Robustness of CMIP6 Historical Global Mean Temperature Simulations: Trends, Long-Term Persistence, Autocorrelation, and Distributional Shape

Simon Michael Papalexiou1,2,3, Chandra Rupa Rajulapati1,4, Martyn P. Clark4,5, and Flavio Lehner6

1Global Institute for Water Security, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, 2Department of Civil, Geological and Environmental Engineering, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, 3Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Czech Republic, 4Coldwater Laboratory, University of Saskatchewan, Canmore, Alberta, Canada, 5Centre for Hydrology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, 6Institute for Atmospheric and Climate Science, ETH Zürich, Zürich, Switzerland

Abstract Multi-model climate experiments carried out as part of different phases of the Coupled Model Intercomparison Project (CMIP) are crucial to evaluate past and future climate change. The reliability of models’ simulations is often gauged by their ability to reproduce the historical climate across many time scales. This study compares the global mean surface air temperature from 29 CMIP6 models with observations from three datasets. We examine (1) warming and cooling rates in five subperiods from 1880 to 2014, (2) autocorrelation and long-term persistence, (3) models’ performance based on probabilistic and entropy metrics, and (4) the distributional shape of temperature. All models simulate the observed long-term warming trend from 1880 to 2014. The late twentieth century warming (1975–2014) and the hiatus (1942–1975) are replicated by most models. The post-1998 warming is overestimated in 90% of the simulations. Only six out of 29 models reproduce the observed long-term persistence. All models show differences in distributional shape when compared with observations. Varying performance across metrics reveals the challenge to determine the “best” model. Thus, we argue that models should be selected, based on case-specific metrics, depending on the intended use. Metrics proposed here facilitate a comprehensive assessment for various applications.

1. Introduction

The decade of the 2010s was the hottest yet in more than 150 years of global mean temperature measurements (NOAA, 2020). The key climate change signatures include intensifying the water cycle (Markonis et al., 2019), modifying the frequency of precipitation extremes (Papalexiou & Montanari, 2019), affecting the water availability (Gosling et al., 2011; Miao & Ni, 2009; Sheffield et al., 2012), alternating the regime of heat waves (Papalexiou, AghaKouchak, et al., 2018; Perkins et al., 2012) and flooding (Min et al., 2011; Rahmstorf & Coumou, 2011), and impacts on human health (Gosling et al., 2009), food security (Piao et al., 2010), ecology (Allen et al., 2010; García et al., 2018), and species biodiversity (Nowak, 2010; Wake & Vredenburg, 2008).

The multi-model climate experiments carried out as part of various phases of the Coupled Model Intercomparison Project (CMIP) are designed to advance our understanding and knowledge of climate variability and climate change. These experiments simulate the current climate and project future conditions (IPCC, 2013). Specifically, the CMIP models are forced with multiple scenarios of greenhouse gas concentrations, representing the forced variability (IPCC, 2013). The CMIP models are initialized using multiple initial states; the different initial states result in different trajectories of the climate system, representing the internal or non-forced variability. The previous phases of CMIP have been very useful to both understand changes in the climate system and help inform climate policy (Aloysius et al., 2016; Forster et al., 2013; Knutti & Sedláček, 2013; Li et al., 2014; Miao et al., 2014; Sillmann et al., 2013; Toreti & Naveau, 2015; Wuebbles et al., 2013). The latest phase of CMIP, CMIP6, is developed to overcome limitations of the CMIP5 and continue to address the ongoing scientific challenges (Eyring et al., 2016, 2019).
2. Data and Methods

2.1. CMIP6 and Observations

The monthly mean near-surface air temperature (tas) outputs from 29 models were obtained from the World Climate Research Program (WCRP) CMIP6 (https://esgf-node.llnl.gov/projects/cmip6/). Here we consider historical simulations (Eyring et al., 2016) corresponding to different variants based on realizations (r), initialization (i) schemes, different physics (p), and forcing (f) indexes. We consider a total of 267 simulations from 29 models provided by 19 modeling groups (see Table 1 for details). Historical simulations refer to the 1850–2014 period except for the AWI-CM-1-1-MR model (1851–2014).

To consider the uncertainty in observations, we use three data sets of observed monthly anomalies: (1) GISS Surface Temperature Analysis (GISTEMP v4; https://data.giss.nasa.gov/gistemp), built by combining the NOAA GHCN v4 database for land (meteorological stations), and the ERSST v5 for ocean areas (Hansen et al., 2010; Lenssen et al., 2019); (2) Berkeley Earth Surface Temperatures (Berkeley; http://berkeleyearth.org/data/), a gridded reconstruction of land surface air temperature records spanning 1850-present, obtained by combining the land air temperature with the HadSST3 sea surface temperature reconstruction (Rohde et al., 2014); and (3) HadCRUT4 (https://crudata.uea.ac.uk/cru/data/temperature/), a global temperature data set, obtained by merging the CRUTEM4 and HadSST3 for the land and ocean, respectively (Jones et al., 2012; Osborn & Jones, 2014). To match the periods between the observed data sets and the historical simulation we consider the period 1880–2014, a total of 135 years.

