# A New Method for Assessing the Performance of General Circulation Models Based on Their Ability to Simulate the Response to Observed Forcing

M. A. ALTAMIRANO DEL CARMEN,[a] F. ESTRADA,[a,b,c] AND C. GAY-GARCÍA[a]

[a] *Centro de Ciencias de la Atmósfera, Universidad Nacional Autónoma de México, Mexico City, México*
[b] *Institute for Environmental Studies, Vrije Universiteit, Amsterdam, Netherlands*
[c] *Programa de Investigación en Cambio Climático, Universidad Nacional Autónoma de México, Mexico City, México*

(Manuscript received 4 July 2020, in final form 15 March 2021)

ABSTRACT: The reliability of general circulation models (GCMs) is commonly associated with their ability to reproduce relevant aspects of observed climate, and thus the evaluation of GCM performance has become a standard practice for climate change studies. As such, there is an ever-growing literature that focuses on developing and evaluating metrics to assess GCM performance. In this paper it is shown that some commonly applied metrics provide little information for discriminating GCMs based on their performance, once uncertainty is included. A new methodology is proposed that differs from common approaches in that it focuses on evaluating GCMs' abilities to reproduce the observed response of surface temperature to changes in external radiative forcing (RF), while controlling for observed and simulated variability. It uses formal statistical tests to evaluate two aspects of the warming trend that are central for climate change studies: 1) if the response to RF produced by a particular GCM is compatible with observations and 2) if the magnitudes of the observed and simulated rates of warming are statistically similar. We illustrate the proposed methodology by evaluating the ability of 21 GCMs to reproduce the observed warming trend at the global scale and for eight subcontinental land domains. Results show that most of the GCMs provide an adequate representation of the observed warming trend for the global scale and for domains located in the Southern Hemisphere. However, GCMs tend to overestimate the warming rate for domains in the Northern Hemisphere, particularly since the mid-1990s.

## 1. Introduction

Most of the scientific basis supporting the attribution of climate change to anthropogenic causes (IPCC 2007, 2013b) is based on atmosphere–ocean general circulation models (GCMs). GCMs are deterministic numerical computational programs, with physical bases that simulate the response of Earth's climate system (Jun et al. 2008) driven by different forcings (natural, anthropogenic, or combined) on time scales from decades to centuries, without explicitly including meteorological observations. These models are able to closely reproduce many of the physical aspects of the current climate including features of forced and unforced variability (Randall et al. 2007; Gleckler et al. 2008). The Coupled Model Intercomparison Project (CMIP), coordinated by the World Climate Research Program, is the most important international effort conducted to advance climate change modeling and to support the IPCC's reports. The performance of GCMs that are part of the CMIP5 has improved with respect to those of the CMIP3, increasing the confidence about climate change detection and attribution at the regional and global scales (Bindoff et al. 2013). GCM performance evaluations have been used as an indicator of their reliability for future projections (Jun et al. 2008; Baumberger et al. 2017), to assign weights to individual models in a multimodel ensemble (Tebaldi and Knutti 2007; Knutti et al. 2010; Christensen et al. 2010; Herger et al. 2018), and to select a subset of the "best" models (Perkins et al. 2007; Maxino et al. 2008; Knutti et al. 2010; Sanderson et al. 2015; Herger et al. 2018), which are then used for conducting climate change assessments such as impact, vulnerability, and adaptation studies (IPCC 2014). However, it is also recognized that the usefulness of a GCM cannot be inferred solely from its degree of agreement with observations (Notz 2015).

In principle, GCMs could be considered independent as they have been developed by different groups and with somewhat different modeling strategies and goals. Most studies, including this one, assume explicitly or implicitly that each GCM is independent from the others and as a random drawn from a distribution with the true climate as its mean (Tebaldi and Knutti 2007; Jun et al. 2008). This implies that the average of an ensemble of GCMs should converge to the true climate as more GCMs are included (Jun et al. 2008). In practice, GCMs are hardly independent as some share common genealogy, numerical schemes for solving equations, parameterizations, and/or components (Jun et al. 2008; Christensen et al. 2010; Masson and Knutti 2011; Steinschneider et al. 2015). In consequence, it is not completely justified to assume GCMs as independent in ensembles of opportunity (Masson and Knutti 2011; Knutti et al. 2013). However, the relationship between models and families of models is not clear and the authors believe that there is no objective and convincing way yet to address this problem.

It is generally accepted that the skill of a GCM to simulate the climate for which there are observational instrumental records is a measure of its performance (Jun et al. 2008; Notz

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-20-0510.s1.

*Corresponding author*: F. Estrada, feporrua@atmosfera.unam.mx

DOI: 10.1175/JCLI-D-20-0510.1

2015; Baumberger et al. 2017). However, simple comparison between simulated and observed climate to assess model skill may not be adequate. The twentieth-century simulations included in the CMIP5 experiment were designed to reproduce the response to observed changes in natural and anthropogenic radiative forcing (i.e., the climate signal), but were not constrained using climate records to replicate relevant aspects of observed variability such as the time of occurrence and phase of natural oscillations. In particular, it is known that low-frequency natural variability can significantly distort the response to radiative forcing (RF), masking it or exaggerating it, depending on the phase of the oscillation (Tsonis et al. 2009; Wu et al. 2011; Estrada et al. 2013b). Initial conditions in GCMs have a much larger effect than previously thought and can significantly distort the warming trend over the observed period and even in future projections (Wallace et al. 2015; Deser et al. 2014). Differences between observed and simulated internal variability are confounding factors that can affect common approaches for evaluating GCM performance (Maraun et al. 2010).

There are a large number of metrics that could be selected to evaluate a wide variety of aspects (Christensen et al. 2010), and new metrics are commonly proposed for particular purposes (Knutti et al. 2010; Herger et al. 2018). However, there is no objective approach for choosing metrics to evaluate GCMs performance and there is little consensus on which metrics are useful to discriminate ''good'' from ''bad'' GCMs (Knutti et al. 2010). An a priori subjective selection of a limited set of metrics with largely unknown interdependencies is difficult to avoid (Christensen et al. 2010). Furthermore, objective methods to evaluate GCM performance could be useful to maximize the value of climate change projections (Knutti et al. 2010). However, empirical evidence supporting this statement is, at best, weak and GCMs with good historical performance could underperform when projecting future climate (Weigel et al. 2010; Notz 2015).

The objective of this paper is to present a new methodology to evaluate GCMs' ability to reproduce the response of the observed mean surface temperature anomaly (MST) to changes in RF at the global and regional scales. Linear least squares regression and formal statistical tests are combined to develop an objective, systematic, and robust method for evaluating GCM performance for climate change applications. The proposed methodology focuses on the central problem of GCM evaluation in the context of climate change: determining the skill of GCMs to simulate the climate system response to changes in external radiative forcings. This objective is achieved using time series models similar to those that have applied for the study of different aspects of climate variability (Tol 1996; Taylor and Buizza 2004), downscaling methods (Estrada et al. 2013a; Estrada and Guerrero 2014), impact assessment (Burke et al. 2015; Hsiang 2016), and climatic change detection and attribution (Tol and de Vos 1993; Harvey and Mills 2002; Qu 2011; Estrada et al. 2013b,c; Estrada and Perron 2014). The proposed methodology underlines the similarities and differences between the evaluation of GCMs and attribution studies, as it is based on evaluating the capacity of models to reproduce observed warming trends.

Section 2 describes the databases of observed and simulated MST used in this study. In section 3 the proposed methodology is presented, and the limitations of classical metrics to determine GCMs performance are discussed. Section 4 shows the results of the proposed methodology applied to the global scale and eight subcontinental land domains distributed across the globe. Conclusions and a summary are given in section 5.

## 2. Data

### a. *Observational data and spatial domains*

We considered observational data of ocean and land monthly average surface temperature (in K) from two gridded observational datasets: 1) the Hadley Centre Climate Research Unit temperature anomalies (HadCRUT4, version 4.6.0.0) on a $5° \times 5°$ global grid, available for the period 1850–2018 (Morice et al. 2012) and 2) the GISS surface temperature analysis (GISTEMP v4) on a $2° \times 2°$ global grid, available for the period 1850–2005 (GISTEMP Team 2021; Hansen et al. 2010; Lenssen et al. 2019). Temperature over land is measured at stations, whereas temperature over the ocean is derived from sea surface temperature and marine air temperature measurements taken by ships and buoys (Jun et al. 2008). Each of these research centers conduct independent analyses of data quality, inhomogeneities, and corrections of instrumental biases at the grid cell level.

To illustrate the proposed methodology, our analysis focuses on the global scale and on eight subcontinental land regions (Fig. 1) that are characterized by different climatic regimes and for which GCM performance has been evaluated in previous research (IPCC 2013a; Qian and Zhang 2015; Chan and Wu 2015). The selected regions are the United States (USA), western Europe (EuW), northern Europe (EuN), Mexico (Mex), and China (Chi) in the Northern Hemisphere (NH) and the Amazon (Ama), southern Africa (SAf), and Australia (Aus) in the Southern Hemisphere (SH). Ocean regions were not included because data tend to be sparser.

The period 1910–2005 was chosen for this study since these are the years for which the HadCRUT4 and GISTEMP datasets have no missing data in more than 70% of the grid points over most of the selected domains. Data completeness is similar for both HadCRUT4 and GISTEMP datasets for the domains located in the NH, with the exception of China. Data gaps are larger in the SH and this is more evident for GISTEMP over regions located in the Amazon and southern Africa than for HadCRUT4 due to differences in data processing mentioned above. The spatial average and time series anomalies of annual MST were calculated with respect to the reference period 1961–90 (Fig. 2).

The observed global and regional MST are influenced by atmospheric and oceanic natural climate variability that can distort the underlying response of the MST to changes in RF. To account for their confounding effects, natural variability modes are considered in our analysis. Modes like the Atlantic multidecadal oscillation (AMO), the Pacific decadal oscillation (PDO), the Southern Oscillation index (SOI), the northern annular mode (NAM), and the North Atlantic Oscillation

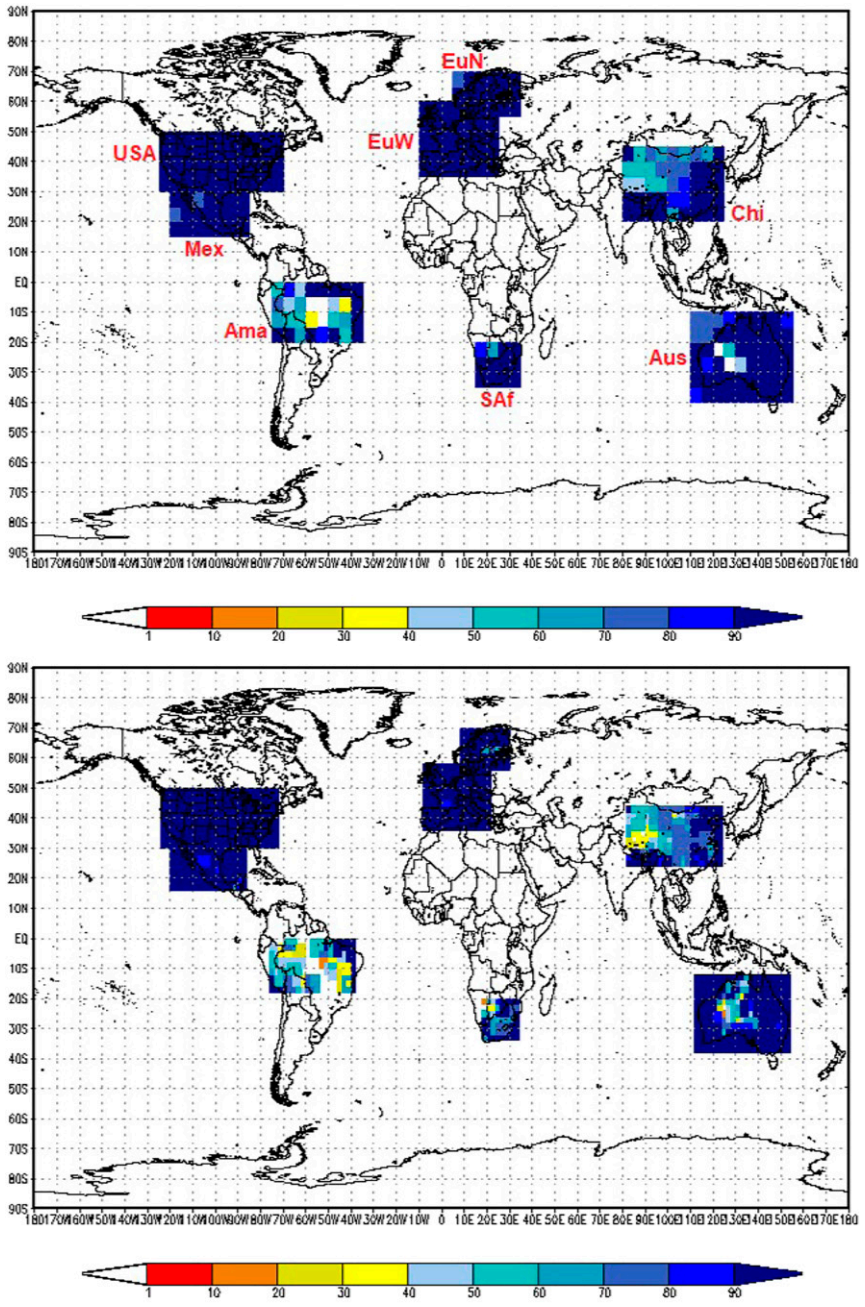


FIG. 1. Data availability for the subcontinental land regions considered in the analysis. The color bar shows the percentage of observed monthly MST data available at each grid point in subcontinental land domains during the 1910–2005 period for (top) HadCRUT4 and (bottom) GISTEMP. Abbreviations are as follows: Ama: the Amazon, Aus: Australia, Chi: China, EuN: northern Europe, EuW: western Europe, Mex: Mexico, SAf: southern Africa, USA: the United States.

