to power spectra, normalization of 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 $f^*(\omega) =$

$f(\omega)/\int_\Omega f(\omega)\mathrm{d}\omega$ and $g^*(\omega) = g(\omega)/\int_\Omega g(\omega)\mathrm{d}\omega$ is defined as $\mathrm{TVD}(f^*, g^*) = 1 - \int_\Omega \min\{f^*(\omega), g^*(\omega)\}\mathrm{d}\omega$. For discrete normalized spectra, the TVD can equivalently be written in terms of the L_1 distance, $\mathrm{TVD}(f^*, g^*) = ||f^* - g^*||_1/2 = \sum_{\omega\in\Omega}|f^*(\omega) - g^*(\omega)|/2$. The distance measure takes on values $0 \leq \mathrm{TVD} \leq 1$, with 0 being the smallest possible discrepancy between spectra and 1 the largest.

Using the posterior spectra samples from the AR model, we can compute posterior distributions for the TVD values 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 series are statistically distinguishable from this reference spectrum. Next, using the maximum a posteriori (MAP) NCEP-2 and the 20CRV2 spectrum for the short- and long-term analyses (respectively), we employ TVD to assess the significance of the discrepancies between the internal variability spectra of the reference observational product and the MIROC5 simulations. We also show TVD values for the comparison between two observational spectra for both the long-term and short-term analyses, which provides a measure of the degree of difference we might expect due to uncertainties in the observation-based estimates of temperature variability.

3.6 Complete data analysis protocol

The protocol for the complete data analysis can be summarized in the following steps:

1. Select $\delta^{\mathrm{mod}}_{\mathrm{base}}$ as the highest optimal baseline discount factor from the control realizations. Set $\delta^{\mathrm{obs}}_{\mathrm{base}}$ to ensure the long-term externally forced variability is comparable in the baselines of the externally forced model simulations and the observational products, as described in Sect. 3.3.

2. To estimate the externally forced baseline and seasonal components, fit univariate DLMs to the observational data using selected $\delta^{\mathrm{obs}}_{\mathrm{base}}$ and $\delta^{\mathrm{obs}}_{\mathrm{seas}} = 1$, as detailed in Sect. 3.1. Fit multivariate DLMs to the simulation data using selected $\delta^{\mathrm{mod}}_{\mathrm{base}}$ and $\delta^{\mathrm{mod}}_{\mathrm{seas}} = 1$, as detailed in Sect. 3.2.

3. Compute the observed and simulated DLM residual time series which primarily represent the natural internal climate variability.

4. Select the order q of the autoregressive models used to capture the temporal structure of the residuals for all observational and simulation types. Fit univariate AR(q) to the observation DLM residuals and hierarchical AR(q) to the simulation DLM residuals, as described in Sect. 3.4.

5. To explore the simulated versus observed spectral differences, estimate the spectral densities from the autoregression coefficients and compute TVDs of estimated spectral densities as detailed in Sects. 3.4 and 3.5.

4 Result from the 30-year assessment of large-scale temperature

In this section, we first apply the previously described methodology to the short-term 30-year 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. 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 $\delta^{\mathrm{mod}}_{\mathrm{base}}$ and $\delta^{\mathrm{obs}}_{\mathrm{base}}$, MAP DLM level one 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). Note, again, the results presented are not meant to be compared directly between each region as our data-driven approach for extracting the various components of the time series does not ensure comparable components between regions, only within regions (see Sect. 3.3).

Figure 2 displays the 95 % posterior intervals for baselines $\eta_{1,t}$, estimated using the DLM model introduced in Sect. 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 (and therefore the baseline should reflect only random noise, as discussed in Sect. 3.3), whereas the reanalysis products and the historical and decadal prediction runs are affected by time-varying anthropogenic and natural forcings. Within each region, the control baselines are a reference to which we can compare the baselines of the externally forced runs. This comparison (within each region) shows secular temperature increases over the period from 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 the 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).

In the global and tropical regions, superimposed on the long-term warming trends in the reanalyses and the decadal prediction and historical runs are short-term (1–2 year) sur-

Table 1. Model baseline discount factor $\delta_{\text{base}}^{\text{mod}}$ and observation baseline discount factor $\delta_{\text{base}}^{\text{obs}}$. DLM smoothed estimates of level one 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.