2.2. Methods

In this study we examine (1) warming and cooling rates in several sub-periods, (2) LTP, (3) multi-model performance ranking based on deviance metrics and entropy, and (4) indices that quantify the distributional shape (L-moments). All of the analyses are based on temperature anomalies, obtained as follows: for a given model we (1) calculate anomalies by subtracting the 1951–1980 mean from the historical time series at each grid cell; (2) estimate the surface area of each grid; (3) weigh each grid time series by the grid area over the global area; (4) sum all weighted monthly anomaly time series of all grids; and (5) calculate the annual mean anomalies as average of all months.
Table 1
Information on the CMIP6 Models Used in This Study

| #    | Model name          | Institution                                                                 | No. of grids (Lat × Lon) | # of simulations per model |
|------|---------------------|----------------------------------------------------------------------------|--------------------------|---------------------------|
| 1    | AWI-CM-1-1-MR       | Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Germany | 192 × 384               | 5                         |
| 2    | BCC-CSM2-MR         | Beijing Climate Center, Beijing, China                                      | 160 × 320               | 3                         |
| 3    | BCC-ESM 1           | Beijing Climate Center, Beijing, China                                      | 64 × 128                | 3                         |
| 4    | CAMS-CSM1-0         | Chinese Academy of Meteorological Sciences, Beijing, China                  | 160 × 320               | 3                         |
| 5    | CESM2-FV2           | National Center for Atmospheric Research, Boulder, USA                      | 96 × 144                | 1                         |
| 6    | CESM2-WACCM-FV2     | National Center for Atmospheric Research, Boulder, USA                      | 96 × 144                | 1                         |
| 7    | CESM2-WACCM         | National Center for Atmospheric Research, Boulder, USA                      | 192 × 288               | 3                         |
| 8    | CESM2               | National Center for Atmospheric Research, Boulder, USA                      | 192 × 288               | 11                        |
| 9    | CanESM5             | Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, BC, Canada | 64 × 128               | 50                        |
| 10   | E3SM-1-0            | Lawrence Livermore National Laboratory, Livermore, USA                      | 180 × 360               | 5                         |
| 11   | EC-Earth3-Veg       | Consortium of various institutions from Spain, Italy, Denmark, Finland, Germany, Ireland, Portugal, Netherlands, Norway, the United Kingdom, Belgium, and Sweden | 256 × 512               | 3                         |
| 12   | EC-Earth3           | Consortium of various institutions from Spain, Italy, Denmark, Finland, Germany, Ireland, Portugal, Netherlands, Norway, UK, Belgium, and Sweden | 256 × 512               | 13                        |
| 13   | FGOALS-f3-L         | Chinese Academy of Sciences, Beijing, China                                 | 180 × 288               | 3                         |
| 14   | FGOALS-g3           | Chinese Academy of Sciences, Beijing, China                                 | 80 × 180                | 3                         |
| 15   | GFDL-ESM 4          | National Oceanic and Atmospheric Administration, GFDL, Princeton, USA       | 180 × 288               | 3                         |
| 16   | GISS-E2-1-G-CC      | NASA Goddard Institute for Space Studies, New York, USA                     | 90 × 144                | 3                         |
| 17   | GISS-E2-1-G         | NASA Goddard Institute for Space Studies, New York, USA                     | 90 × 144                | 28                        |
| 18   | GISS-E2-1-H         | NASA Goddard Institute for Space Studies, New York, USA                     | 90 × 144                | 18                        |
| 19   | INM-CM4-8           | Institute for Numerical Mathematics, Russian Academy of Science, Moscow, Russia | 120 × 180              | 1                         |
| 20   | INM-CM5-0           | Institute for Numerical Mathematics, Russian Academy of Science, Moscow, Russia | 120 × 180              | 10                        |
| 21   | IPSL-CM6A-LR        | Institut Pierre Simon Laplace, Paris, France                                | 143 × 144               | 32                        |
| 22   | MCM-UA-1-0          | Department of Geosciences, University of Arizona, Tucson, USA               | 80 × 96                 | 2                         |
| 23   | MIROC6              | Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute, National Institute for Environmental Studies, and RIKEN Center for Computational Science, Japan | 128 × 256              | 10                        |
| 24   | MPI-ESM 1-2-HR      | Max Planck Institute for Meteorology                                         | 192 × 384               | 10                        |
| 25   | MRI-ESM 2-0         | Meteorological Research Institute, Tsukuba, Japan                           | 160 × 320               | 6                         |
| 26   | NESM3               | Nanjing University of Information Science and Technology, Nanjing, China    | 96 × 192                | 5                         |
| 27   | NorCPM1             | NorESM Climate modeling Consortium, Norway                                   | 96 × 144                | 30                        |
| 28   | NorESM2-LM          | NorESM Climate modeling Consortium, Norway                                   | 96 × 144                | 3                         |
| 29   | SAM0-UNICON         | Seoul National University, Seoul, Republic of Korea                         | 192 × 288               | 1                         |