(NAO) have a stronger influence in the NH regions (Hu et al. 2003; Englehart and Douglas 2004; Brönnimann et al. 2007; Riaz et al. 2017; Brunetti and Kutiel 2011; de Beurs et al. 2018; Dong et al. 2019). The dipole mode index or Indian Ocean dipole (IOD), southern annular mode (SAM), North Pacific index (NPI), AMO, and SOI influence SH regions (Mason and Jury 1997; Power et al. 1999; Tyson and Preston-Whyte 2000; Hendon et al. 2007; Fogt et al. 2011; Ashcroft et al. 2014; Lakhraj-Govender and Grab 2018). Annual time series of these natural variability indices were obtained from the

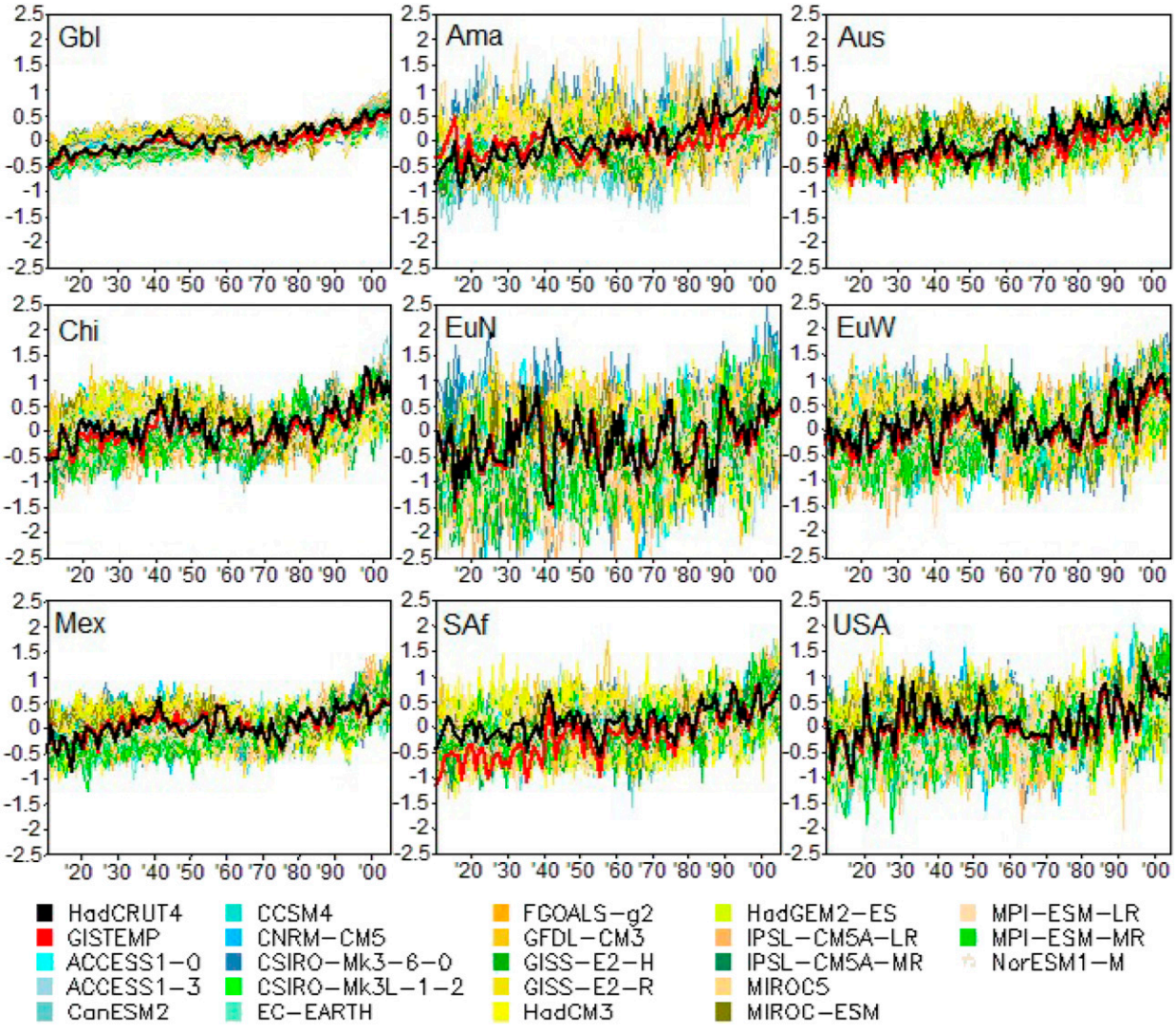


FIG. 2. Annual MST time series for observed (HadCRUT4 and GISTEMP), GCM realizations (light color lines), and GCM ensemble (bold color lines) for nine domains in the 1910–2005 period. Abbreviations are as in Fig. 1.

following sources: AMO (NOAA; Enfield et al. 2001), PDO (JISAO; Mantua et al. 1997), SOI (CRU; Ropelewski and Jones 1987), NAO (CRU; Jones et al. 1997), NAM (NCAR 2019), NPI (Hurrell et al. 2019), IOD (NOAA; Saji and Yamagata 2003), and SAM (MESNZ 2017).

Observed radiative forcing time series were obtained from GISS-NASA (Hansen et al. 2011), commonly used in the literature in the estimation of the transient climate response (Gregory and Forster 2008; Schwartz 2012) and attribution studies (Kaufmann et al. 2011; Estrada et al. 2013b; Pasini et al. 2017; Estrada and Perron 2019). The radiative forcing from well-mixed greenhouse gases, land use change, ozone, stratospheric $H_2O$, aerosols, black carbon, solar irradiance, and snow albedo are aggregated into total radiative forcing (TRF), which summarizes all variables that have a trending behavior (Estrada et al. 2013b). The radiative forcing from stratospheric aerosols (VOLC) is also considered to account for the effects of volcanic eruptions.

### b. *GCM output*

We use 107 realizations of 2-m surface temperature from 21 GCMs included in the CMIP5 historical experiment. We selected GCMs that had at least two realizations (Table 1; Taylor et al. 2012). The multimodel mean, the 21 ensemble means from each GCM, and the 107 individual model runs were included in the analysis. The sample was chosen to match that of the observations (1910–2005). All simulations are available in standard NetCDF format at https://esgf-node.llnl.gov/search/cmip5/ and at https://cera-www.dkrz.de/WDCC/ui/cerasearch/. The spatial resolution of model output varies across GCMs and thus simulations were regridded to two common grids that correspond to those of the HadCRUT4 and GISTEMP

TABLE 1. GCMs included in the CMIP5 historical experiments that had at least two realizations. Resolutions are provided at https://portal.enes.org/data/enes-model-data/cmip5/resolution.

| GCM | Country | Atmospheric resolution (lat × lon) | Realizations |
|---|---|---|---|
| Access1.0 | Australia | 1.25° × 1.875° | 3 |
| Access1.3 | Australia | 1.25° × 1.875° | 3 |
| CanESM2 | Canada | 2.7906° × 2.8125° | 5 |
| CCSM4 | United States | 0.942 406° × 1.25° | 6 |
| CNRM-CM5 | France | 1.4008° × 1.40 625° | 10 |
| CSIRO-Mk3.6.0 | Australia | 1.8653° × 1.875° | 10 |
| CSIRO-Mk3L-1 | Australia | 3.1857° × 5.625° | 3 |
| EC-EARTH | European consortium | 1.1215° × 1.125° | 7 |
| FGOALS | China | 2.7906° × 2.8125° | 3 |
| GFDL CM3 | United States | 2.0° × 2.5° | 5 |
| GISS-E2-H | United States | 2.0° × 2.5° | 6 |
| GISS-E2-R | United States | 2.0° × 2.5° | 6 |
| HadCM3 | United Kingdom | 2.5° × 3.75° | 10 |
| HadGEM2-ES | United Kingdom | 1.25° × 1.875° | 4 |
| IPSL-CM5A-LR | France | 1.894 737° × 3.75° | 6 |
| IPSL-CM5A-MR | France | 1.267 606° × 2.5° | 3 |
| MIROC5 | Japan | 1.4008° × 1.406 25° | 5 |
| MIROC-ESM | Japan | 2.7906° × 2.8125° | 3 |
| MPI-ESM-LR | Germany | 1.8653° × 1.875° | 3 |
| MPI-ESM-MR | Germany | 1.8653° × 1.875° | 3 |
| NorESM1-M | Norway | 1.8947° × 2.5° | 3 |

datasets using bilinear interpolation (Jun et al. 2008). Annual MST anomalies with respect to 1961–90 were obtained using monthly frequency data and the spatial average was calculated for each GCM and region (Fig. 2).

Binary masks of monthly missing/available grid points obtained from each observational dataset (HadCRUT4 and GISTEMP) were applied to GCMs output in order to mimic observed and simulated data. This allows the assessment of GCMs simulations where and when observations are available and thus reduces biases introduced by observational coverage as much as possible (Hegerl et al. 2007; Knutson et al. 2013; Cowtan et al. 2015).

All GCMs included in the CMIP5 produce numerical experiments that are dependent on a set of initial conditions and external forcing scenarios to simulate, for example, past, present, and future climates. The ensemble mean was calculated for each GCM to 1) produce a clearer climate signal since averaging over realizations dampens variability and provides a better representation of the model's response to changes in radiative forcing (Jun et al. 2008; Knutti et al. 2010; Annan and Hargreaves 2011), and 2) reduce the variability in simulations that would otherwise contribute to the error component in any statistical model (Jun et al. 2008; Deser et al. 2014). For calculating the ensemble mean each realization is equally weighted, which can be considered a more transparent strategy to combine GCM outputs (Weigel et al. 2010; Herger et al. 2018), since the only difference between simulations from the same GCM using the same external forcing and physics configuration is the sets of initial conditions. These initial conditions are, for any practical reason, to be considered random (Maraun et al. 2010) and as such there is no reason to assign lower or higher weights to any particular run.

## 3. Methodology

This section is composed of two parts. First, we analyze some of the metrics that are commonly used to evaluate GCM performance, and we show that such metrics are not helpful to discriminate GCMs based on their skill to reproduce observed climate and that they may not be informative for climate change applications. The results motivate the need to develop new metrics that are more robust and adequate for climate change applications. The second part of this section proposes a new methodology to evaluate GCM performance that tackles weaknesses identified in classical metrics and that focuses on the ability of GCMs to reproduce the observed response to changes in external radiative forcing.

### *a. Assessment of classical metrics for evaluating GCM performance*

The process of developing, evaluating, and combining GCM performance metrics is not straightforward. Rather, there is a considerable amount of subjectivity in the selection of metrics and in their interpretation (Christensen et al. 2010). Some of the metrics that have been proposed to evaluate GCMs performance include the magnitude of model biases during the observed period, comparison of trend slope sign and magnitudes, or composites of a large number of model performance diagnostics (Weigel et al. 2010).