NCEP-2; ERA-I; M-Decadal; M-Historical; M-Control

	Global	Tropical	Northern	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

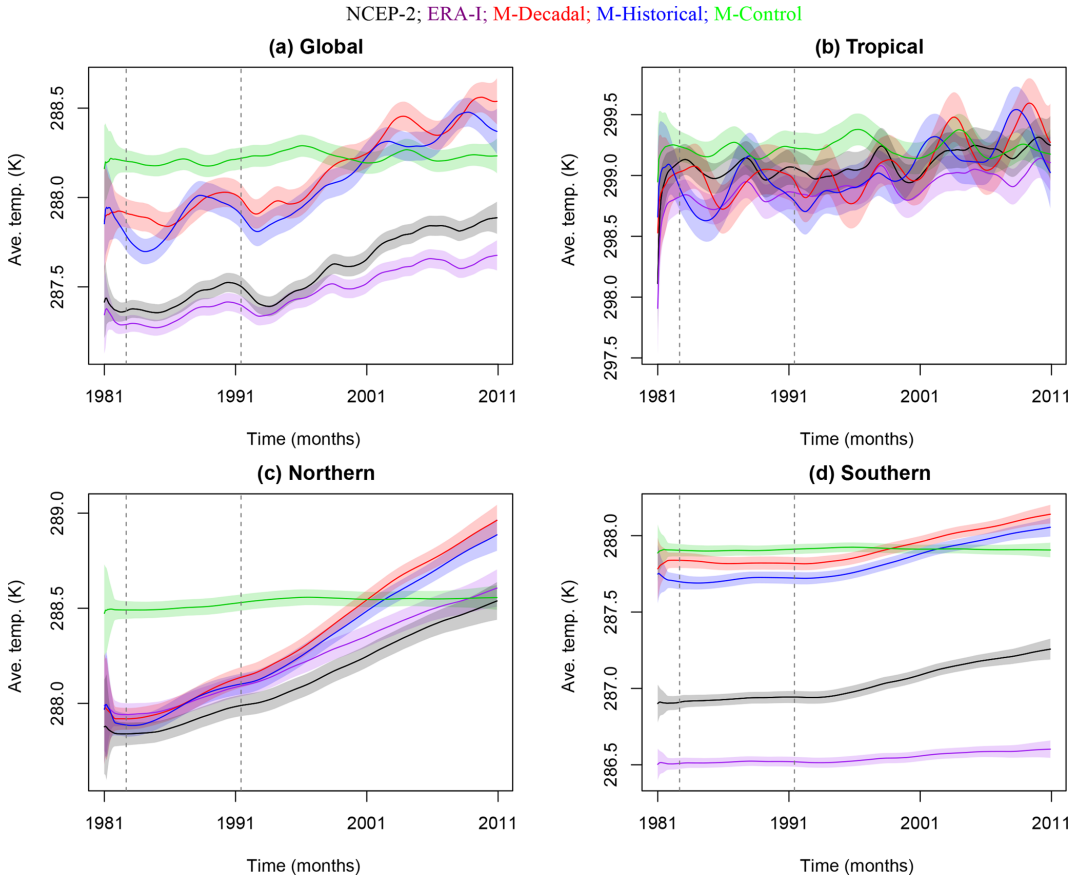

Figure 2. Baseline temperature estimates which capture long-term externally forced changes (as well as short-term cooling responses to volcanic eruptions). Different line colors denote the type of simulation and the reanalysis product. The top panels represent global **(a)** and tropical **(b)** regions; the bottom panels represent the Northern **(c)** and Southern hemispheres **(d)**. Vertical lines indicate the volcanic eruptions of El Chichón in 1982 and Pinatubo in 1991.

face 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 more pronounced in the global-spatial average, and are more noisy in the smaller-scale tropical averages. The surface cooling signals caused by El Chichón and Pinatubo are markedly smaller in the Northern Hemisphere and Southern Hemisphere averages than for the global domain. This is the result of the Northern Hemisphere and Southern Hemisphere baselines being estimated with discount factors close to one (see Table 1). As mentioned briefly in Sect. 3.3, 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, thus resulting in higher baseline discount factors in the hemispheric

regions. The selection of high discount factors suggests that the externally forced longer-term variability in both hemispheres was very close to linear. 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. Any shorter-term forced variability not captured by the baselines will be reflected in the residuals. As a sensitivity study (also mentioned in Sect. 3.3), we considered more flexible baselines which incorporated more variability, and therefore resulted in less variability in the residuals. Although we found the results of the analysis on the residuals robust with respect to the specification of variability of the baseline, further investigation of the differences in the amplitude of the global-average and hemispheric-average volcanic signals may be of interest.

Figure 2 also yields many other features of interest, such as differences in the 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 two 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 Southern Hemisphere bias is that it may arise due to 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 the 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 Southern Hemisphere. Additionally, the land surface temperature is better sampled in 1981 than the SST. This could also be a cause of bias, as there is a larger contribution from the land surface temperature in the Northern Hemisphere and tropical regions than in the Southern Hemisphere.

The 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 Southern Hemisphere imprint of structural 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. Further, 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 $\alpha^k_{1,t}$ for $k = 1, 2$ (i.e., for the amplitudes of the annual and semiannual cycles, respectively). Amplitudes were estimated using the DLM model in Sect. 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 zero and indistinguishable from one another; therefore, they are not shown. Results for the annual and semiannual cycles are more interesting. Consider the reanalyses first. For all four spatial domains, and for both $k = 1$ 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.

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 Southern Hemisphere, 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 4 illustrates the 95 % probability intervals on the posterior spectra, normalized with respect to white noise on the log scale. Spectra were estimated using the methods presented in Sect. 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 the 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