LTP (introduced in the 1940s by Kolmogorov, studied extensively by Mandelbrot, Wallis, and Van Ness in the 1960s, and popularized in geosciences by Beran, 1998) occurs when autocorrelation slowly approaches zero. Such behavior is characterized by persistence around specific values (not necessarily close to the longer-term mean) for extended periods. For example, if global temperature variability is a process with strong persistence, then this could imply that temperature anomalies persist at low or high values for several decades. During such periods, temperature might not be representative of a “true” longer-term climatology, the forced response, or any other longer-term climatological attractor—a challenge inherent to any detection and attribution study. For these reasons, competent models should not only reproduce the probability laws of a process but also its correlation structure (Papalexiou, 2018; Papalexiou, Markonis, et al., 2018; Papalexiou & Serinaldi, 2020) Therefore, the comparison of autocorrelation properties between simulations and observations is crucial. LTP in models and observations is quantified by the Hurst coefficient (H) for the 1880–2014 period; H < 0.5 indicates anti-persistence, H = 0.5 denotes time independence, and as H increases from 0.5 to 1 the persistence also increases. Several methods exist to calculate the LTP (for complete account see Beran, 1998; Beran et al., 2013); here we use the slope of the fitted line in a logarithmic plot between time scale (k) and standard deviation of scaled time series (σ(k)). For a time series {x(t)} the jth value of the scaled time series, at scale k, is given by the following.

\[
x_{j}^{(k)} = \frac{1}{k} \sum_{t=(j-1)k+1}^{j} x_t
\]

(1)

The ACSs from models and observations, used to assess how well models replicate the observed ACS, are estimated by calculating the celebrated Pearson correlation coefficient up to lag 10 years. We rank each
model by calculating the square ratio error, defined as \( \text{SRE} = \sum_{i=1}^{N_m} (\rho_i/\hat{\rho}_i - 1)^2 \), between the simulated ACS and the observed ACS up to a lag of 10 years; models with the smallest and highest SRE are ranked as 1 and 29, respectively.

The deviance and entropic measures used in this study are (1) Model Deviance Criteria (MDC), (2) relative entropy, or the Kullback-Leibler divergence criteria (KLDC) (Kullback & Leibler, 1951), (3) Approximate Entropy (ApEn) (Pincus, 1991), and (4) Sample Entropy (SampEn) (Richman & Moorman, 2000). The MDC calculates the deviance of the inverse cumulative distribution functions (also known as quantile functions) between models and observations by estimating the Root Mean Square Error (RMSE) given by

\[
\text{RMSE} = \left( \sum (Q_o(u) - Q_s(u))^2 \right)^{1/2}
\]

where \( u \) varies from 0.01 to 0.99 in steps of 0.01 and \( Q_o(u) \) and \( Q_s(u) \) denote the quantile values of observed and simulated anomalies for the corresponding \( u \). The lower the RMSE the better the match between observations and simulations.

The KLDC also calculates the difference between the probability distribution of model and observations; the lower the KLDC values the better agreement in the probability distribution of simulations and observations (Shukla et al., 2006). Here we calculate the relative entropy following Delsole and Tippett (2007) and Kleeman (2002), that is,

\[
\text{KLDC} = \left( \log \frac{\sigma_s^2}{\sigma_o^2} + \frac{\sigma_o^2}{\sigma_s^2} - 1 \right) + \left( \frac{(\mu_s - \mu_o)^2}{\sigma_s^2} \right)
\]

where \( \mu \) and \( \sigma \) are mean and variance, respectively, and subscripts \( o \) and \( s \) indicate observations and simulations. This equation can be decomposed into dispersion and signal components; it evaluates both the predictability of the spread (dispersion) and the evolution of the mean (signal) of the model prediction.

