Evaluation of CMIP5 Ability to Reproduce 20th Century Regional Trends in Surface Air Temperature and Precipitation over CONUS

Jinny Lee[1], Duane Waliser[2], Huikyo Lee[2], Paul Loikith[3], Kenneth E. Kunkel[4]

[1]Department of Civil and Environmental Engineering, Center for Hydrometeorology and Remote Sensing, University of California, Irvine, CA

[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

[3]Department of Geography, Portland State University, Portland, OR

[4]Cooperative Institute for Climate and Satellites, NCSU and NOAA, Asheville, NC

May 8, 2019

Submitted to Climate Dynamics

Corresponding Author: Jinny Lee, jinny.lee@uci.edu.

Abstract

The ability of the 5th phase of the Coupled Model Intercomparison Project (CMIP5) to reproduce 20th-century climate trends over the seven CONUS regions of the National Climate Assessment (NCA) is evaluated. This evaluation is carried out for summer and winter for three time periods, 1895-1939, 1940-1979, and 1980-2005. The evaluation includes all 206 CMIP5 historical simulations from 48 unique models and their multi-model ensemble (MME), as well as a gridded in-situ dataset of surface air temperature and precipitation. Analysis is performed on both individual members and the MME, and considers reproducing the correct sign of the trends by the members as well as reproducing the trend values. While the MME exhibits some trend bias in most cases, it reproduces historical temperature trends with reasonable fidelity for summer for all time periods and all regions, including at the CONUS scale, except the Northern Great Plains from 1895-1939 and Southeast during 1980-2005. Likewise, for DJF, the MME reproduces historical temperature trends across all time periods over all regions, including at the CONUS scale, except the Southeast from 1895-1939 and the Midwest during 1940-1979. Model skill was highest across all of the seven regions during JJA and DJF for the 1980-2005 period. The quantiatively best result is seen during DJF in the Southwest region with at least 74% of the ensemble members correctly reproducing the observed trend across all of the time periods. No clear trends in MME precipitation were identified at these scales due to high model precipitation variability.

KEYWORDS: CMIP5, model evaluation, surface air temperature, multi-model ensemble

1. INTRODUCTION

An emerging urgence in climate change studies is the investigation of regional–scale variability and trends of key climate variables such as surface air temperature and precipitation (IPCC, 2013; Melillo et al., 2014). Given that regional topography and surface characteristics are influential in how the regional climate responds to both global climate change influences as well as local projections of large-scale natural variability, it is essential that these features are examined in our global climate models (Kunkel and Liang 2005). Continued analysis of long-term observed trends and low-frequency variability in key climate variables is instrumental to understanding the characteristics of the observed changes, and in combination with validated models, forcing data and model responses, better quantify attribution.

The Intergovernmental Panel on Climate Change Fifth Assessment Report (IPCC AR5) concludes that much of the warming that is experienced globally since the 1950s is extremely likely to have been caused by an increase in anthropogenic greenhouse gas concentrations in the atmosphere. Future global climate projections, based on analysis of GCM projections suggest an increase in globally-averaged surface temperature by at least 2°C by 2100 (IPCC, 2013) with consequential implications. As just one example, the increased temperature will lead to substantial melting of glaciers and ice sheets in polar regions and a rise in sea level. Several studies have been conducted on global temperature changes (e.g. Crowley (2000), Knutson et al., (2006), Knutson et al., (2013)), all of which agree that a majority of regions across the globe have experienced warming over the last century that beyond their natural climate variability.

The National Climate Assessment Climate Science Special Report (CSSR; Wuebbles et al. 2017) states that the United States experienced an average warming of 1.2-1.8°F since 1895. However, this warming is not uniform over the entire country nor is it constant. Much of the northern parts of the nation have experienced more warming whereas the Southeast has seen small increases in temperature. In addition, most of the warming has occurred since 1970. The NCA3 report further suggests that future climate projections, based on the results of 16 GCMs, show an additional increase in temperature of 2-4°F over the next few decades with 3-5°F increase by the end of this century under a lower emissions scenario and 5-10°F under a higher emissions scenario.