To illustrate some of the limitations of these commonly applied metrics, we evaluate the performance of GCMs in reproducing the observed annual MST time series from HadCRUT4 and GISTEMP for the global and USA domains. The metrics chosen for this illustration are the Pearson linear correlation coefficient and the root-mean-square error (RMSE). These metrics were calculated for the annual mean MST series

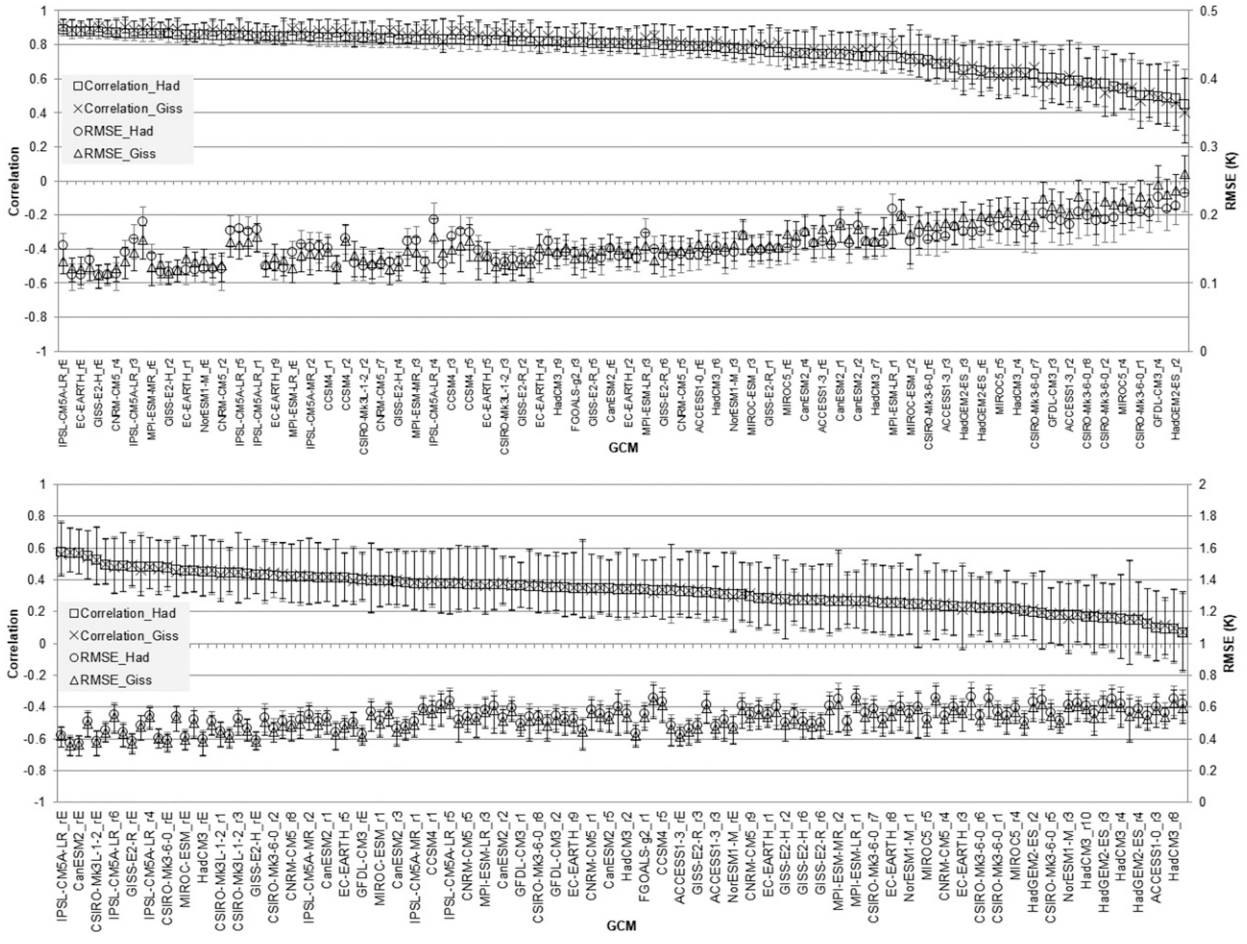


FIG. 3. Correlation and RMSE values between observed (HadCRUT or GISTEMP) and GCM monthly MST series from 1910 to 2005 and 95% confidence intervals for (top) the global domain and (bottom) the USA region. The suffixes r1, . . . , r*n* denote the *n*th realization from the same GCM while rE refers to the ensemble mean from a given GCM. Some of the 129 labels are suppressed to improve the figure's clarity.

from 107 individual realizations of 21 GCMs and for the ensemble mean (rE) of each GCM. As is common in the evaluation of climate models for climate change applications, the trend was not removed because it is the main component of interest and the association between observed and simulated variations around the trend is expected to be close to zero because of the CMIP5 twentieth-century experiment characteristics, which focus on reproducing the response of the climate system to changes in observed radiative forcing; these are ''free-running simulations'' (i.e., with no nudging or data assimilation) and thus internal model variability and observations have no reason to be related (Maraun et al. 2010; Estrada et al. 2012; Deser et al. 2014; Sun et al. 2019).

Figure 3 shows the calculated correlation and RMSE value for all simulations and ensemble means, in decreasing order from highest to lowest correlation values. This figure reveals that 1) these metrics are hardly independent as higher correlation values are associated with lower RMSE values (this is of importance because the lack of independence between metrics is seldom accounted for in practice and can generate reinforcement biases) and 2) ensemble means tend to show higher correlation values than any individual realization; as discussed in the following paragraph, confidence intervals for the correlation coefficient tend to be smaller for models with higher correlation values (i.e., ensemble means).

While the point estimates of these metrics may clearly suggest that some GCMs have better performance than others, the uncertainty in these estimates needs to be accounted for to infer how different these values really are. This is not commonly done when classical metrics are applied, but it should be a standard practice, as it is in other fields. We constructed 95% confidence intervals for the estimated correlation coefficients and the RMSE values by calculating the standard error using the bootstrap method, which is based on resampling with replacement to approximate the empirical distribution of sample estimates (Efron and Tibshirani 1998).

The results show that most of the confidence intervals calculated for the correlation coefficients and RMSE values overlap. Although we show results for only two domains, similar results are found for all other domains. This illustrates

that these metrics do not provide strong statistical evidence for supporting the use of one GCM over another, for ranking models, or for assigning different weights. Based on these metrics, GCMs show a similar skill for simulating the observed MST annual time series. Note that ignoring the confidence intervals, these metrics could lead to very different conclusions regarding model ranking and selection of the "best" GCMs and weights, depending on the particular realization (r1, r2, ...) that is chosen for each model. It is important to remember that realizations from the same model differ only in the initial conditions, and that those sets of values are chosen randomly. As such, using these conventional metrics, without accounting for their uncertainty, could be as effective as randomly choosing a set of models, ranking them, or assigning weights. This example illustrates the lack of robustness of conventional metrics to discern the differences in performance of a set of GCMs and it shows that results are sensitive not only to the selected metric but to the particular realizations that are chosen.

Furthermore, these commonly used metrics offer no information about how well the GCMs can reproduce the climate system response to changes in external radiative forcing. These metrics, like many others, are not meaningful to evaluate the change in climate, as their objective is to compare climatological (static) states, not how they evolve. In addition, most of the classical metrics do not consider the effects of factors such as internal variability and differences in initial states. In particular, low-frequency oscillations can considerably distort trends in climatic variables, either in observations or in GCM simulations (Swanson et al. 2009; Wu et al. 2011), and initial conditions can have similar effects on simulations (Wallace et al. 2015; Deser et al. 2014). Below, we present a methodology based on ordinary least squares regression that focuses on assessing GCM performance in reproducing the response to RF embedded in the observations.

### b. *Performance evaluation based on regression models*

A climatic variable $y_t$, observed or simulated, can be represented as a response to external forcing and natural variability. If a linear functional form is assumed, this can be written as

$$y_t = f(F_t, V_t) = c + \xi(\Delta F_t) + V_t, \tag{1}$$

where $c$ is a reference climatology related to the preindustrial level of the external forcing, $\Delta F_t$ is the change in external forcing, and $V_t$ includes high- and low-frequency climatic oscillations produced by physical variability modes as well as the climate system's persistence. To assess the performance of a particular GCM, an observed variable $y_t^o$ can be expressed as a function of GCM output:

$$y_t^o = f(y_{t,i}^m) = f[\xi^m(F_{t,i}), \eta_{t,i}^m]. \tag{2}$$

Here $y_{t,i}^m$ represents $i$ realizations of a particular $m$ GCM while $\xi^m(F_{t,i})$ and $\eta_{t,i}^m$ are the response to external forcing and model internal variability, respectively. Equation (2) can be approximated by the following regression:

$$y_t^o = \hat{\alpha}_i^m + \hat{\beta} y_{t,i}^m + \varepsilon_t, \tag{3}$$

where $\hat{\alpha}_i^m$ is the bias (difference in means) between observations and GCM realizations, $\hat{\beta}$ is the slope parameter, and $\varepsilon_t$ are the regression's residuals. The residuals of regression (3) can be expressed as

$$\varepsilon_t = \mu_t + \upsilon_t, \tag{4}$$

where $\mu_t$ is the difference in response to RF between observed and simulated variables and $\upsilon_t$ is the difference between observed and modeled internal variability. If the GCM is able to adequately reproduce the observed response, $\mu_t$ should be a stationary variable. Note that $\mu_t$ would contain nonstationarities if either the observed responses to changes in RF, $\xi(\Delta F_t)$, and the simulated response to RF [i.e., $\xi^m(\Delta F_t)$] had different features or if the bias $\hat{\alpha}_i^m$ showed structural changes or changepoints in its value during the sample period. The existence of differences in the trend function between observed and simulated variables can be investigated by means of structural change tests on $\hat{\alpha}_i^m$ and $\hat{\beta}$, and by testing whether the functional form of the regression is linear or not.

Another potential source of nonstationarities is $\upsilon_t$, which contains a variety of oscillations with different frequencies. It has been shown that low-frequency natural variability can distort the trend of climatic variables (Swanson et al. 2009; Wu et al. 2011). Large differences in phase and amplitude of low-frequency oscillations between observed and simulated variability can produce nonstationarities in the residuals of the regression.

To minimize the effects of internal variability, the independent variable $y_{t,i}^m$ in Eq. (3) can be replaced by the ensemble mean of the $n$ available realizations of model $m$:

$$y_{t,i}^m = \overline{y_t^m} = \frac{1}{n}\sum_{i=1}^{n} y_{t,i}^m = c^m + \xi^m(\Delta F_t) + \psi_t^m. \tag{5}$$

Assuming that different realizations of the same GCM are independent from each other, it is expected that as $n$ increases, $\psi_t^m$ would approach a white-noise process. Furthermore, regression (3) can be extended to include the observed variability modes that influence the dependent variable, particularly those related to low-frequency oscillations:

$$y_t^o = \hat{\alpha}^m + \hat{\beta}\overline{y_t^m} + \sum_{j=1}^{k} \hat{\theta}_j X_{t,j} + \sum_{l=1}^{p} \hat{\varphi}_l y_{t-l}^o + \zeta_t. \tag{6}$$

Here, $X_{t,j}$ refers to the $j$th natural variability mode $j = 1, \ldots, k$; also, $\hat{\theta}_j$ are the corresponding coefficients, $y_{t-l}^o$ are lagged values of the dependent variable used to represent the persistence of the observed climatic variable, and $\hat{\varphi}_l$ are the associated parameters. Note that $\sum_{l=1}^{p} \hat{\varphi}_l y_{t-l}^o$ also allows us to correct for autocorrelation in the residuals of the regression (Wooldridge 2013).

Equation (6) can be expressed as $y_t^o - \sum_{j=1}^{k} \hat{\theta}_j X_{t,j} = \hat{\alpha}^m + \hat{\beta}\overline{y_t^m} + \sum_{l=1}^{p} \hat{\varphi}_l y_{t-l}^o + \zeta_t$ and thus the term $\sum_{j=1}^{k} \hat{\theta}_j X_{t,j}$ can be interpreted as a physically based filter that allows for a clearer representation of the observed warming trend. Other techniques, such as polynomial regressions, could be used for accounting for the effects of natural variability oscillations. However, here we favor the use of variability modes that are

known to affect global and regional temperatures and for which some physical mechanisms have been proposed in the literature (see references in section 2a). As described below, the selection of the natural variability terms $X_{t,j}$ to be included in regression (6) is determined by means of an auxiliary regression in which $\mathrm{TRF}_t$ is used instead of $\overline{y_t^m}$. The lag length $p$ is such that the residuals show no autocorrelation according to the Breusch-Godfrey test, as is recommended in the literature to avoid biases in the estimated coefficients (Wilkins 2018; Greene 2012; Keele and Kelly 2006). Note that $\zeta_t$ would be stationary if $\overline{y_t^m}$ adequately represents the observed response to RF, if the ensemble average removes all large nonstationarities produced by internal variability, and if the set $X_{t,j}$ does not omit relevant variables that contain low-frequency oscillations. Note that linear regression models relating observational and simulated variables such as Eq. (6) are used for a wide range of purposes such as in attribution studies and statistical downscaling (Glahn and Lowry 1972; Hegerl and Zwiers 2011; Bindoff et al. 2013).