The third and fourth criteria are based on entropy measures. Approximate Entropy (Pincus, 1991) (ApEn) and sample entropy (SampEn) are measures of persistence or regularity in a system (time series), thereby assessing the randomness of data without any previous knowledge about the source generating the data set. Low entropy reflects a persistent or a system that is predictable, with patterns (or values) that repeat throughout the series, while high entropy means a stochastic system or independence in the data series. More details on entropy measures, information theory and complexity are found in Delgado-Bonal and Marshak (2019). Let \( x_n, n = 1,2,...,N \), be a time series of length \( N \). For a nonnegative integer \( m \) (also called as the embedded dimension), blocks are defined such that \( u(i) = [x_i, x_{i+1}, ..., x_{i+m-1}] \) and \( u(j) = [x_j, x_{j+1}, ..., x_{j+m-1}] \). The distance between these two blocks is calculated as \( d(u(i), u(j)) = \max_k = 1,2,...,m (|x_{i+k} - x_{j+k}|) \). Each distance is compared with a threshold \( r \) (also called as resolution or tolerance), and the correlation sum \( C^m_m(r) \) is calculated as

\[
C^m_m(r) = \frac{\sum_{j=1}^{N-m} \text{count}(d(u(i), u(j)) \leq r)}{N_m}
\]

where \( \text{count}(d(u(i), u(j)) \leq r) = \) number of \( j \leq N_m \) such that \( d(u(i), u(j)) \leq r \) and \( N_m = N - m + 1 \). This is repeated with all reconstructed blocks, and the probability of a pattern (of length \( m \)) in the time series is obtained by the following.

\[
\phi^m_m(r) = \frac{1}{N_m} \sum_{i=1}^{N_m} \log C^m_m(r)
\]

The ApEn is defined as \( \text{ApEn}(m, r, N) = \phi^m_m(r) - \phi^{m+1}_m(r) \), with \( m \geq 1 \). Low values of ApEn imply that the system is highly repetitive and predictable; in contrast, high values imply randomness and unpredictability. ApEn in some cases might be biased as it strongly depends on the sample size; Sample Entropy (SampEn), however, is more consistent in this respect (Chen et al., 2005; Yentes et al., 2013). The SampEn is defined as...
\[
\text{SampEn}(m, r, N) = -\log\left(\frac{A^{m+1}(r)}{A^m(r)}\right)
\]

where \(A^m(r)\) is the probability that two sequences are similar for \(m\) points and is given by the following.

\[
A^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} A^m_i(r)
\]

\[
A^m_i(r) = \sum_{j=1, j \neq i}^{N-m} \text{count}(d(u(i), u(j)) \leq r)
\]

A typical value for \(m\) is 2, while \(r\) is estimated by identifying the \(r\) resulting in highest entropy (Delgado-Bonal & Marshak, 2019) (see also supporting information Figure S21; here \(r = 0.28\) times the standard deviation of the time series).

Finally, we compare models and observations using L-moment ratios (L-skewness \(\tau_3\) and L-kurtosis \(\tau_4\)), which offer more robust estimates than their product-moments counterparts (skewness and kurtosis); the sample estimate are based on formulas found in the literature (Hosking, 1990; Papalexiou & Koutsoyiannis, 2016; Vogel & Fennessey, 1993). L-ratios allow exploring important statistical properties of the simulated or observed time series and provide information regarding the underlying distribution. While the MDC or KLDC compares the overall distribution, L-moment ratios \(\tau_3, \tau_4\) focus on properties that are more intuitive and conceptually easy to understand. Particularly, L-skewness reveals if the sample has positive or negative skewness showing if the distribution has longer right or left tail, respectively. L-kurtosis shows how heavy the distribution tails are; this is crucial as distribution with heavy (or heavier) tails can generate values that deviate significantly from the mean value.

3. Results

3.1. Comparison of Observed and Simulated Temperature Evolution

The comparison between observed and simulated anomalies from the CMIP6 models, and their multi-model mean (Figure 1), shows general agreement in the temporal evolution of global mean temperatures. The CMIP6 models capture the variability and the overall trend in the observed anomalies over the historical period. Most major episodic cooling and warming periods in observations are also evident in models: intermittent cooling after volcanic eruptions (Flato et al., 2013); the mid-twentieth century plateau during 1941–1950, attributed to sulfate aerosol increases (Meehl et al., 2004) and ENSO activity (Brönnimann, 2005); and the strong warming over the 1964–2004 period, attributed primarily to anthropogenic forcing (Giorgi et al., 2011; IPCC, 2007; Meehl et al., 2004). These features are also visible at monthly and seasonal resolution and are reproduced qualitatively by all models (Figures S1–S16).