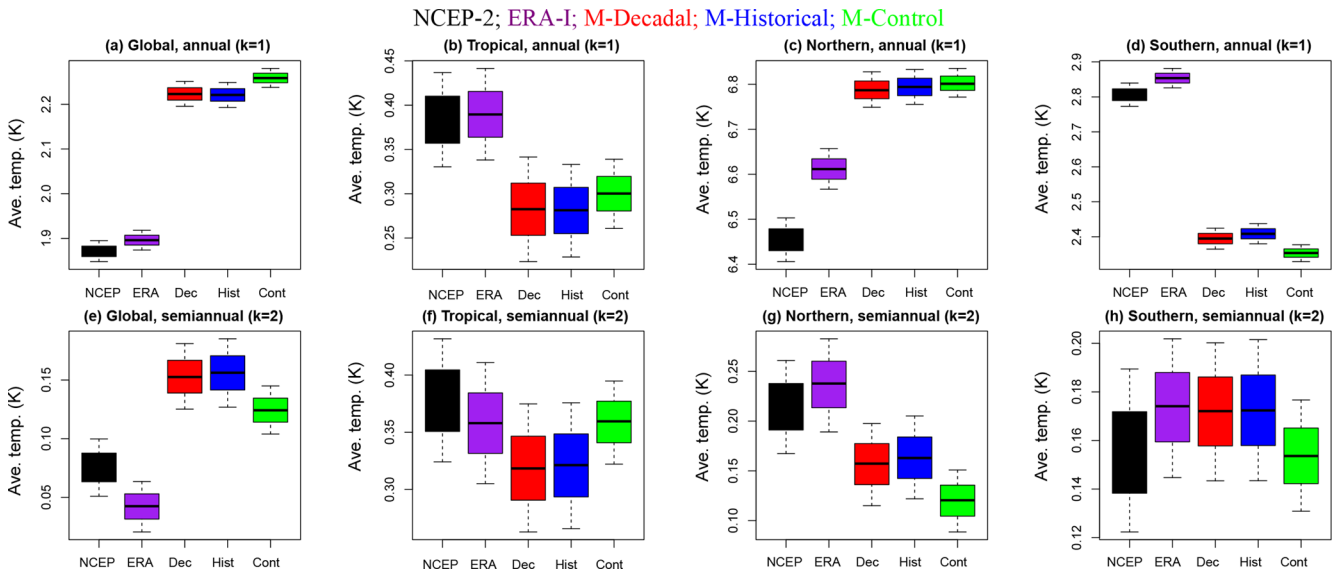

Figure 3. Posterior amplitude samples for harmonics $k = 1, 2$. Varying model versus reanalysis differences are seen across domains; however, we do not find any cases in which there are significant amplitude differences between the three types of model simulation. Whiskers indicate the maximum and minimum values and boxes indicate the 95 % posterior intervals. Different colors denote the type of simulation and the reanalysis product.

variability acting on a range of different timescales, such as the Madden–Julian Oscillation, ENSO, and the Interdecadal Pacific Oscillation.

Other features of Fig. 4 are also noteworthy. First, within each region, the spectra for the three different types of MIROC5 simulation are very similar. This suggests that the DLM method applied here has consistently estimated the internally generated component of surface temperature within each region from (1) the significant 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., 2018). 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., 2018). We caution, however, that the inference on "observed" estimates of internal variability on multi-decadal timescales is limited by the relatively short (30-year) time-period.

Recall from Sect. 3.4 that the presence of complex roots points towards the existence of quasi-cyclical temperature variations. The results in the fifth row of Table 1 indicate that complex roots are only consistently obtained 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 the sixth 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.

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 Fig. 5. Figure 5a–d show 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 sim-