The validity of these future climate projections weighs heavily on the accuracy of the historical simulations. There have been several studies which utilize GCMs to examine changes in the temperature and precipitation over a specific region in the U.S. (e.g. Kunkel and Liang (2005), Barnett et al., (2008), Cayan et al., (2013)) or to project regional climate change (e.g. Meehl and Tebaldi (2004), Kunkel et al., (2013), Wuebbles et al., (2014)). Similarly, another study has utilized Regional Climate Models (RCMs) to evaluate past climate trends in North America (e.g. Bukovsky, 2012). One key finding in the evaluations of observed change over the contiguous United States (CONUS) is the lack of warming in the Southeast over the second half of the twentieth century. This feature, termed the "warming hole," is well studied (Kunkel et al., (2006), Meehl et al., (2012), Kumar et al., (2013)) and research suggests that models that are more skillful in simulating North Atlantic sea surface temperatures are better at reproducing the "warming hole," but the number of models that actually do so are small.

The purpose of this research study is to evaluate the ability of the CMIP5 simulations and their ensemble mean to reproduce, with regional fidelity, the recent historical near surface air temperature and precipitation trends over the continental US. This includes discriminating across the seven NCA regions across CONUS, considering 3 multi-decade periods, and considers model fidelity of trends both in terms of reproducing the correct sign as well as the magnitudes of the trends. Past studies have considered a subset of this objective using only a portion of the available simulations participating in the fifth phase of the Coupled Model Intercomparison Project (CMIP5, Taylor et al., 2012) against observed data. For example, Janssen et al., (2016) used 94 available simulations to examine regional extreme precipitation events in the United States. Kumar et al., (2013) analyzed 196 simulations to find a relationship between the "warming hole" and natural climate variability.

Temperature and precipitation were chosen as the key climate variables for this study in part because an enhanced anthropogenic greenhouse effect will result in warming temperatures as a first order impact and a warmer atmosphere will alter precipitation intensity and patterns due to the thermodynamic relationship between temperature and precipitation. Secondly, temperature and precipitation change will have/are having profound impacts on society and the environment. While there are many other important variables that climate change will alter, temperature and precipitation are arguably the most important for impacts and therefore the most critical for climate models to capture trends in.

This study goes beyond previous studies' objectives and approaches by utilizing the entire CMIP5 historical suite available (up to 206 simulations), considering regional trends across all of CONUS, and considering fidelity of both signs and magnitudes of regional trends. Moreover, this study utilizes a new long-record in-situ based time series for its observational reference (See Section 2). By performing a comprehensive analysis of the model fidelity in reproducing recent regional historical trends along the lines described above, we aim to yield additional quantitative information on the uncertainty associated with the multi-model ensemble projections of future U.S. climate, specifically tailored to the seven regions over CONUS defined by the United States Global Change Research Program (USGCRP) NCA. While the recent NCA Climate Science

Special Report (CSSR) highlights observed trends in and simulated future projections of temperature and precipitation over CONUS, a first order evaluation of the fidelity with which the climate models used in the report are able simulate historical temperature and precipitation trends is not included (Easterling et al. 2017). Therefore, this comprehensive evaluation of these trends gives context to the uncertainty associated with future projections of temperature and precipitation across CONUS as presented in the NCA CSSR.

The remainder of this paper is structured as follows. Section 2 describes the data and methodology used in our evaluation of simulated trends. Section 3 presents comparisons between the observed and simulated trends in temperature and precipitation. Section 4 summarizes our key findings. Section 5 entails the conclusions of our study.