3.2. Reproduction of Warming, Cooling, and Hiatus Periods in CMIP6 Simulations

Superimposed on the secular warming trend since 1880 are periods of accelerated and slowed warming relative to the long-term trend. Examples of accelerated warming periods include the early twentieth century and late twentieth century. Examples of slowed warming periods, or “hiatus” periods, include the mid-twentieth century and the 2000s (Li & Baker, 2016; Medhaug et al., 2017; Trenberth & Fasullo, 2013). We evaluate whether models reproduce the warming rates during these periods. The historical time period is divided into four slices following Zhu et al. (2018): (1) a cooling period during the late nineteenth century (1880–1909; Time Slice 1 in Figure 2); (2) a warming phase during the early twentieth century (1909–1942; Time Slice 2); (3) a hiatus during mid-twentieth century (1942–1975; Time Slice 3); and (4) a warming phase in the late twentieth century (1975–2014; Time Slice 4). For each time slice we estimate the warming/cooling rates (slopes of trend lines) from observations, individual model simulations, and the multi-model mean (Figures 2 and S17).

The observed warming (ranging from 0.17°C to 0.18°C per decade for the three observational datasets) during the late twentieth century (Time Slice 4), attributed mainly to anthropogenic activities, is represented fairly well in the multi-model mean (0.24°C per decade). Specifically, the observed rate till 1998 is well simulated by most of the models. We also investigate the much-debated post-1998 warming slowdown in
Figure 1. Observed and CMIP6 simulated anomalies of global mean surface temperature. Shaded area shows the ensemble of models with more than one simulation, and for models with one simulation are shown as thin lines, multi-model mean (thick red line) and historical observations from GISTEMP (thick black line), Berkeley Earth (dashed black line), and HadCRUT4 (dotted black line). Anomalies are estimated for the 1951–1980 baseline period.

Figure 2. Warming, cooling, and hiatus periods in observed (gray) and multi-model mean (orange) global mean temperature anomalies (relative to 1951–1980). Black and red lines show the trend during different time periods from observations and multi-model mean, respectively. Plots on right show warming/cooling rates (°C decade\(^{-1}\)) for individual models during the time slices; black and red lines represent the rate in observations and multi-model mean, respectively. The vertical lines show the range of values for models with more than one ensemble member.
observations and models. The CMIP6 multi-model mean warming rate is 0.23°C per decade, substantially larger than the observed rate of 0.1°C per decade. Only 10% of all model simulations show such a small warming rate. The CMIP5 multi-model mean was 0.2°C per decade over approximately the same time period (Knutson et al., 2016). The hiatus from 1942 to 1975 (Time Slice 3), attributed to a combination of external forcing (anthropogenic aerosols, volcanic eruptions, solar variability) and internal variability, is replicated by most of the models (except AWI-CM-1-1MR, BCC-ESM 1, E3SM 1-0, and NorESM2-LM for which none of the associated simulations [with at least 3] are within the range of observed trends) with near zero trend over the period. For these two periods (Time Slices 3 and 4), the previous versions of CMIP are also in agreement with the observations (IPCC, 2013).

Interestingly, the early twentieth century warming (Time Slice 2) tends to be underestimated by most models. This acceleration, potentially resulting from increase in greenhouse gases, natural forcing, and internal climate variability (Delworth & Knutson, 2000; Hegerl et al., 2018; Meehl et al., 2004), ranges from 0.13°C to 0.15°C per decade for observational data sets, and only 6% of all simulations capture this range. The simulated warming rate of the early twentieth century was also lower in CMIP3 and CMIP5 simulations compared to the observed one (IPCC, 2007, 2013). While CMIP6 on average shows a weak warming trend during the late nineteenth century (Time Slice 1), observations are ambiguous with regard to the trend given their uncertainties for that time period (mostly originating from reduced global coverage); a decreasing rate is reproduced in 20% of the models’ simulations. The observed rates for these two periods are not simulated well by most models in the previous versions of CMIP (IPCC, 2013; Knutson et al., 2016; Zhu et al., 2018).

3.3. Discrepancies in Autocorrelations and Persistence Between Observations and Simulations

We first investigate autocorrelation properties of global temperature anomalies, as LTP is characterized by the temporal decay of ACS. Marked differences are observed among the ACSs of the observed and individual models (Figure 3). Even though the simulations track the observations relatively closely (as seen in sections 3.1 and 3.2), most individual simulations underestimate the observed ACS. We rank each model using SRE, and for models with more than one simulation, we calculate the ensemble ACS by averaging the ACS of all simulations in the ensemble. The four best and worst performing models are shown in Figure 3 (results

![Figure 3](https://example.com/figure3.png)
for all models are presented in Table S1 and Figure S18). Note that there are large differences among the simulations; some perform exceptionally well, and others show unrealistic ACSs. In some models all simulations are consistent in reproducing the observed ACS (e.g., EC-Earth3-Veg and AWI-CM-1-1-MR); however, there are models for which some but not all simulations are consistent (e.g., CanESM5 and EC-Earth3) and also models with no consistent simulations (e.g., MIROC6 and NorESM2-LM). We stress that the autocorrelation properties are affected by the presence of trends (most prominently, the slow decay rate of the ACS). The results in terms of model validation could potentially also be affected if the trends, either among the models or between models and observations, differ significantly; yet, the preceding analysis has shown that these trends are very similar. For models with sufficient ensemble members (Milinski et al., 2019), one could detrend the data with the ensemble mean, but for observations this is not possible without making subjective choices on the trend. Further, we used two available large ensembles with relatively different trends (MPI-ESM and CanESM5) to assess the effect of detrending on the results and find no systematic difference in ACS between models before and after detrending (see Figure S20). We thus choose to compare ACS without detrending.