The evaluation of the performance of a particular GCM is determined by analyzing the regression's coefficients and residuals: an accurate representation of the observed response to RF requires the coefficients $\hat{\beta}$ and $\hat{\alpha}^m$ to be not statistically different from unity and zero, respectively, and that the noise component $\zeta_t$ behaves as a (second order) stationary process. Stationarity of the noise component requires that the coefficients $\hat{\beta}$ and $\hat{\alpha}^m$ are stable and that the regression's functional form is correct. For evaluating the assumptions of linearity in the functional form the Ramsey RESET test (Ramsey 1969) is used and for testing parameter stability we use the Quandt–Andrews structural change test for a break in the trend function occurring at an unknown date (Andrews 1993). It should be noted that even if the assumptions mentioned above are satisfied, $\zeta_t$ could have serial correlation and heteroskedasticity, which may affect significance tests in regression (6). Moreover, bias in the slope coefficients can occur in models with lagged dependent variable terms when autocorrelation is still present in the error component. Even in such a case, the bias effect would be small unless the autocorrelation in the error component is very high (Keele and Kelly 2006). However, as has been shown in the literature, this problem can be avoided by ensuring that the error component is free of autocorrelation and this can be achieved by adding more lagged terms of the dependent variable to avoid the occurrence of the "omitted variables" problem (Keele and Kelly 2006; Wilkins 2018). Robust standard errors help to circumvent heteroskedasticity problems. The reader is referred to Estrada et al. (2013a) for a discussion on evaluating the statistical adequacy of regression models and the use of statistical tests in the context of climate change scenario generation.

The proposed regression approach has important advantages over other methods found in the literature. Among these are 1) the statistical significance of the bias parameter $\hat{\alpha}^m$ and the ability to evaluate restrictions on the value of the parameter $\hat{\beta}$ using standard $t$ tests and a Wald test. It also allows testing joint hypothesis about parameter values and for evaluating differences in parameter (metric) values across GCMs. 2) Furthermore, it takes into account the effects of observed natural variability, as well as the effects of internal variability in GCMs that could distort the observed and simulated response to RF. This allows comparing both responses controlling for the effects of low-frequency oscillations contained in the observed and simulated series. 3) Finally, a variety of statistical tests for analyzing the regression's residuals are available in the literature. These tests can provide empirical evidence about the existence, types, and sources of nonstationarities in $\zeta_t$ and thus help to better evaluate and compare the performance of various GCMs.

The proposed methodology is implemented in two steps. First, an auxiliary regression model based on Eq. (6), $y_t^o = \hat{\alpha}^o + \hat{\beta}^o \mathrm{TRF}_t + \sum_{j=1}^{k} \hat{\theta}_j^o X_{t,j} + \sum_{l=1}^{p} \hat{\varphi}_l^o y_{t-l}^o + \mu_t$, is estimated for observed temperatures using $\mathrm{TRF}_t$ as a proxy for $\Delta F_t$, and a set of $X_{t,j}$ to account for some of the most prominent sources of high- and low-frequency oscillations in $V_t$. For each domain, the best-fit statistically adequate model (i.e., one that satisfies the statistical assumptions of the linear regression model) is selected. For selecting the independent variables to be included in the model, we followed a general approach for empirical modeling based on Spanos (2019). The initial specification of the regression model includes the variables suggested by the literature and then that specification is modified according to the specification problems that are detected by the results of a battery of formal misspecification tests (see last row in Tables A and B and last nine columns in Tables C–T in the online supplemental material available on the Journals Online website at https://doi.org/10.1175/JCLI-D-20-0510.s1). The resulting model contains the independent variables that produce a statistically adequate model (see Estrada et al. 2013a; Estrada and Guerrero 2014). Once such a model specification is found, then the model is reexamined to evaluate if a more parsimonious specification that still satisfies the linear regression assumptions can be obtained by excluding variables that are not significant at a given significance level (e.g., 10% significance level). In the second step, the natural variability terms selected by the auxiliary regression are used to estimate regression (6), in which $\overline{y_t^m}$ replaces $\mathrm{TRF}_t$. The lag length $p$ was selected based on the Breusch–Godfrey test to ensure that the error component is free of autocorrelation to avoid potential biases in the estimated coefficients.

It is important to note that the estimation of coefficients is independent in these two steps, in contrast to other estimation approaches such as two-stage least squares in which the coefficients estimated in the first stage are used in the second stage estimation. Once regression (6) has been estimated, the performance of the GCM is determined by evaluating two things. The first is the similarity of the trend in observed and simulated MST time series by testing the stability of parameter $\hat{\beta}$ and the adequacy of the regression's functional form (linear). If the conditions of parameter stability and correct functional form are not satisfied, there is empirical evidence against the ability of the GCM to adequately reproduce the observed response to RF. The second is if the parameters are stable and the functional form is correct then the values of the estimated parameters are analyzed. If the confidence intervals for $\hat{\beta}$ include the unity (i.e., the estimated parameter value is not different from unity at a given significance level), then the GCM is able to

adequately simulate the observed rate of warming; the GCM overestimates the observed rate of warming if $\hat{\beta} < 1$ (i.e., the estimated parameter value is smaller than unity for a given significance level); conversely, if $\hat{\beta} > 1$ (i.e., the estimated parameter value is greater than unity for a given significance level) the GCM underestimates the observed rate of warming. As mentioned above, this approach allows to formally evaluate other metrics using the same regression, such as the existence of bias. Bias can be formally evaluated by performing a *t* test on $\hat{\alpha}^m$, as well as by testing for a level shift in $\hat{\alpha}^m$ by means of stability tests. Wald tests (Greene 2012) can be used to evaluate the individual or joint significance of parameters such as $\hat{\alpha}^m$ or $\hat{\beta}$ as well as to test if they are statistically equal to a certain value. The application presented in the next section centers on evaluating the ability of GCMs to reproduce the observed response to RF, which we consider the most relevant feature for evaluating models for climate change applications.

## 4. An analysis of GCM performance over global and regional domains

In this section we present an application of the proposed methodology for the global domain and eight subcontinental land regions (Fig. 1). We consider the annual mean surface temperature time series from HadCRUT4, GISTEMP, and the ensembles of simulations produced by 21 GCMs, as well as the multimodel ensemble (see sections 2a and 2b). The regression models include two groups of independent variables: 1) those that approximate the warming trend, namely TRF and the ensemble mean of each GCM, and 2) the set of variables $X_{t,j}$ that include the main natural variability modes (AMO, PDO, SOI, NAM, NAO, NPI, IOD, and SAM; see section 2a), stratospheric aerosols (VOLC), and the persistence of MST, $y^o_{t-1}$.

Following the methodology described in section 3, regression models were estimated for MST. For each domain, a regression model was estimated using TRF and some physically relevant natural variability indices for each region (see section 2a). We note that results are robust to using an alternative radiative forcing dataset (Miller et al. 2014). Explanatory variables were selected based on the global/regional influence reported in the literature (see section 2a). Equation (7) illustrates the final model obtained for global temperatures from the HadCRUT4 and GISTEMP datasets:

$$y^o_t = \hat{\alpha}^m + \hat{\beta}_1 \mathrm{TRF}_t + \hat{\theta}_1 \mathrm{AMO}_t + \hat{\theta}_2 \mathrm{AMO}_{t-1} + \hat{\theta}_3 \mathrm{PDO}_{t-1} + \hat{\theta}_4 \mathrm{SOI}_t + \hat{\theta}_5 \mathrm{VOLC}_t + \sum_{l=1}^{2} \hat{\varphi}_l y^o_{t-l} + \zeta_t \quad (7)$$

where TRF, AMO, PDO, SOI, and VOLC were defined previously, and *t–j* refers to the *j* annual lag of the variable. Figure 4a shows the specification and parameter estimates for all domains included in this study, while parameter values are reported in Tables A and B in the supplemental material. While the selection of variability modes is based on physical considerations, lag length is determined empirically. In cases where the regression model includes lagged dependent terms, the response to external forcing is calculated as $\hat{\beta}_1/(1 - \sum_{l=1}^{p} \hat{\varphi}_l)$, which represents the long-run climate response. The total (long run) coefficients of natural variability modes provided in Fig. 4b help interpreting and comparing the estimated effects with previous (static) estimates. However, most of the estimates of the effects of natural variability modes in the literature have been produced using univariate methods such as correlation and simple linear regression that do not consider possible indirect effects between variability modes. In contrast, the coefficients reported here represent the partial effects of each variable, which account for the effects of the other variables included in the regression. Unless independent variables are uncorrelated, the univariate and multivariate estimates of the magnitude, the significance, and even the sign of the effects can be different. The effect that in the univariate setting may be commonly attributed to a particular natural variability mode can be better represented by a combination of other modes in the multivariate setting (this is sometimes referred to as the omitted variable problem; see Greene 2012).

AMO has a significant influence at the global scale and over most of the domains located in the Northern Hemisphere (Bindoff et al. 2013; Steinman et al. 2015; Guan et al. 2015). In such regions, the positive phase of AMO is associated with higher temperatures and its influence is largest over Europe and North America (Fig. 4b). AMO is characterized by a low-frequency oscillation that has been shown to obscure the warming trend by masking or exacerbating it depending on its phase (Swanson et al. 2009; Wu et al. 2011). PDO and NAO have also been proposed as variability modes that can distort the global warming trend and, as expected, the regression models for the global domain include the AMO, PDO, and SOI, which also have a global effect on temperatures (Guan et al. 2015; Li et al. 2013b; Cohen and Barlow 2005). Figure 4b provides empirical evidence of the effect of observed natural variability modes on MST, the estimated models show that AMO has a significant influence in most of the domains (7 out of 9), followed by SOI, NAO, and PDO. The remaining variability modes have influence over particular regional domains. As shown in the literature, NAO and NAM show relevant effects over regions such as Europe and Mexico (Fig. 4b; Li et al. 2013a; Vihma et al. 2019), while SAM and IOD mainly influence regions in the Southern Hemisphere (Wang and Cai 2013).