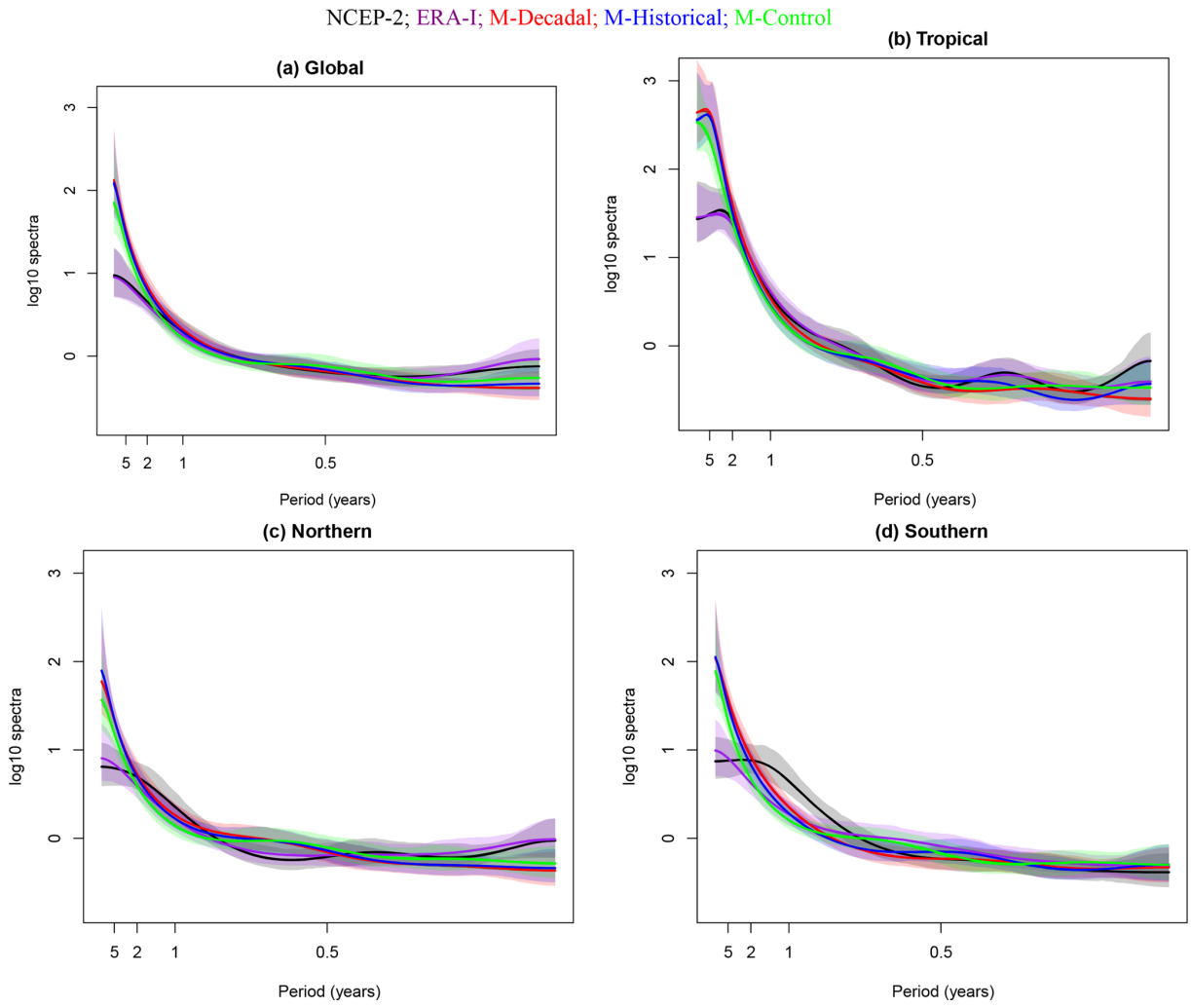

Figure 4. MAP AR $\log_{10}$ spectra normalized with respect to white noise for each climate simulation by region with the 95 % posterior intervals shaded. Instead of the frequency ω, the x axis is labeled at select years ($2\pi/12\omega$). Different line colors denote the type of simulation and the reanalysis product. The top panels represent global **(a)** and tropical **(b)** regions; the bottom panels represent the Northern Hemisphere **(c)** and Southern Hemisphere **(d)**.

ulations 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. 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 abovementioned finding that the discrepancy between low-frequency temperature variability in the reanalyses and the MIROC5 simulations is largest in the tropics.

Figure 5e–h display results for the comparison between the model spectra and the NCEP 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 abovementioned 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 Result from the 63-year assessment of large-scale temperature

To examine the sensitivity of our results to record length we apply a similar analysis to a time period of 63 years, Jan-

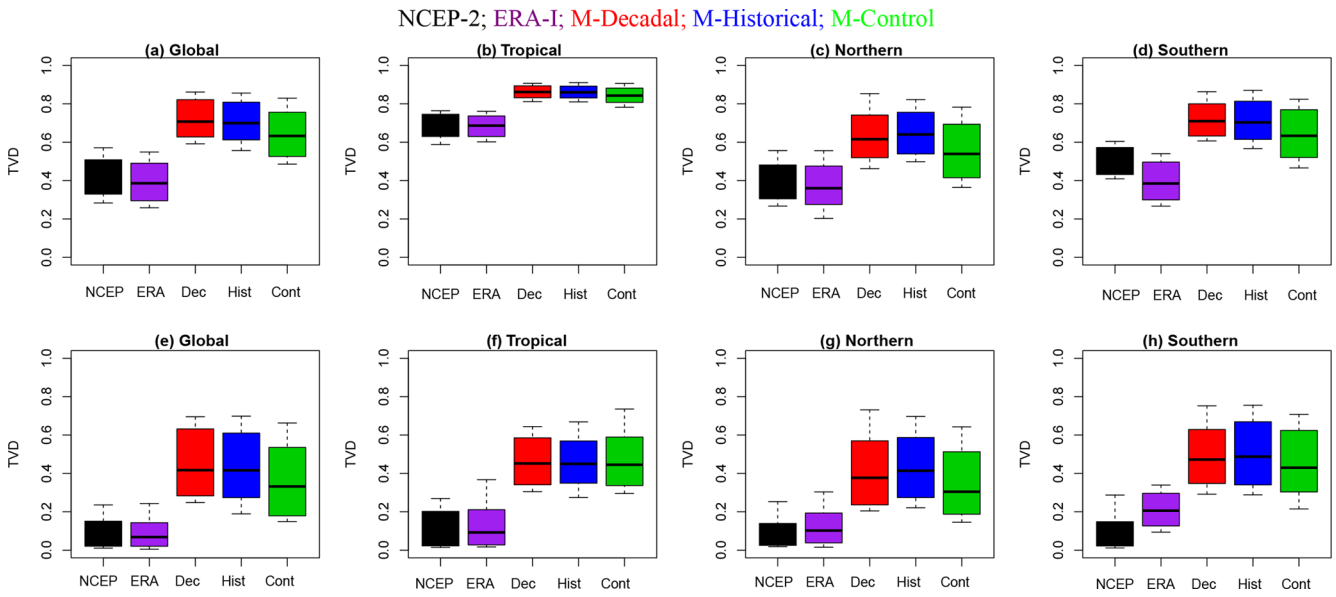

Figure 5. **(a–d)** TVD calculated from $\boldsymbol{\phi}$ samples with white noise as the reference. **(e–h)** TVD calculated from $\boldsymbol{\phi}$ samples with the MAP NCEP spectrum as the reference. Larger TVD values indicate higher discrepancy between spectral densities. Whiskers indicate the maximum and minimum values and boxes indicate the 95 % posterior intervals. The scenario or observational product is indicated using colors. From left to right: global **(a, e)**, tropical **(b, f)**, Northern Hemisphere **(c, g)**, and Southern Hemisphere **(d, h)**.