Next, we investigate the LTP of anomalies as quantified by the scaling of standard deviation versus time scale. LTP is crucial as it governs how slowly the autocorrelation decays with lag or how standard deviation decreases at larger scales which reveals the fluctuation potential at multidecadal scales (Figure 3). For example, for a given annual standard deviation $\sigma$, the decadal $\sigma^H$, assuming LTP with $H = 0.95$, is expected to be 254.8% larger than the decadal standard deviation assuming independence. Observed anomalies show a strong LTP with $H = 0.95$ on an average; model simulations show $H$ is ranging from 0.75 to 0.96 (Figure 3 shows for four best and worst models; see Figure S19 for all models). Results show that there are models reproducing high persistence as seen in observations. This indicates that such models replicate the long-term variability due to several factors such as volcanic, anthropogenic activities, and natural variability. Models reproducing observed autocorrelations are expected to better mimic climate’s temporal dynamics and allow for realistic local (decadal) trends and clustering.

### 3.4. Probabilistic and Entropic Ranking of Climate Models

Since the inception of CMIP, models have been ranked according to their performance in representing aspects of the climate system deemed important for determining the forced response, such as emerging constraints based on past warming, ocean heat uptake, or cloud and radiation characteristics (Brient & Schneider, 2016; Jiménez-de-la-Cuesta & Mauritsen, 2019; Steinacher & Joos, 2016); this practice is considered to lead to more trustworthy projections (Wang et al., 2019). The previous section revealed discrepancies between the CMIP6 simulations in reproducing the observed autocorrelations. Equal-weight multi-model means are widely used; however, some simulations clearly represent observations—in this case global temperature variability—more realistically than others, providing motivation for a performance-based model weighting or ranking.

Here models are ranked by four probabilistic and entropic criteria in comparison with the observations (best performing model is ranked with 1). The first two, the MDC and KLDC, compare the probability distributions between observations and simulations; two time series might have the same MDC or KLDC values but different autocorrelation properties. The other two, ApEn (Pincus, 1991) and SampEn (Richman & Moorman, 2000), are measures of temporal complexity (Delgado-Bonal & Marshak, 2019). They measure the “information flow” by quantifying the regularity or the unpredictability of fluctuations; two time series might have the same ApEn and SampEn values but different probability distributions. Note that most models have many simulations (except for four models having one simulation; see Table 1), and we use three observational data sets; thus, the number of total combinations is of the order of $10^{20}$; thus, we randomly select one simulation of each model and compared with a randomly selected observational data set. We repeat this process thousands of times and eventually calculate the average rank for each model.

Analysis shows that different criteria rank different models as best or worst (Figure 4). For example, based on the MDC and KLDC the best performing models are the NorCPM1 and NESM3, respectively; while based on ApEn and SampEn measures, the best performing models are the GISS-E2-H and SAM0-UNICON. In general, the relative entropy (KLDC) and the MDC methods are consistent with each other; this is expected as both methods compare the observed and simulated anomalies without considering the temporal structure. Likewise, the entropy-based methods (ApEn and SampEn) that focus on temporal patterns are also
ranking models similarly. Since the first two methods have different foci than the last two, their results are not expected to be the same. Six models show contradicting results across the different performance metrics; for example, INM-CM4-8 is considered to be one of the best models according to the relative entropy and MDC methods; yet, ApEn and SampEn rank this model poorly. Perhaps unsurprisingly, no model performs best for all methods; if we consider the lowest average rank from all methods as an overall performance metric then the NorCPM1 would appear as best performing. Other well-performing models with low average ranks are the GISS-E2-1-H, NESM3, MPI-ESM 1-2-HR, INM-CM4-8, and FGOALS-g3.

3.5. Inconsistencies in Distributional Shape

The higher-order L-moment ratios (L-skewness [$\tau_3$], L-kurtosis [$\tau_4$]) quantify the distributional shape properties (see also section 2.2). The metrics $\tau_3$ and $\tau_4$ are measures of skewness and kurtosis and indicate the potential of a process to generate values that differ notably from the mean and appear as extremes. To understand the distributional properties of the anomalies, we estimate the $\tau_3$ and $\tau_4$ in simulations and observations. For models with more than one ensemble member, we calculate the $\tau_3$ and $\tau_4$ metrics by averaging across all members of the ensemble.