Following section 3b, GCM performance is evaluated by replacing $\mathrm{TRF}_t$ with the simulated response to RF, $\overline{y^m_t}$, in the regression models. In all cases, $\overline{y^m_t}$ refers to the ensemble mean from each GCM, as well as to the multimodel ensemble. Note that the CMIP5 is an ''ensemble of opportunity'' and the number of ensemble member available varies depending on the GCM (Taylor et al. 2012). This can affect the results of statistical analyses that are used to compare or combine GCMs, such as model performance evaluation and the generation of probabilistic scenarios (Stephenson et al. 2012; Knutti et al. 2010; Tebaldi and Knutti 2007). As discussed in Notz (2015), this is further complicated by the fact that there is only a single realization of observed climate that not necessarily represents the mean of the data generating process. The following equation for the global domain illustrates the application of regression (6) using the HadCRUT4 and GISTEMP dataset (i.e., both regression models share the same specification):

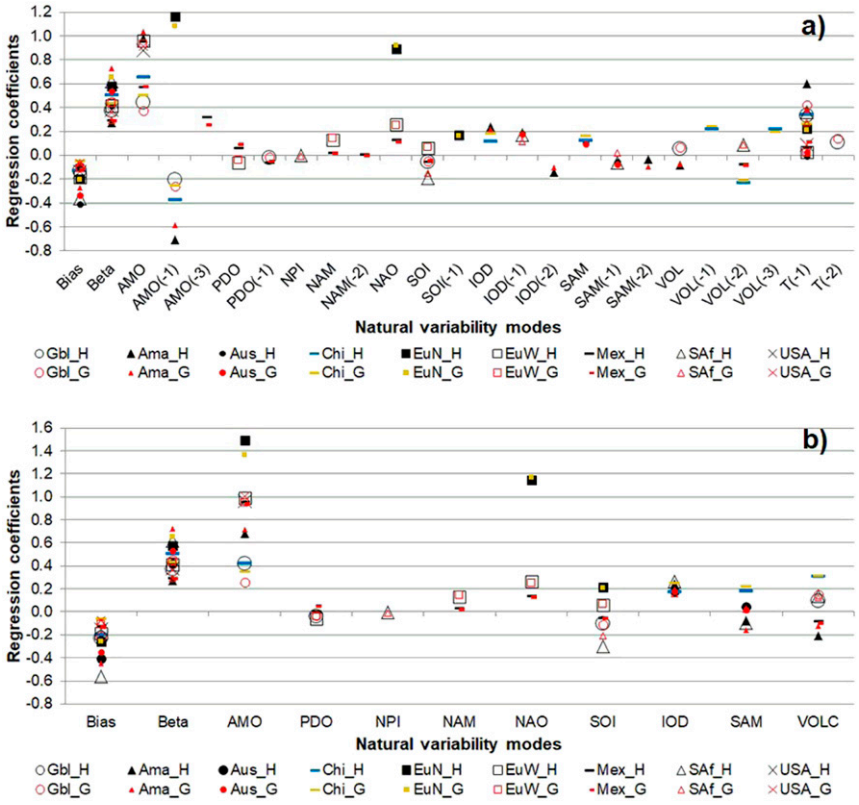


FIG. 4. Estimated coefficient and total (long run) effects from the auxiliary regression $y_t^o = \hat{\alpha}^o + \hat{\beta}^o \mathrm{TRF}_t + \sum_{j=1}^{k} \hat{\theta}_j^o X_{t,j} + \sum_{l=1}^{p} \hat{\varphi}_l^o y_{t-l}^o + \mu_t$, for all domains and observational datasets. (a) Estimated values; the intercept (bias), the long-run response to changes in external forcing, and the individual coefficients of the variables included in the regressions for the nine domains. (b) Estimates of the total (long run) effects of each of the variables included in the regression models. Letters H and G denote the HadCRUT4 and GISTEMP datasets, respectively. Abbreviations are as in Fig. 1.

$$y_t^o = \hat{\alpha}^m + \hat{\beta}_1 \overline{y_t^m} + \hat{\theta}_1 \mathrm{AMO}_t + \hat{\theta}_2 \mathrm{AMO}_{t-1} + \hat{\theta}_3 \mathrm{PDO}_{t-1} + \hat{\theta}_4 \mathrm{SOI}_t + \hat{\theta}_5 \mathrm{VOLC}_t + \sum_{l=1}^{2} \hat{\varphi}_l y_{t-l}^o + \zeta_t. \quad (8)$$

The stability of parameter $\hat{\beta}_1$ and the adequacy of the regression's functional form were evaluated using the Quandt–Andrews and Ramsey REST tests, respectively. These two tests provide empirical evidence to evaluate if the simulated and observed warming trends have similar features and magnitudes. Supplementary Tables C–T report the estimated parameters and results of a battery of other tests for evaluating the statistical adequacy of the regression models for each region and temperature datasets. (These results are summarized in Figs. 5 and 8.)

Figure 5 shows the range of values of the estimated regression parameters for the nine domains, and for the HadCRUT4 and GISTEMP datasets. This figure includes the results for the mean of 21 GCM ensembles, as well as for the multimodel mean. The right panel of Fig. 5 shows only the parameter values of regressions for which parameter $\hat{\beta}_1$ is found to be stable and the assumption of linearity was satisfied. Note that with very few exceptions in the EuN domain, the $\hat{\beta}_1$ coefficients are positive. This figure also shows that, in comparison with Fig. 4b, the parameter magnitudes and signs associated with natural variability modes are similar to those found for the auxiliary regression, which suggests that these results are robust to using $\mathrm{TRF}_t$ or $\overline{y_t^m}$ to represent the climate response to changes in external RF.

If no structural break is present and the functional form is correct, then we test for under/overestimation of the observed warming rate using Wald tests. In the case of regressions that include a number $p$ of lagged terms of the dependent variable to correct for autocorrelation, the $\hat{\beta}_1$ coefficient does not represent the total (long-run) effect but only the instantaneous change. The total effect is given by $\hat{\beta}_1/(1-\sum_{l=1}^{p}\hat{\varphi}_l)$ and thus if the simulated and observed rates of warming are the same then $\hat{\beta}_1/(1-\sum_{l=1}^{p}\hat{\varphi}_l)$ should be statistically equal to 1. The Wald test is formulated as $H_0: \hat{\beta}_1 = (1-\sum_{l=1}^{p}\hat{\varphi}_l)$ and $H_A: \hat{\beta}_1 \neq (1-\sum_{l=1}^{p}\hat{\varphi}_l)$. In most cases $p = 1$, as only the first lagged term of the dependent variable is needed to account for autocorrelation in the error component. These tests were applied to all regions and datasets. The significance level for all tests was set at 5%.

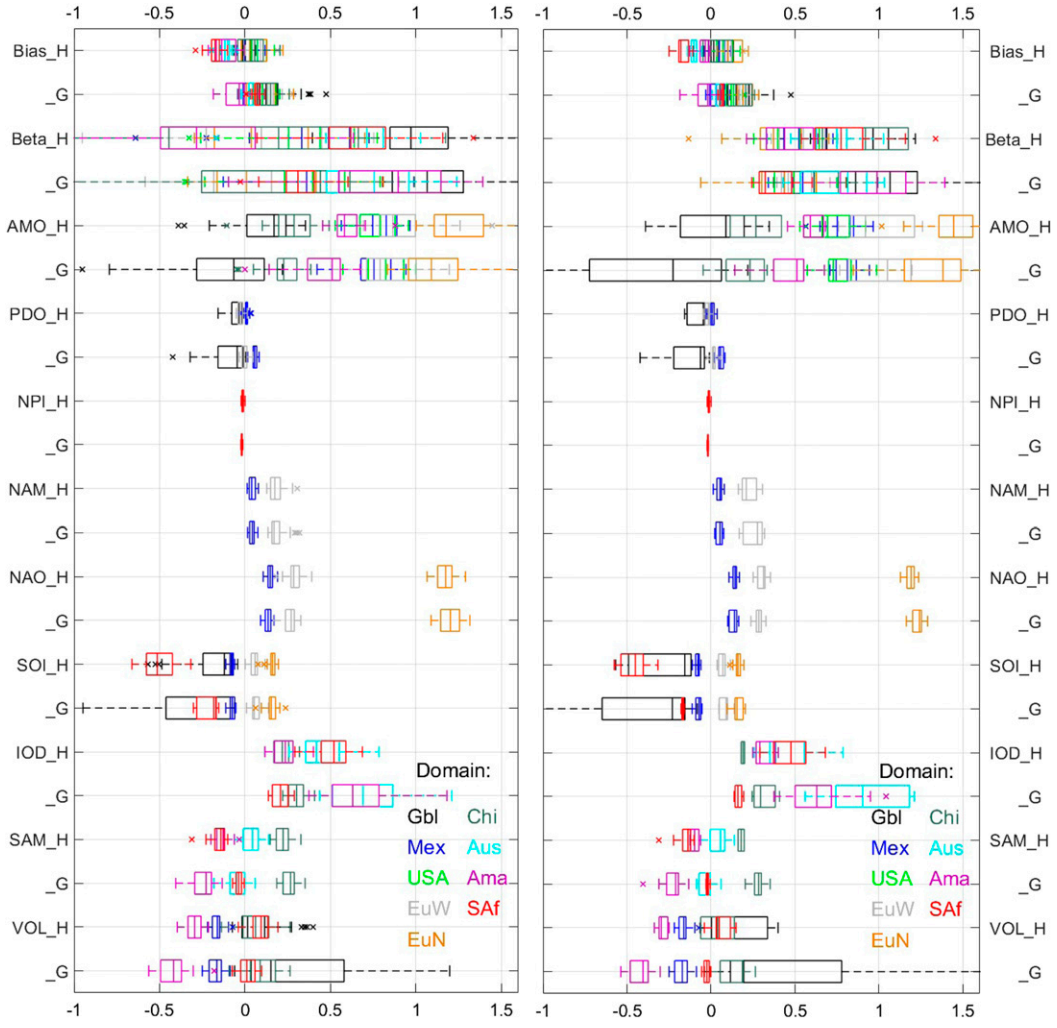


FIG. 5. Estimated total (long run) coefficients from regression (6), $y_t^o = \hat{\alpha}^m + \hat{\beta}\overline{y_t^m} + \sum_{j=1}^{k}\hat{\theta}_j X_{t,j} + \sum_{l=1}^{p}\hat{\varphi}_l y_{t-l}^o + \zeta_t$, for all domains, observational datasets, and GCM ensemble means. (left) For all GCMs, the estimated values, the intercept (bias), the long run response to changes in external forcing, and the long run coefficients of the variables included in the regressions for the nine domains. (right) The same coefficient values as the left panel, but only for those regressions that satisfy the parameter stability and linearity assumptions. Letters H and G denote the HadCRUT4 and GISTEMP datasets, respectively. Abbreviations are as in Fig. 1.

Figure 6 compares the 21 GCM ensemble means (gray lines) and the fitted regression models that satisfied the parameter stability and linearity assumptions (blue lines) as well as those for which $\hat{\beta}_1 = 1$ (black lines), for all regions and CGM ensemble means. Furthermore, the two panels show model fit for the cases where natural variability modes are included in the regressions (Fig. 6a) and when these variables are omitted (Fig. 6b). This figure shows the importance of including this variability modes for improving model fit and illustrates how these modes can modulate the warming trend. This is particularly clear for regions such as the Amazon, Europe, and China for which results suggest that variability modes had an important contribution to the warming rate experienced during the first decade of this century.

Figure 7 illustrates the cases in which 1) regression models satisfy the parameter stability and linearity assumptions and $\hat{\beta}_1 = 1$ (Fig. 7a), 2) the models overestimate the warming trend (Fig. 7b), 3) the parameter stability assumption is not satisfied (Fig. 7c), and 4) the linearity assumption is not met (Fig. 7d). Note that this figure is shown only for illustrative purposes as in most cases visual inspection would not reveal if assumptions are not satisfied or if the observed response to changes in external RF is correctly reproduced by GCMs. Formal statistical tests, such as the ones presented here, provide a much more objective and reliable way to investigate these features. In agreement with previous results (Bindoff et al. 2013), this figure also shows that in all cases natural variability modes are not able to reproduce the observed warming in these regressions, particularly during the last 20 years of the sample, and that this