uary 1950 to December 2012. Excluding the decadal experiment which does not include realizations covering more than 30 years, we use MIROC5 simulation data from the historical and control experiments. We compare the model results to the results obtained from the 20CRV2 and BEST observational products. Again, we would like to reiterate that the results are not meant to be directly compared between each region. Table 2 provides summaries of the DLM and AR estimated model parameters and posterior inferences for each spatial domain. This includes the baseline discount factors $\delta_{\text{base}}^{\text{mod}}$ and $\delta_{\text{base}}^{\text{obs}}$, MAP DLM level one 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 6 displays the 95 % probability intervals for the DLM estimated baselines $\eta_{1,t}$. The control baseline is seen to be flatter than the historical baseline, as the control runs do not incorporate changes in the external forcings. However, the historical runs and observational products are affected by changes in anthropogenic forcing, resulting in the temperature increases seen over the period from 1950 to 2012. Cooling effects of volcanic eruptions are also evident in the baselines with lower discount factors (global, tropical, and Southern Hemisphere), similarly to what is observed in the short-term analysis. In this longer analysis, the Southern Hemisphere exhibited more variability, resulting in a lower baseline discount factor (see Table 2) than in the short-term analysis, which has an optimal value close to one (see Table 1). It is not unexpected for a larger time-span to reflect more temperature variability than a more narrow time-span. The variability in the Northern Hemisphere remains less dynamic than that of the other regions, as is the case for the short-term analysis. This is reflected in a baseline discount factor that is close to one (see Table 2). Again, we see the most variability in the smaller-spatial-scale tropical averages.

Figure 6 also illustrates a distinct discrepancy between the two observational products. This is most noticeable in the global region and the Northern and Southern hemispheres, where 20CRV2 is consistently warmer than BEST. The discrepancy is present, but not very strong in the tropics, where BEST is warmer than 20CRV2. It can also be seen that the baseline for the historical simulations is cooler than that of the observational products in the Northern Hemisphere, and warmer than the observational products in the Southern Hemisphere, whereas for the other two regions there are no relevant discrepancies.

Figure 7 shows the 95 % posterior intervals of the seasonal amplitudes of the annual and semiannual cycles. As in the 30-year analysis, for all four spatial domains, the harmonics corresponding to the trimestral and quarterly cycles are very close to zero and are indistinguishable from one another; therefore, they are not shown. The observational products exhibit differences in both annual and semiannual cycles, for the global and Southern Hemisphere domains. BEST exhibits lower amplitudes than 20CRV2 in the global domain, with the opposite being true in the Southern Hemisphere. Differences between the observational products are also seen in the annual cycle in the tropics, where BEST is

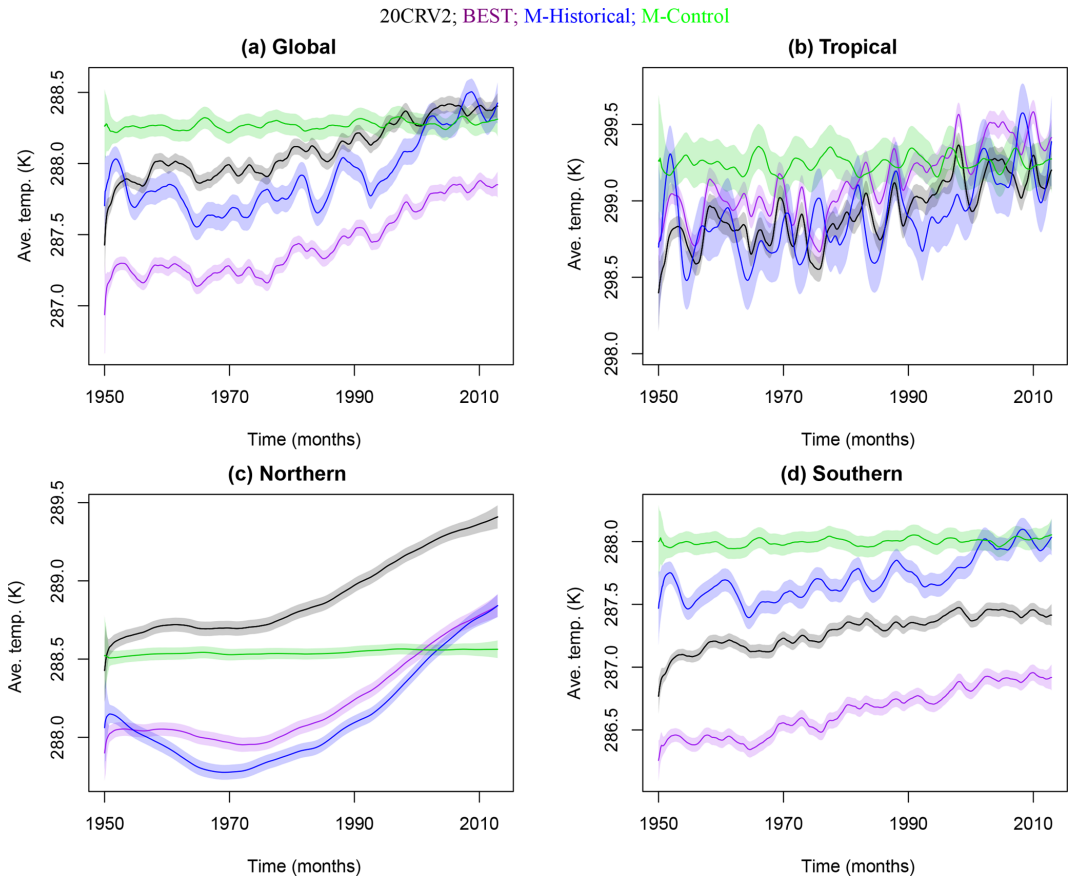

Figure 6. Same as in Fig. 2 but for the long-term analysis.