The $\tau_3$ and $\tau_4$ values have large variations across the models (Figure 5). Specifically, $\tau_3$ ranges from 0.03 (for CAMS-CSM1-0) to 0.3 (for SAM0-UNICON), indicating symmetric and positively skewed distributions, respectively; $\tau_4$ varies from 0.11 (for EC-Earth3) to 0.24 (for E3SM-1-0) suggesting that some models might have heavier tails than the normal distribution. Interestingly, the observed ($\tau_3, \tau_4$) points for the observations (GISTEMP: [0.184, 0.104]; Berkeley: [0.108, 0.09]; HadCRUT4: [0.122, 0.108]) lie outside the range of...
simulated points in the \( \tau_3 - \tau_4 \) plane (Figure 5a; see Figure S22 for monthly scales). We calculate L-moments for the aggregated time series (aggregated using equation 1), that is, timeseries showing multiyear anomalies up to decadal scale. We note that \( \tau_3 \) and \( \tau_4 \) change with aggregation scale in both the observations and simulations (see Figure 5b, time scales ranging from annual [green] to decadal [blue]). Typically, temperature anomalies are assumed to follow the normal distribution under stationarity (theoretically supported by central limit theorem); yet, the noted variation could be explained by nonstationarity and local trends, such as external forcings discussed in section 1. Some models have the same \( \tau_3 \) as the observations, indicating similar skewness; yet, they do not match the observations’ \( \tau_4 \), specifically at low scales (green color; Figure 5b). Investigation of the \((\tau_3, \tau_4)\) points in the \( \tau_3 - \tau_4 \) plane (at the aggregated scales) shows a clear clustering of the observation-based points (triangles) and mean of all models (circles). However, some individual models (given by the range around the multi-model mean estimate in Figure 5b) match well the observations of GISTEMP.

4. Conclusions

Climate models are developed to help advance our knowledge and understanding of climate variability and climate change. A rigorous evaluation of model simulations reveals their credibility. However, performance assessment depends on the evaluation criteria used; for example, typically, no single model is top-performing in all criteria. More specifically, models are assessed by comparing their historical simulations with observations and focusing on properties such as the mean, standard deviation, or the deviance between probability density functions. However, other distributional measures such as shape characteristics, and important joint distribution properties such as the autocorrelation and LTP at higher scales, are not generally examined.

Here, an extensive assessment of global mean surface air temperature is presented. Studying changes in global mean surface temperature is important for analyzing global response of various natural, managed, and human systems to warming rate (IPCC, 2018). The IPCC Special Report 15 (SR15) highlights the impacts of 1.5°C global warming above preindustrial levels and illustrates the impacts and risks of different levels of global warming for people, economies, and ecosystems across different sectors and regions. Given the importance of studying the changes in global mean surface temperature, and the ability to simulate historical and future trends by climate models, the simulations have to be assessed with observed ones in a comprehensive manner. We therefore evaluate the global mean surface temperature from the sixth phase of the Coupled Model Intercomparison Project (CMIP6) in comparison...
with three observation data sets (GISTEMP, Berkeley and HadCRUT4) at annual and monthly scales. We use 267 simulations from 29 models provided by 19 modeling groups and assess the proximity of each model with three observation data sets using performance measures and robust criteria in terms of distributional properties. Particularly, we focus on (1) warming and cooling trends, (2) ACS and LTP, (3) probabilistic and entropic measures, and (4) distributional shape characteristics as quantified by L-moments.

1. The overall historical warming is captured by most of the models and by the multi-model mean. The rapid warming in the observed anomalies during the late twentieth century and the warming hiatus from 1942 to 1975 is well replicated by most of the models, confirming the important role of external forcing for producing these variations around the secular warming trend. Also, between 1998 and 2014, the observed trend is smaller (0.1°C per decade) than most of the models (90% of simulations overestimated this trend).
2. Only six out of 29 models reproduce LTP accurately. The ACS decays slowly, which is primarily due to the long-term trend; yet even so, it is generally underestimated by most models.
3. The models were ranked, from best to worst, using the MDC and three entropy-based criteria. No single model performs best in all four criteria, leaving some subjective choices up to the model users when determining the feasibility of models for a given application.
4. The simulated L-skewness and L-kurtosis are found to be higher in most models compared to observational estimates. This indicates that improvement is needed to capture the distributional shape characteristics. The current tendency of models to produce higher skewness could be related to their tendency to warm more than observations over the late twentieth century (Tokarska et al., 2020). Their tendency for higher kurtosis is harder to diagnose. It could be related to a combination of lack of low frequency variability paired with a tendency to overestimate the cooling after volcanic eruptions, although the latter issue could well be due to the exceptional coincidence of El Nino events during observed volcanic eruptions during the twentieth century (Lehner et al., 2016).