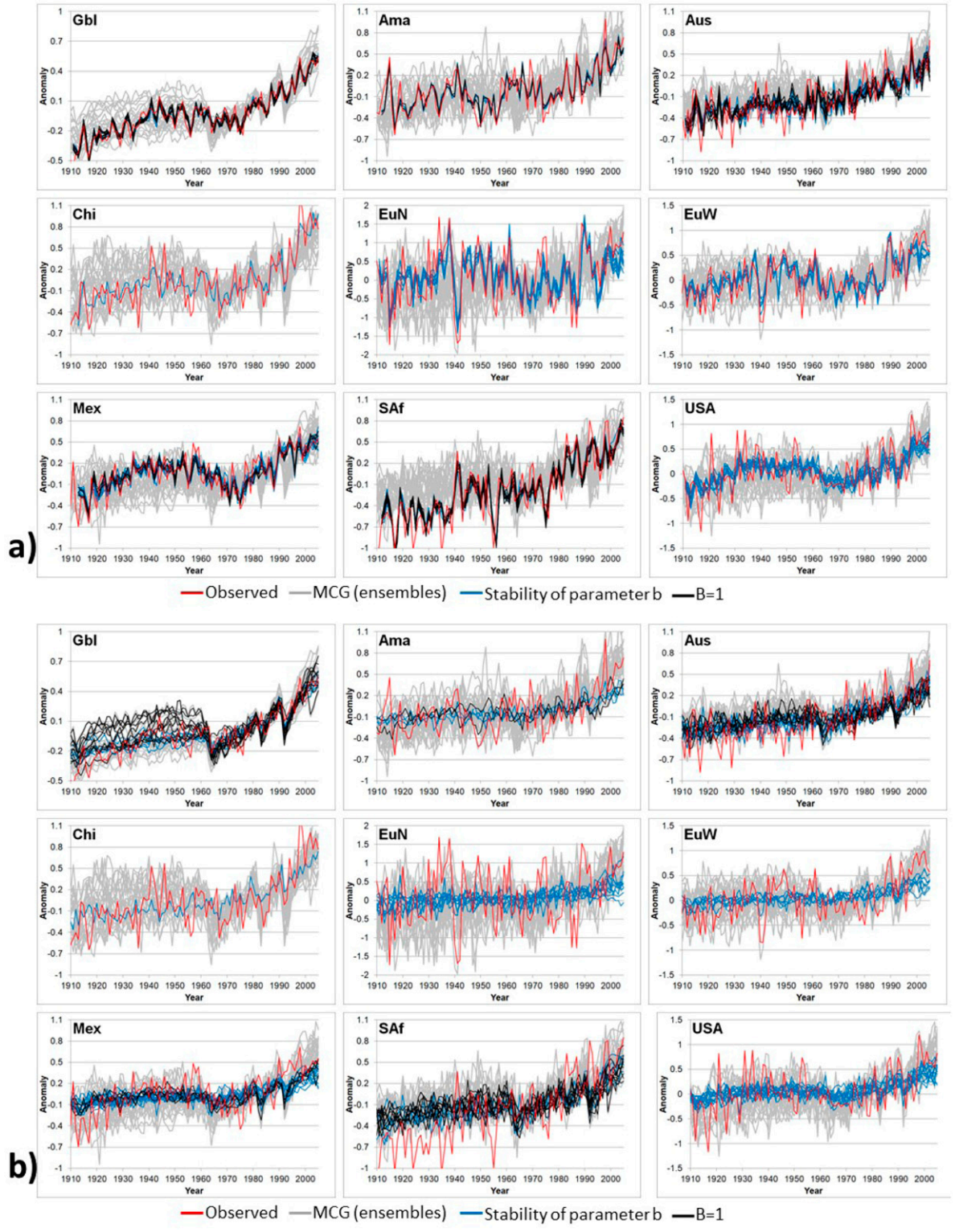


FIG. 6. Fitted temperature series (a) with and (b) without including the effects of natural variability modes. Gray lines show GCM ensemble means, while blue and black lines show the fitted temperature series from regression models that satisfy the parameter stability and linearity assumptions, and those that satisfy these assumptions and for which $\hat{\beta} = 1$, respectively. Abbreviations are as in Fig. 1.

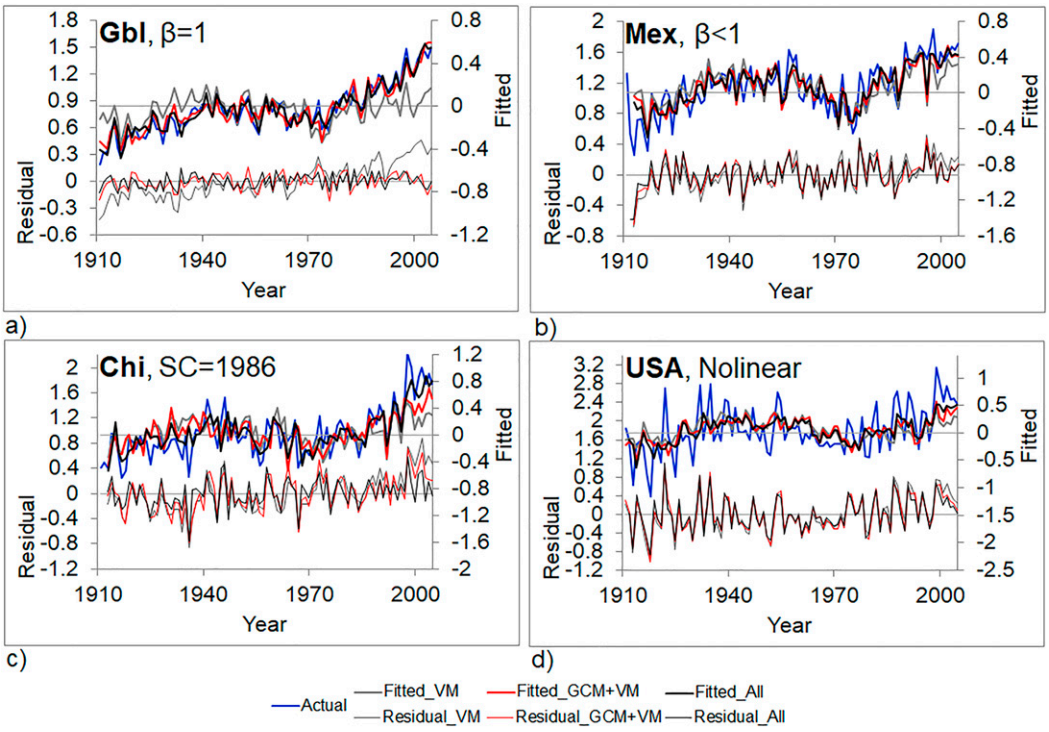


FIG. 7. Illustrative cases of satisfactory GCM performance and when the trend or magnitude of the observed response to external radiative forcing is not correctly reproduced. Three specifications are considered for regression (6) in these examples: 1) only natural variability modes are included in the estimated regressions (Fitted_VM), 2) both natural variability modes and the response to changes in external RF are included (Fitted_GCM+VM), and 3) observed temperature persistence is added to GCM+VM (Fitted_All). The observations are from the HadCRUT4 dataset and the GCM model used is NorESM1-M. SC refers to the estimated date of occurrence of the structural change. Region abbreviations are as in Fig. 1.

is only achieved if the response to changes in external RF are included.

Considering the HadCRUT4 (GISTEMP) dataset, only 12 (8) of the 21 GCMs plus the multimodel ensemble mean, produce regressions in which parameter $\hat{\beta}_1$ is stable and are able to reproduce the observed warming trend in at least 50% of the selected domains (black and gray entries in Fig. 8). GCMs that can successfully reproduce the trend of the observed response to RF in both observed MST datasets for at least half of the domains are CanESM2, EC-EARTH, FGOALS-g2, GISS-E2-H, HadCM3, MPI-ESM-LR, and the multimodel mean. The domains for which at least half of the models can reproduce the warming trend are the global scale (Gbl) and the regions USA, Mex, Aus, SAf, and Ama. In contrast, EuN, EuW, and Chi are the domains for which GCMs more commonly fail to reproduce the characteristics of the warming trend, whether this is due to the presence of structural breaks and/or to incorrect functional form.

It is worth noting that these results depend on the temperature database used, as also happens with traditional metrics. The differences in results between datasets are more common in regions where there are more data gaps. Differences in spatial coverage and temporal continuity, as well as in data and gap-filling processes, can generate disparities in the warming trends contained in each dataset. For instance, GISTEMP tends to show higher warming in most domains during the second part of the twentieth century when compared to HadCRUT4 (Fig. 2). These differences are larger in regions located in the Southern Hemisphere where data coverage is sparser, and smaller in regions with fewer data gaps such as the United States and Europe. Differences in data coverage and quality likely influence results shown in Fig. 8.

The ability of GCM simulations to reproduce the magnitude of the observed warming also varies between regions and depends on the observational database that is used, the ability of current GCMs to adequately simulate the spatial distribution of warming, and factors related to RF. For the majority of GCMs (>50%) that are not able to reproduce the warming trend (i.e., parameter stability is not satisfied; see Fig. 8) this problem occurs in the following domains: EuN and EuW (GISTEMP) and Ama, Chi, EuN, and EuW (HadCRUT4). In such domains, GCMs tend to simulate higher rates of warming than observed.

The lack of agreement between observed and modeled warming rates has been discussed in the literature, and three main hypotheses can be identified: low-frequency natural variability and feedback processes, unaccounted external radiative forcing factors or changes in their rate of growth,

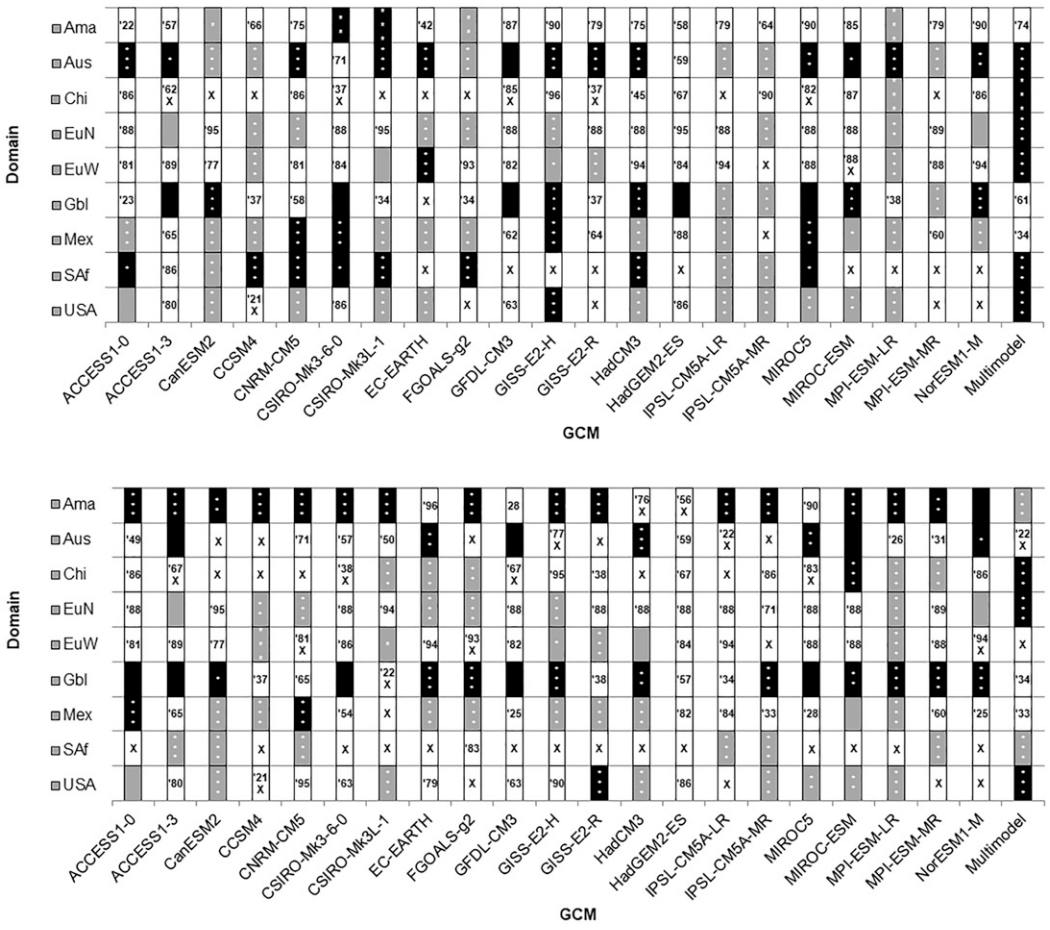


FIG. 8. Wald test results for parameters $\hat{\beta}_1$ in regression (8). The results for the (top) HadCRUT4 and (bottom) GISTEMP annual MST time series. Entries colored in black and gray indicate that the null hypothesis $\hat{\beta}_1 = \left(1 - \sum_{l=1}^{p} \hat{\varphi}_l\right)$ is accepted or rejected, respectively. Entries in white indicate either the presence of a structural break in $\hat{\beta}_1$ or incorrect functional form. Figures indicate the last two digits of the estimated break date, and the character X denotes incorrect functional form. One, two, and three asterisks denote statistical significance of parameter $\hat{\beta}_1$ at the 10%, 5%, and 1% levels, respectively. Abbreviations are as in Fig. 1.

and deficiencies in temperature datasets (see Estrada and Perron 2017). However, the observed slowdown was most likely caused by a combination of multiple factors and cannot be attributed to any particular one. For some of the GCMs, the overestimation of the warming trend starts in the 1970s, but this becomes more pronounced from the mid-1990s onward, as documented by the results of the structural change test that we applied (Fig. 8). This finding is in agreement with what is reported by Fyfe et al. (2013), who conclude that GCMs from CMIP5, with the prescribed forcings, do not reproduce the slowdown from 1998 to 2012. Moreover, most of these GCMs tend to overestimate warming trends during recent decades compared with observations (Kim et al. 2012). This overestimation of the warming trend could be related to the limited ability of GCMs to adequately simulate regional feedback processes such as the Arctic amplification, which became more pronounced since the 1990s, and a variety of local and remote feedback processes related to it (Gillett et al. 2008; Cohen et al. 2019). The existence of unaccounted RF factors or important changes in their rate of growth has also been proposed as a possible explanation for the observed slowdown in the warming trend during the late twentieth century (Estrada et al. 2013b; Steinman et al. 2015).