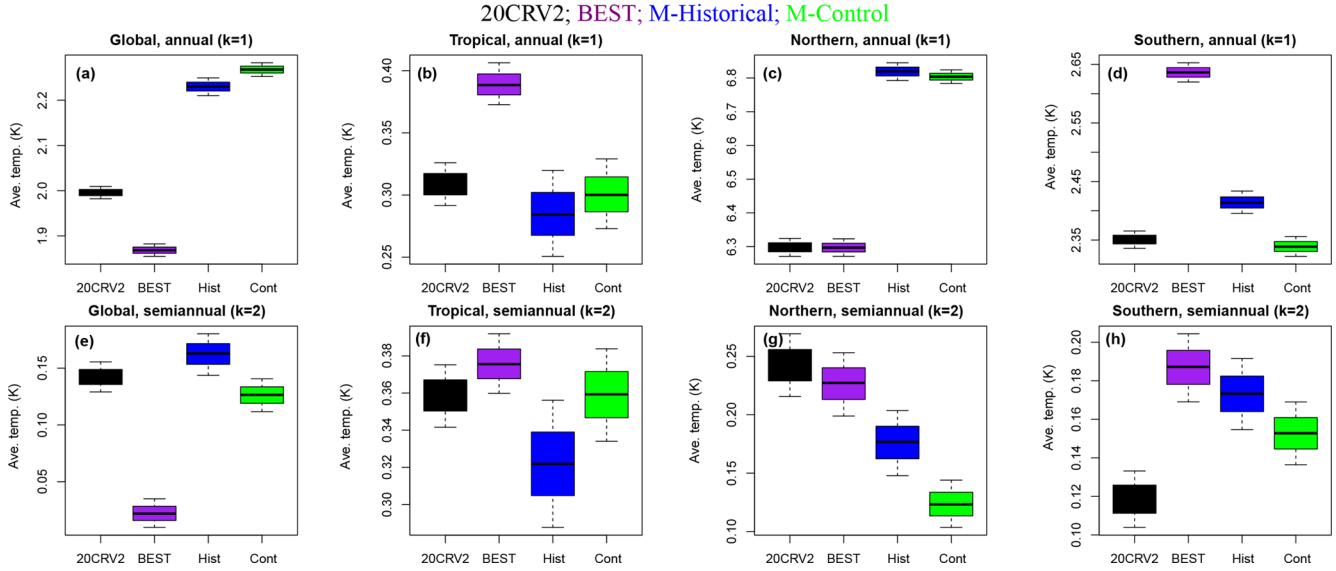

Figure 7. Same as in Fig. 3 but for the long-term analysis.

Table 2. Model baseline discount factor $\delta_{\text{base}}^{\text{mod}}$ and observation baseline discount factor $\delta_{\text{base}}^{\text{obs}}$. DLM smoothed estimates of residual 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.

20CRV2; BEST; M-Historical; M-Control

	Global	Tropical	Northern	Southern
$(\delta_{base}^{mod}, \delta_{base}^{obs})$	(0.91, 0.94)	(0.88, 0.92)	(0.99, 0.99)	(0.92, 0.95)
DLM MAP V	0.02, 0.02, 0.16, 0.22	0.07, 0.07, 0.54, 0.61	0.07, 0.07, 0.34, 0.4	0.02, 0.03, 0.17, 0.28
AR order q	4	8	5	5
AR MAP σ^2	0.009, 0.010, 0.038, 0.069	0.007, 0.005, 0.024, 0.033	0.026, 0.025, 0.105, 0.188	0.009, 0.016, 0.042, 0.080
Maximum modulus	0.46, 0.69*, 0.94*, 0.93*	0.87, 0.89, 0.93, 0.92	0.65, 0.75*, 0.93*, 0.91*	0.61, 0.67*, 0.94*, 0.95*
Complex root, max modulus, ρ	0.46, 0.42, 0.46, 0.49	0.87, 0.89, 0.93, 0.92	0.65, 0.43, 0.51, 0.55	0.61, 0.51, 0.45, 0.53
Corresponding wavelength, λ	314.16, 5.07, 4.11, 4.11	23.27, 26.18, 57.12, 62.83	20.27, 3.81, 4.83, 4.91	24.16, 4.30, 3.83, 5.03