This study offers a detailed comparison on CMIP6 historical global temperature simulations. It provides additional metrics and performance assessments of the global temperature anomalies and reveals the challenge to determine the “best” model. Rather, it appears that models need to be chosen for a purpose, and their performance should be assessed based on that purpose. These metrics and assessment criteria are not limited to temperature and can be used for any hydro-meteorological variable (such as precipitation and wind speed) and at any spatial and temporal scales. Many real-world applications might demand accurate reproduction of the probability distribution—that should be the case if extremes are investigated. Other cases may need to focus on the temporal evolution of time series that can affect the clustering of high or low values. The fact is that informed decisions and model selection should not only be based on the typically used model comparison measures, such as mean and variance, but go beyond. Here we propose alternative measures that we deem will prove valuable in comparing and selecting simulations for actual applications; that could be risk assessment studies using temperature, precipitation, and so forth, at local or regional scales, or studies aiming to develop climate change adaptation strategies. As in all model evaluation efforts, uncertainties in observations themselves have to be considered carefully.

Conflict of Interest
The authors declare no competing financial interests.

Data Availability Statement
All CMIP6 model simulations and observed data are available online (https://esgf-node.llnl.gov/projects/cmip6, https://data.giss.nasa.gov/gistemp, http://berkeleyearth.org/data, and https://crudata.uea.ac.uk/cru/data/temperature).

References
Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., et al. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. Forest Ecology and Management, 259(4), 660–684. https://doi.org/10.1016/j.foreco.2009.09.003
Stott, P. A., & Forest, C. E. (2007). Ensemble climate predictions using climate models and observational constraints. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365*(1857), 2029–2052. https://doi.org/10.1098/rsta.2007.2075

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances, 6*(12), eaaz9549. https://doi.org/10.1126/sciadv.aaz9549

Toreti, A., & Naveau, P. (2015). On the evaluation of climate model simulated precipitation extremes. *Environmental Research Letters, 10*(1), 014012. https://doi.org/10.1088/1748-9326/10/1/014012

Trenberth, K. E., & Fasullo, J. T. (2013). An apparent hiatus in global warming? *Earth’s Future, 1*, 19–32. https://doi.org/10.1002/2013EF000165

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances, 6*(12), eaaz9549. https://doi.org/10.1126/sciadv.aaz9549

Toreti, A., & Naveau, P. (2015). On the evaluation of climate model simulated precipitation extremes. *Environmental Research Letters, 10*(1), 014012. https://doi.org/10.1088/1748-9326/10/1/014012

Trenberth, K. E., & Fasullo, J. T. (2013). An apparent hiatus in global warming? *Earth’s Future, 1*, 19–32. https://doi.org/10.1002/2013EF000165

Vogel, R. M., & Fennessey, N. M. (1993). L moment diagrams should replace product moment diagrams. *Water Resources Research, 29*(6), 1745–1752. https://doi.org/10.1029/93wr00341

Wake, D. B., & Vredenburg, V. T. (2008). Are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proceedings of the National Academy of Sciences, 105*(Supplement 1), 11,466–11,473. https://doi.org/10.1073/pnas.0801921105

Wang, H.-M., Chen, J., Xu, C.-Y., Chen, H., Guo, S., Xie, P., & Li, X. (2019). Does the weighting of climate simulations result in a better quantification of hydrological impacts? *Hydrology and Earth System Sciences, 23*(10), 4033–4050. https://doi.org/10.5194/hess-23-4033-2019

Wuebbles, D., Meehl, G., Hayhoe, K., Karl, T. R., Kunkel, K., Santer, B., et al. (2013). CMIP5 climate model analyses: Climate extremes in the United States. *Bulletin of the American Meteorological Society, 95*(4), 571–583. https://doi.org/10.1175/BAMS-D-12-00172.1

Yentes, J. M., Hunt, N., Schmid, K. K., Kaipust, J. P., McGrath, D., & Stergiou, N. (2013). The appropriate use of approximate entropy and sample entropy with short data sets. *Annals of Biomedical Engineering, 41*(2), 349–365. https://doi.org/10.1007/s10439-012-0668-3

Zhu, X., Dong, W., Wei, Z., Guo, Y., Gao, X., Wen, X., et al. (2018). Multi-decadal evolution characteristics of global surface temperature anomaly data shown by observation and CMIP5 models. *International Journal of Climatology, 38*(3), 1533–1542. https://doi.org/10.1002/joc.5264