Eleven out of 22 GCMs are able to reproduce both the trend of the observed response to RF and the magnitude of the warming rate for at least 30% of the selected domains. In the case of the HadCRUT4 dataset, these models are CNRM-CM5, HadCM3, CSIRO-Mk3.6.0, CSIRO-Mk3L-1, GISS-E2-H, MIROC5, and the multimodel mean, while for the GISTEMP dataset these are ACCESS1.0, ACCESS1.3, NorESM1-M, MIROC-ESM, and multimodel mean. The domains for which at least 40% of the GCMs are able to reproduce both the trend of the response to RF and the warming rate of the observational datasets are Gbl, Aus, SAf, and Ama, while for the domains in the Northern Hemisphere most of the

GCMs tend to significantly overestimate the warming rate ($\hat{\beta}_1 < 1$), due to the factors mentioned above.

## 5. Conclusions

In this paper we show that GCM evaluation, selection, and ranking based on classical performance metrics can be misleading and the differences between these metrics can be random and meaningless. This is clearly shown when comparing the metrics' confidence intervals instead of just point estimates. Compared to commonly used metrics, the proposed methodology introduces relevant improvements. It allows us to formally evaluate two of the most relevant aspects for climate change projections: 1) if the trend of the response to RF from a particular GCM is compatible with observations and 2) if the magnitude of the response to RF is similar to that in observations. The proposed method allows us to evaluate the performance of GCMs in reproducing the observed warming trend in a multivariate setting in which the effects of natural variability are accounted for.

The methodology allows us to formally test for these two characteristics and to evaluate the statistical significance of differences between observations and GCMs, as well as between different GCMs. This new approach is based on formal statistical tests that provide empirical evidence that allows us to classify GCMs into groups that 1) are able or unable to adequately reproduce the observed warming trend and 2) under/overestimate or accurately estimate the magnitude of the response to RF. These tests can also be applied to jointly evaluate metrics. Furthermore, confounding factors that may distort the response to RF, such as natural observed variability and the internal GCMs variability, are controlled for. These improvements in GCM evaluation can be of particular importance for applications such as impact, vulnerability, and adaptation assessments and for detecting areas of opportunity to improve current GCMs.

We apply the proposed methodology to nine spatial domains and show that from the GCMs considered (21 models plus the multimodel mean) in this study, only 40% of them are able to reproduce the observed warming in the Gbl, Aus, SAf, and Ama regions, in terms of both its trend and rate of increase. Less than 40% of them are able to reproduce the trend and magnitude of warming in Chi, EuN, EuW, Mex, and USA regions, and most GCMs overestimate the warming rate. While most of the classical performance metrics only provide relative measures of how well GCMs are able to reproduce the observed climate, the proposed method is based on stricter and more informative criteria to discriminate models that can and cannot reproduce two of the most relevant aspects of performance for climate change studies. The proposed method indicates which of these criteria GCMs fail to satisfy. Moreover, the proposed methodology allows us to identify that most of the GCMs tend to overestimate the warming in regions of the Northern Hemisphere, and that these models' simulations tend to show significant discrepancies with the observed magnitude of the warming trend, particularly since the mid-1990s. Several explanations for the reduced warming rate during that period have been proposed (Estrada et al. 2013b; Guan et al. 2015; Steinman et al. 2015; Fyfe et al. 2016; Estrada and Perron 2017), and the lack of fit of models during this period has been discussed in the literature (Dai et al. 2015; Fyfe et al. 2016).

*Acknowledgments.* This study was developed as part of a PhD project in the Postgraduate Program in Earth Sciences of the National Autonomous University of Mexico, with a CONACYT scholarship. The authors are grateful to René Lobato Sánchez, Víctor Manuel Mendoza Castro, and Ignacio Arturo Quintanar Isaias for their helpful feedback and recommendations.

*Data availability statement.* The data that support the findings of this study are available from the corresponding author upon request.

## REFERENCES

Andrews, D., 1993: Tests for parameter instability and structural change with unknown change point. *Econometrica*, **61**, 821–856, https://doi.org/10.2307/2951764.

Annan, J., and J. Hargreaves, 2011: Understanding the CMIP3 multimodel ensemble. *J. Climate*, **24**, 4529–4538, https://doi.org/10.1175/2011JCLI3873.1.

Ashcroft, L., D. Karoly, and J. Gergis, 2014: Southeastern Australian climate variability 1860–2009: A multivariate analysis. *Int. J. Climatol.*, **34**, 1928–1944, https://doi.org/10.1002/joc.3812.

Baumberger, C., R. Knutti, and G. Hirsch-Hadorn, 2017: Building confidence in climate model projections: An analysis of inferences from fit. *Wiley Interdiscip. Rev.: Climate Change*, **8**, e454, https://doi.org/10.1002/wcc.454.

Bindoff, N., and Coauthors, 2013: Detection and attribution of climate change: From global to regional. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 867–952.

Brönnimann, S., E. Xoplaki, C. Casty, A. Pauling, and J. Luterbacher, 2007: ENSO influence on Europe during the last centuries. *Climate Dyn.*, **28**, 181–197, https://doi.org/10.1007/s00382-006-0175-z.

Brunetti, M., and H. Kutiel, 2011: The relevance of the North-Sea Caspian Pattern (NCP) in explaining temperature variability in Europe and the Mediterranean. *Nat. Hazards Earth Syst. Sci.*, **11**, 2881–2888, https://doi.org/10.5194/nhess-11-2881-2011.

Burke, M., S. Hsiang, and E. Miguel, 2015: Global non-linear effect of temperature on economic production. *Nature*, **527**, 235–239, https://doi.org/10.1038/nature15725.

Chan, D., and Q. Wu, 2015: Attributing observed SST trends and subcontinental land warming to anthropogenic forcing during 1979–2005. *J. Climate*, **28**, 3152–3170, https://doi.org/10.1175/JCLI-D-14-00253.1.

Christensen, J., E. Kjellström, F. Giorgi, G. Lenderink, and M. Rummukainen, 2010: Weight assignment in regional climate models. *Climate Res.*, **44**, 179–194, https://doi.org/10.3354/cr00916.

Cohen, J., and M. Barlow, 2005: The NAO, the AO, and global warming: How closely related? *J. Climate*, **18**, 4498–4513, https://doi.org/10.1175/JCLI3530.1.

——, and Coauthors, 2019: Divergent consensuses on Arctic amplification influence on midlatitude severe winter weather. *Nat. Climate Change*, **10**, 20–29, https://doi.org/10.1038/s41558-019-0662-y.

Cowtan, K., and Coauthors, 2015: Robust comparison of climate models with observations using blended land air and ocean sea

surface temperatures. *Geophys. Res. Lett.*, **42**, 6526–6534, https://doi.org/10.1002/2015GL064888.

Dai, A., J. Fyfe, S. Xie, and X. Dai, 2015: Decadal modulation of global surface temperature by internal climate variability. *Nat. Climate Change*, **5**, 555–559, https://doi.org/10.1038/nclimate2605.

de Beurs, K., G. Henebry, B. Owsley, and I. Sokolik, 2018: Large scale climate oscillation impacts on temperature, precipitation and land surface phenology in Central Asia. *Environ. Res. Lett.*, **13**, 065018, https://doi.org/10.1088/1748-9326/aac4d0.

Deser, C., A. S. Phillips, M. A. Alexander, and B. V. Smoliak, 2014: Projecting North American climate over the next 50 years: Uncertainty due to internal variability. *J. Climate*, **27**, 2271–2296, https://doi.org/10.1175/JCLI-D-13-00451.1.

Dong, X., S. Zhang, J. Zhou, J. Cao, L. Jiao, Z. Zhang, and Y. Liu, 2019: Magnitude and frequency of temperature and precipitation extremes and the associated atmospheric circulation patterns in the Yellow River basin (1960–2017), China. *Water*, **11**, 2334, https://doi.org/10.3390/w11112334.

Efron, B., and R. Tibshirani, 1998: *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 436 pp.

Enfield, D., A. Mestas-Nuñez, and P. Trimble, 2001: The Atlantic Multidecadal Oscillation and its relationship to rainfall and river flows in the continental U.S. *Geophys. Res. Lett.*, **28**, 2077–2080, https://doi.org/10.1029/2000GL012745.

Englehart, P. J., and A. V. Douglas, 2004: Characterizing regional-scale variations in monthly and seasonal surface air temperature over Mexico. *Int. J. Climatol.*, **24**, 1897–1909, https://doi.org/10.1002/joc.1117.

Estrada, F., and V. Guerrero, 2014: A new methodology for building local climate change scenarios: A case study of monthly temperature projections for Mexico City. *Atmósfera*, **27**, 429–449, https://doi.org/10.20937/ATM.2014.27.04.08.

——, and P. Perron, 2014: Detection and attribution of climate change through econometric methods. *Bol. Soc. Mat. Mex.*, **20**, 107–136, https://doi.org/10.1007/s40590-014-0009-7.

——, and ——, 2017: Extracting and analyzing the warming trend in global and hemispheric temperatures. *J. Time Ser. Anal.*, **38**, 711–732, https://doi.org/10.1111/jtsa.12246.

——, and ——, 2019: Causality from long-lived radiative forcings to the climate trend. *Ann. N. Y. Acad. Sci.*, **1436**, 195–205, https://doi.org/10.1111/nyas.13923.

——, B. Martínez-López, C. Conde, and C. Gay-García, 2012: The new national climate change documents of Mexico: What do the regional climate change scenarios represent? *Climatic Change*, **110**, 1029–1046, https://doi.org/10.1007/s10584-011-0100-2.

——, V. Guerrero, and C. Gay-García, 2013a: A cautionary note on automated statistical downscaling methods for climate change. *Climatic Change*, **120**, 263–276, https://doi.org/10.1007/s10584-013-0791-7.

——, P. Perron, and B. Martínez-López, 2013b: Statistically derived contributions of diverse human influences to twentieth-century temperature changes. *Nat. Geosci.*, **6**, 1050–1055, https://doi.org/10.1038/ngeo1999.

——, ——, C. Gay-García, and B. Martínez-López, 2013c: A time-series analysis of the 20th century climate simulations produced for the IPCC's fourth assessment report. *PLOS ONE*, **8**, 1–10, https://doi.org/10.1371/journal.pone.0060017.

Fogt, R., D. Bromwich, and K. Hines, 2011: Understanding the SAM influences on the South Pacific–ENSO teleconnection. *Climate Dyn.*, **36**, 1555–1576, https://doi.org/10.1007/s00382-010-0905-0.

Fyfe, J., P. Gillett, and F. Zwiers, 2013: Overestimated global warming over the past 20 years. *Nat. Climate Change*, **3**, 767–769, https://doi.org/10.1038/nclimate1972.

——, and Coauthors, 2016: Making sense of the early-2000s warming slowdown. *Nat. Climate Change*, **6**, 224–228, https://doi.org/10.1038/nclimate2938.

Gillett, N., D. Stone, P. Stott, T. Nozawa, A. Karpechko, G. Hegerl, M. Wehner, and P. Jones, 2008: Attribution of polar warming to human influence. *Nat. Geosci.*, **1**, 750–754, https://doi.org/10.1038/ngeo338.

GISTEMP Team, 2021: GISS surface temperature analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies, accessed 28 January 2019, https://data.giss.nasa.gov/gistemp/.

Glahn, H., and D. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gleckler, P., K. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, https://doi.org/10.1029/2007JD008972.

Greene, W., 2012: *Econometric Analysis.* 7th ed. Prentice Hall, 1239 pp.

Gregory, J., and P. Forster, 2008: Transient climate response estimated from radiative forcing and observed temperature change. *J. Geophys. Res.*, **113**, D23105, https://doi.org/10.1029/2008JD010405.

Guan, X., J. Huang, R. Guo, and P. Lin, 2015: The role of dynamically induced variability in the recent warming trend slowdown over the Northern Hemisphere. *Sci. Rep.*, **5**, 12669, https://doi.org/10.1038/srep12669.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, https://doi.org/10.1029/2010RG000345.

——, M. Sato, P. Kharecha, and K. V. Schuckmann, 2011: Earth's energy imbalance and implications. *Atmos. Chem. Phys.*, **11**, 13 421–13 449, https://doi.org/10.5194/acp-11-13421-2011.

Harvey, D., and T. Mills, 2002: Unit roots and double smooth transitions. *J. Appl. Stat.*, **29**, 675–683, https://doi.org/10.1080/02664760120098739.