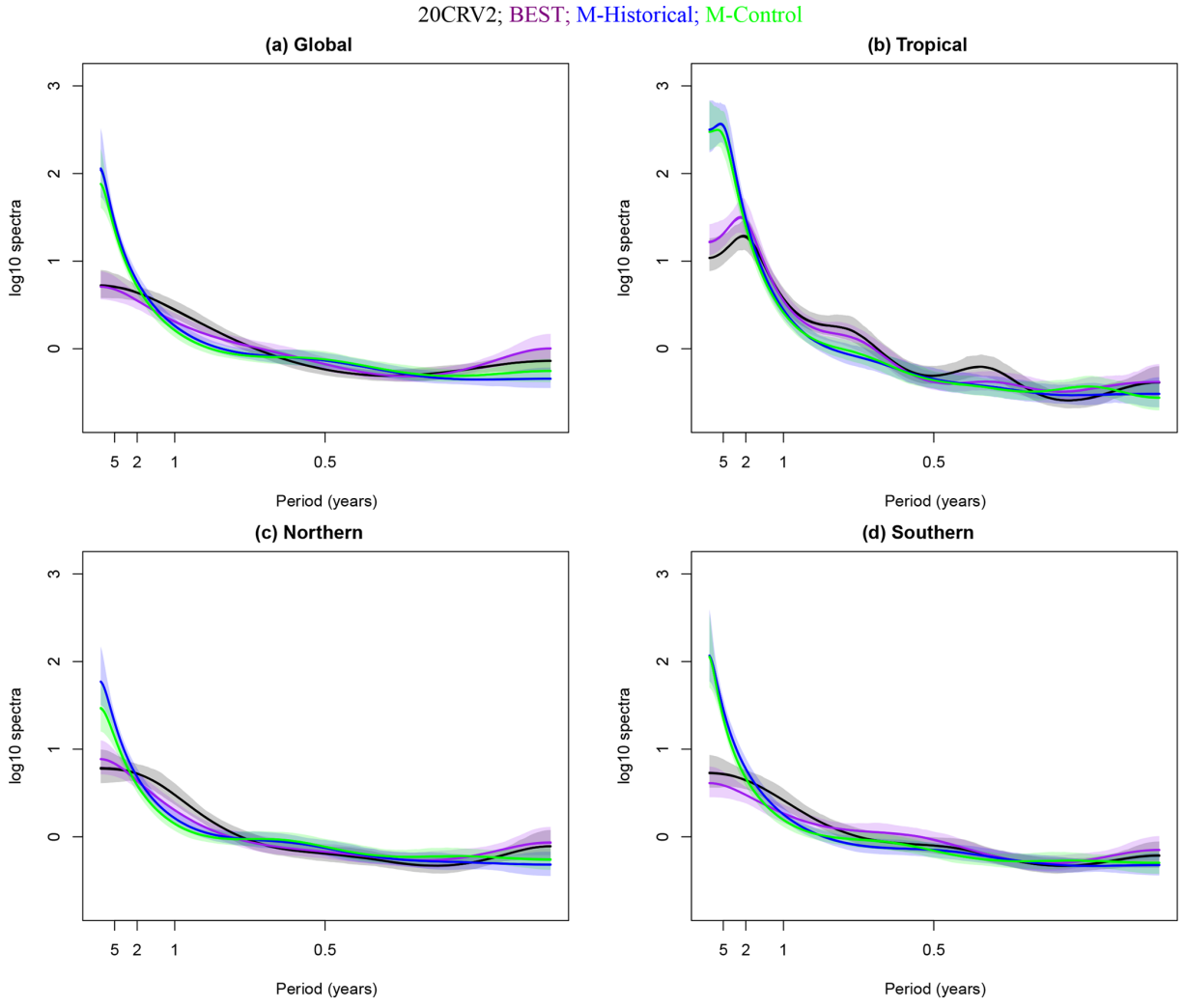

Figure 8. Same as in Fig. 4 but for the long-term analysis.

distinctly higher. In the Northern Hemisphere the two observational products are indistinguishable. In general, the model versus observation differences in amplitude are less clear, although pronounced differences are seen in the annual cycle of the globe and Northern Hemisphere.

Figure 8 shows the 95 % posterior intervals for the spectra, normalized with respect to white noise on the log scale. Once again, the tropics exhibit the least smooth spectra, where an AR of an order higher than in the other cases ($q = 8$, see Table 2) provides the best fit for the residuals. The observa-

tional spectra are indistinguishable within all spatial regions. Furthermore, the model run spectra are also indistinguishable within all spatial regions and exhibit notably higher power in the low frequencies than the observational spectra. This is consistent with the previous 30-year analysis. However, the model versus observational differences seen here are stronger than in the 30-year analysis.

We explore possible quasi-cyclical temperature variations by considering the complex roots of the AR characteristic polynomial. Table 2 again indicates that complex roots with moduli exceeding 0.8 are only obtained for the tropical domain – most likely reflecting temperature variability associated with the occurrence of El Niños and La Niñas. The wavelengths for the tropical quasi-periodic variability are approximately 23.27 months (1.94 years) for 20CRV2, 26.18 months (2.18 years) for BEST, 57.12 months (4.76 years) for the historical simulation, and 62.8 months (5.23 years) for the control run. Notice that even in a 63-year record there is an absence of quasi-periodic behavior on long timescales, as was the case for the analysis of the 30-year record. To avoid redundancy, we have omitted the TVD summaries, as they share many similarities with the 30-year analysis.

6 Conclusions

We developed a model diagnostic statistical methodology and protocol that can be generally applied to the assessment and comparison of simulations from CMIP models. The methods presented here have two main advantages: the protocol developed is a consistent and statistically principled approach to jointly estimate the components of the temperature time series, and the results incorporate probabilistic uncertainty as the approach is model-based. Within this protocol, 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). The three sets of numerical experiments considered were initialized decadal predictions (not included in the long-term analysis), control runs, and uninitialized simulations of historical climate change. Removal of the seasonal and baseline components from the raw temperature data yielded residuals that primarily provided information on unforced natural internal climate variability. We characterized this internal variability by fitting univariate and multivariate autoregressive (AR) models to the residuals. As 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 illustrated our approach in both a short-term and long-term analysis using one selected climate model (MIROC5). In the short-term 30-year analysis, estimation of the various temperature components was performed for two reanalysis data sets and for three different types of experiments. Similarly, in a long-term 63-year analysis, estimation of these components was performed for a 20th century reanalysis, gridded in situ data, and two of the three different simulated experiments. We found significant differences between the observational product data and the model-generated simulations in all three temperature components (seasonal, baseline, and internal variability), for both the short- and long-term analyses.

From a climate perspective, two results of our analyses 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 simulation runs yielded very similar spectral estimates of internal temperature variability within each region – 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 model spectra is reassuring, and implies that our statistical analysis methods are consistently extracting comparable components of internal variability for the simulation data within each region. The second intriguing result emerged from the comparison of the model and observational product 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), we find no evidence that the MIROC5 model systematically underestimates the amplitude of low-frequency internal variability inferred from observational product data. Our methodology and results present a novel approach for obtaining data-driven estimates of the amplitude of observed multi-decadal temperature variability, thereby providing a more solid observational "target" for model evaluation purposes.

Code and data availability. The data used in the short-term analysis MIROC5, NCEP Reanalysis 2, and ERA-Interim are available from https://esgf-node.llnl.gov/search/cmip5/ (last

access: 13 August 2018), https://www.esrl.noaa.gov/psd/ (last access: 13 August 2018), and https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim (last access: 13 August 2018), respectively. For the long-term analysis, observational record data from the BEST and 20CRV2 are available from http://berkeleyearth.org/data/ (last access: 13 August 2018) and https://www.esrl.noaa.gov/psd/data/gridded/data.20thC_ReanV2.monolevel.mm.html (last access: 13 August 2018), respectively. The R package for DLMs, "dlm", is available online (https://cran.r-project.org/web/packages/dlm/index.html, last access: 13 August 2018).

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors wish to thank Ben Santer, Ana Kupresanin, and Francisco Beltran at LLNL for helpful conversation. We also thank the editor, associate editor, and referees for their comments. Bruno Sansó was partially funded by the National Science Foundation (grant no. DMS-1513076). Raquel Prado was partially funded by the National Science Foundation (grant no. SES-1461497).

Review statement. This paper was edited by Chris Forest and reviewed by three anonymous referees.

References

Alvarez, E., C., P., Euan, 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, Q. J. Roy. Meteor. Soc., 137, 1381–1399, 2011.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The twentieth century reanalysis project, Q. J. Roy. Meteor. Soc., 137, 1–28, 2011.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., 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, Stat. Med., 37, 2855–2873, https://doi.org/10.1002/sim.7695, 2018.

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., Homeyer, C. R., Knox, J. A., Krüger, K., Lambert, A., Long, C. S., Martineau, P., Molod, A., Monge-Sanz, B. M., Santee, M. L., Tegtmeier, S., Chabrillat, S., Tan, D. G. H., Jackson, D. R., Polavarapu, S., Compo, G. P., Dragani, R., Ebisuzaki, W., Harada, Y., Kobayashi, C., McCarty, W., Onogi, K., Pawson, S., Simmons, A., Wargan, K., Whitaker, J. S., and Zou, C.-Z.: Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems, Atmos. Chem. Phys., 17, 1417–1452, https://doi.org/10.5194/acp-17-1417-2017, 2017.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B.: Bayesian data analysis, Chapman and Hall/CRC, New York, USA, 2013.

Gibson, P. B., Perkins-Kirkpatrick, S. E., Alexander, L. V., and Fischer, E. M.: Comparing Australian heat waves in the CMIP5 models through cluster analysis, J. Geophys. Res.-Atmos., 122, 3266–3281, 2017.

Imbers, J., Lopez, A., Huntingford, C., and Allen, M.: Sensitivity of climate change detection and attribution to the characterization of internal climate variability, J. 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., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-year reanalysis project, B. Am. Meteorol. Soc., 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), B. Am. Meteorol. Soc., 83, 1631–1643, 2002.

Kay, J., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S., Danabasoglu, G., Edwards, J., and Holland, M.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, B. Am. Meteorol. Soc., 96, 1333–1349, 2015.

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, Cambridge, UK and New York, USA, 2013.

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

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., and Stockdale, T.: Decadal prediction: can it be skillful?, B. Am. Meteorol. Soc., 90, 1467–1485, 2009.

Perkins-Kirkpatrick, S. E., Fischer, E. M., Angélil, O., and Gibson, P.: The influence of internal climate variability on heatwave frequency trends, Environ. Res. Lett., 12, 044005, https://doi.org/10.1088/1748-9326/aa63fe, 2017.

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

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickhams, C., and Mosher, S.: Berkeley Earth Temperature Averaging Process, Geoinfor Geostat: An Overview, 1, 20–100, https://doi.org/10.4172/2327-4581.1000103, 2013.

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, J. Geophys. Res.-Atmos., 106, 28033–28059, 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., Perlwitz, J., Solomon, S., Stott, P. A., Taylor, K. E., Terray, L., Thorne, P. W., Wehner, M. F., Wentz, F. J., Wigley, T. M. L., Wilcox, L. J., and Zou, C.-Z.: Identifying human influences on atmospheric temperature, P. Natl. Acad. Sci., 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, 6399, https://doi.org/10.1126/science.aas8806, 2018.

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

Solomon, S., Daniel, J. S., Neely, R. R., Vernier, J.-P., Dutton, E. G., and Thomason, L. W.: The persistently variable background stratospheric aerosol layer and global climate change, Science, 333, 866–870, 2011.

Taylor, K. E.: A summary of the CMIP5 experiment design, available at: 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, B. Am. Meteorol. Soc., 93, 485–498, 2012.

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., and Masui, T.: The representative concentration pathways: an overview, Climatic Change, 109, 5, https://doi.org/10.1007/s10584-011-0148-z, 2011.

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

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