Hegerl, G., and F. Zwiers, 2011: Use of models in detection and attribution of climate change. *Wiley Interdiscip. Rev.: Climate Change*, **2**, 570–591, https://doi.org/10.1002/wcc.121.

——, and Coauthors, 2007: Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 663–745.

Hendon, H., D. W. J. Thompson, and M. C. Wheeler, 2007: Australian rainfall and surface temperature variations associated with the Southern Hemisphere annular mode. *J. Climate*, **20**, 2452–2467, https://doi.org/10.1175/JCLI4134.1.

Herger, N., G. Abramowitz, R. Knutti, O. Angélil, K. Lehmann, and B. M. Sanderson, 2018: Selecting a climate model subset to optimise key ensemble properties. *Earth System Dyn.*, **9**, 135–151, https://doi.org/10.5194/esd-9-135-2018.

Hsiang, S., 2016: Climate econometrics. *Annu. Rev. Resour. Econ.*, **8**, 43–75, https://doi.org/10.1146/annurev-resource-100815-095343.

Hu, Z., S. Yang, and R. Wu, 2003: Long-term climate variations in China and global warming signals. *J. Geophys. Res.*, **108**, 4614, https://doi.org/10.1029/2003JD003651.

Hurrell, J., and Coauthors, 2019: The climate data guide: North Pacific (NP) Index by Trenberth and Hurrell; monthly and winter. National Center for Atmospheric Research, accessed 11 February 2020, https://climatedataguide.ucar.edu/climate-data/north-pacific-np-index-trenberth-and-hurrell-monthly-and-winter.

IPCC, 2007: *Climate Change 2007: The Physical Science Basis.* Cambridge University Press, 996 pp.

——, 2013a: *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, 1535 pp.

——, 2013b: Summary for policymakers. *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, 29 pp.

——, 2014: Summary for policymakers. *Climate Change 2014: Impacts, Adaptation and Vulnerability*. Cambridge University Press, 32 pp.

Jones, P., T. Jónsson, and D. Wheeler, 1997: Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and South-West Iceland. *Int. J. Climatol.*, **17**, 1433–1450, https://doi.org/10.1002/(SICI)1097-0088(19971115)17:13<1433::AID-JOC203>3.0.CO;2-P.

Jun, M., R. Knutti, and D. Nychka, 2008: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Stat. Assoc.*, **103**, 934–947, https://doi.org/10.1198/016214507000001265.

Kaufmann, R., H. Kauppi, M. Mann, and J. Stock, 2011: Reconciling anthropogenic climate change with observed temperature 1998–2008. *Proc. Natl. Acad. Sci. USA*, **108**, 11 790–11 793, https://doi.org/10.1073/pnas.1102467108.

Keele, L., and N. Kelly, 2006: Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Polit. Anal.*, **14**, 186–205, https://doi.org/10.1093/pan/mpj006.

Kim, H.-M., P. Webster, and J. Curry, 2012: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. *Geophys. Res. Lett.*, **39**, L10701, https://doi.org/10.1029/2012GL051644.

Knutson, T., F. Zeng, and A. Wittenberg, 2013: Multimodel assessment of regional surface temperature trends: CMIP3 and CMIP5 twentieth-century simulations. *J. Climate*, **26**, 8709–8743, https://doi.org/10.1175/JCLI-D-12-00567.1.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate*, **23**, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1.

——, D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophys. Res. Lett.*, **40**, 1194–1199, https://doi.org/10.1002/grl.50256.

Lakhraj-Govender, R., and S. Grab, 2018: Assessing the impact of El Niño–Southern Oscillation on South African temperatures during austral summer. *Int. J. Climatol.*, **39**, 143–156, https://doi.org/10.1002/joc.5791.

Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.*, **124**, 6307–6326, https://doi.org/10.1029/2018JD029522.

Li, J., C. Sun, and F. Jin, 2013a: NAO implicated as a predictor of Northern Hemisphere mean temperature multidecadal variability. *Geophys. Res. Lett.*, **40**, 5497–5502, https://doi.org/10.1002/2013GL057877.

——, and Coauthors, 2013b: El Niño modulations over the past seven centuries. *Nat. Climate Change*, **3**, 822–826, https://doi.org/10.1038/nclimate1936.

Mantua, N., S. Hare, Y. Zhang, J. Wallace, and R. Francis, 1997: Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.*, **78**, 1069–1079, https://doi.org/10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2.

Maraun, D., and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, **48**, RG3003, https://doi.org/10.1029/2009RG000314.

Mason, S., and M. Jury, 1997: Climatic variability and change over southern Africa: A reflection on underlying processes. *Prog. Phys. Geogr.*, **21**, 23–50, https://doi.org/10.1177/030913339702100103.

Masson, D., and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Lett.*, **38**, L08703, https://doi.org/10.1029/2011GL046864.

Maxino, C., B. McAvaney, A. Pitman, and S. Perkins, 2008: Ranking the AR4 climate models over the Murray-Darling basin using simulated maximum temperature, minimum temperature and precipitation. *Int. J. Climatol.*, **28**, 1097–1112, https://doi.org/10.1002/joc.1612.

MESNZ, 2017: Southern annular mode annual values, 1887–2016. Ministry for the Environment and Statistics New Zealand, accessed 11 February 2020, https://data.mfe.govt.nz/table/89383-southern-annular-mode-annual-values-18872016/metadata/.

Miller, R., and Coauthors, 2014: CMIP5 historical simulations (1850–2012) with GISS ModelE2. *J. Adv. Model. Earth Syst.*, **6**, 441–478, https://doi.org/10.1002/2013MS000266.

Morice, C., J. Kennedy, N. Rayner, and P. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset. *J. Geophys. Res.*, **117**, D08101, https://doi.org/10.1029/2011JD017187.

NCAR, 2019: The climate data guide: Hurrell wintertime SLP-based Northern Annular Mode (NAM) Index. National Center for Atmospheric Research, accessed 28 January 2019, https://climatedataguide.ucar.edu/climate-data/hurrell-wintertime-slp-based-northern-annular-mode-nam-index.

Notz, D., 2015: How well must climate models agree with observations? *Philos. Trans. Roy. Soc.*, **A373**, 20140164, https://doi.org/10.1098/rsta.2014.0164.

Pasini, A., P. Racca, S. Amendola, G. Cartocci, and C. Cassardo, 2017: Attribution of recent temperature behaviour reassessed by a neural network method. *Sci. Rep.*, **7**, 17681, https://doi.org/10.1038/s41598-017-18011-8.

Perkins, S., A. Pitman, N. Holbrook, and J. McAneney, 2007: Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *J. Climate*, **20**, 4356–4376, https://doi.org/10.1175/JCLI4253.1.

Power, S., T. Casey, C. Folland, A. Colman, and V. Mehta, 1999: Inter-decadal modulation of the impact of ENSO on Australia. *Climate Dyn.*, **15**, 319–324, https://doi.org/10.1007/s003820050284.

Qian, C., and X. Zhang, 2015: Human influences on changes in the temperature seasonality in mid- to high-latitude land areas. *J. Climate*, **28**, 5908–5921, https://doi.org/10.1175/JCLI-D-14-00821.1.

Qu, Z., 2011: A test against spurious long memory. *J. Bus. Econ. Stat.*, **29**, 423–438, https://doi.org/10.1198/jbes.2010.09153.

Ramsey, J., 1969: Tests for specification errors in classical linear least squares regression analysis. *J. Roy. Stat. Soc.*, **31**, 350–371, https://www.jstor.org/stable/2984219.

Randall, D., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 589–662.

Riaz, S. M. F., M. J. Iqbal, and S. Hameed, 2017: Impact of the North Atlantic Oscillation on winter climate of Germany. *Tellus*, **69A**, 1406263, https://doi.org/10.1080/16000870.2017.1406263.

Ropelewski, C., and P. Jones, 1987: An extension of the Tahiti-Darwin Southern Oscillation index. *Mon. Wea. Rev.*, **115**, 2161–2165, https://doi.org/10.1175/1520-0493(1987)115<2161:AEOTTS>2.0.CO;2.

Saji, N., and T. Yamagata, 2003: Possible impacts of Indian Ocean Dipole mode events on global climate. *Climate Res.*, **25**, 151–169, https://doi.org/10.3354/cr025151.

Sanderson, B., R. Knutti, and P. Caldwell, 2015: A representative democracy to reduce interdependency in a multimodel ensemble. *J. Climate*, **28**, 5171–5194, https://doi.org/10.1175/JCLI-D-14-00362.1.

Schwartz, S., 2012: Determination of Earth's transient and equilibrium climate sensitivities from observations over the twentieth century: Strong dependence on assumed forcing. *Surv. Geophys.*, **33**, 745–777, https://doi.org/10.1007/s10712-012-9180-4.

Spanos, A., 2019: *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data*. Cambridge University Press, 29 pp.

Steinman, B., M. Mann, and S. Miller, 2015: Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures. *Science*, **347**, 988–991, https://doi.org/10.1126/science.1257856.

Steinschneider, S., R. McCrary, L. Mearns, and C. Brown, 2015: The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning. *Geophys. Res. Lett.*, **42**, 5014–5044, https://doi.org/10.1002/2015GL064529.

Stephenson, D., M. Collins, J. Rougier, and R. Chandler, 2012: Statistical problems in the probabilistic prediction of climate change. *Environmetrics*, **23**, 364–372, https://doi.org/10.1002/env.2153.

Sun, J., K. Zhang, H. Wan, P. Ma, Q. Tang, and S. Zhang, 2019: Impact of nudging strategy on the climate representativeness and hindcast skill of constrained EAMv1 simulations. *J. Adv. Model. Earth Syst.*, **11**, 3911–3933, https://doi.org/10.1029/2019MS001831.

Swanson, K., G. Sugihara, and A. Tsonis, 2009: Long-term natural variability and 20th century climate change. *Proc. Natl. Acad. Sci. USA*, **106**, 16 120–16123, https://doi.org/10.1073/pnas.0908699106.

Taylor, J., and R. Buizza, 2004: A comparison of temperature density forecasts from GARCH and atmospheric models. *J. Forecasting*, **23**, 337–355, https://doi.org/10.1002/for.917.

Taylor, K., R. Stouffer, and G. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.*, **365**, 2053–2075, https://doi.org/10.1098/rsta.2007.2076.

Tol, R., 1996: Autoregressive conditional heteroscedasticity in daily temperature measurements. *Environmetrics*, **7**, 67–75, https://doi.org/10.1002/(SICI)1099-095X(199601)7:1<67::AID-ENV164>3.0.CO;2-D.

——, and A. de Vos, 1993: Greenhouse statistics-time series analysis. *Theor. Appl. Climatol.*, **48**, 63–74, https://doi.org/10.1007/BF00864914.

Tsonis, A. A., K. L. Swanson, G. Sugihara, and P. A. Tsonis, 2009: Climate change and the demise of Minoan civilization. *Climate Past*, **6**, 525–530, https://doi.org/10.5194/cp-6-525-2010.

Tyson, P., and R. Preston-Whyte, 2000: *The Weather and Climate of Southern Africa.* Oxford University Press, 396 pp.

Vihma, T., and Coauthors, 2019: Effects of the tropospheric large-scale circulation on European winter temperatures during the period of amplified Arctic warming. *Int. J. Climatol.*, **40**, 509–529, https://doi.org/10.1002/joc.6225.

Wallace, J., C. Deser, B. Smoliak, and A. Phillips, 2015: Attribution of climate change in the presence of internal variability. *Climate Change: Multidecadal and Beyond*, C.-P. Chang et al., Eds., World Scientific, 1–29.

Wang, G., and W. Cai, 2013: Climate-change impact on the 20th-century relationship between the southern annular mode and global mean temperature. *Sci. Rep.*, **3**, 2039, https://doi.org/10.1038/srep02039.

Weigel, A., R. Knutti, M. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *J. Climate*, **23**, 4175–4191, https://doi.org/10.1175/2010JCLI3594.1.

Wilkins, A., 2018: To lag or not to lag?: Re-evaluating the use of lagged dependent variables in regression analysis. *Political Sci. Res. Methods*, **6**, 393–411, https://doi.org/10.1017/psrm.2017.4.

Wooldridge, J., 2013: *Introductory Econometrics: A Modern Approach*. Cengage Learning, 881 pp.

Wu, Z., N. Huang, J. Wallace, B. Smoliak, and X. Chen, 2011: On the time-varying trend in global-mean surface temperature. *Climate Dyn.*, **37**, 759–773, https://doi.org/10.1007/s00382-011-1128-8.