2. DATA AND METHODOLOGY

a. Data

CMIP5 provides a state-of-the-art multi-model dataset produced by climate modeling groups in an effort to further our knowledge of climate variability and climate change (Taylor et al. 2012). The simulations of the CMIP5 GCMs estimate future climate change and provide a scientific basis for the IPCC AR5 (IPCC, 2013). To evaluate the models' capability to reproduce observed trends in surface air temperature and precipitation, we analyzed the historical simulations of the 20th century with time-evolving external forcings. There are 21 different modeling centers from 14 different countries that have contributed a total of 206 historical simulations with spatial resolutions varying between 0.5° to 4° (Table 1). The specific simulation period varies across the models, but most of the models provide output for the 111 years between 1895 and 2005. The 206 simulations are made up of contributions from 48 unique GCMs, with some models contributing ensembles as small as one simulation or as large as 25 simulations that differ in initial conditions and/or parameterizations. The historical simulations are driven by various forcings that attempt to be, but are not entirely, identical across all modeling groups (Kunkel et al. 2006), and these differences may account for some differences in the model behavior. The multi-model mean temperature and precipitation trends of the CMIP5 suite are used to examine the ensemble as a whole, as previous research has shown that this approach can capture the observed temperature and precipitation trends well for some regions, time periods, and ensemble constructions (e.g. Kumaret al., 2013). In addition, in parts of the current study, each of the 206 simulations is treated as if they are independent as done in Kunkel et al.'s (2006) analysis of central United States temperature trends in the twentieth-century. The present study used monthly mean near-surface air temperature and monthly total precipitation from the 206 historical simulations. Detailed information about the models used in this study is presented in the Appendix.

This analysis uses observed monthly mean near-surface air temperature and monthly total precipitation data as reference data to compare to those of the GCM simulations. The reference dataset is a new National Center for Environmental Information (NCEI) climate monitoring product based on the Global Historical Climatology Network-Daily (GHCN-Daily) dataset (Menne et al., 2012), referred to as nClimGrid, hereafter. nClimGrid includes monthly averaged minimum, maximum, and mean near-surface air temperature, as well as total monthly precipitation data. Data availability spans from 1895 to present over CONUS and Alaska. nClimGrid primarily utilizes the Cooperative Observer (COOP) program to which area-weighted averages of grid points are used to interpolate latitude-longitude spacing of 5 km (Vose et al. 2014; Kunkel et al. 2015). This new gridded dataset is distinct in two significant ways: it uses a larger station network with greatly improved spatial coverage and density, particularly over the West; and the interpolation process accounts for topographic effects (e.g., from mountain ranges, temperature inversions, and coastal influences), meaning the resultant grids depict actual temperatures rather than climate anomalies. Bias adjustments were computed to account for historical changes in observation time as well as historical changes in station location, temperature instrumentation, and siting conditions. Because of these substantial changes to the official temperature record of the United States, a new evaluation of model simulation of historical trends is warranted. Considering the simulation period of the models, we used the reference data over CONUS for the three time periods, 1895-1939 (Period 1), 1940-1979 (Period 2), and 1980-2005 (Period 3).

b. Methodology

Spatially-averaged regional means were calculated for the NCA regions over CONUS (Fig. 1a). To ensure that each NCA subregion is of equal size and comprised of the same number of grid points among member models, all data is first interpolated to a 2° latitude/longitude grid mesh. Performing the data interpolation before spatial averaging does not affect the results of the current study. The Great Plains region used in the NCA3 (Melillo et al., 2014) was divided into a northern and southern component, following expectations this will be formalized for the next NCA report, resulting in a total of seven regions (Table 2) as presented in Janssen et al. (2014): Northwest, Southwest, Northern Great Plains, Southern Great Plains, Midwest, Northeast, and Southeast.

Trends in temperature and precipitation are computed for summer (June through August; JJA) and winter (December through February; DJF) applying the method of least squares to regionally averaged temperature and precipitation data for the two seasons in the reference and the CMIP5 models. The linear trends are computed for the reference, the multi-model ensemble (MME), and ensemble members for three time blocks (1895-1939, 1940-1979, 1980-2005). The choice of the time blocks is based on the observed warming and cooling trends, and closely mimics those in Kunkel et. al (2006). These time periods were chosen as a result of observed increasing temperature trends during 1895-1939 and 1980-2005 in contrast to a decreasing temperature trend during 1940-1979 (Figure 1b). Four metrics are used to evaluate model performance: a trend bias of the MME, trend biases of ensemble members, percent of ensemble members producing trends of the same sign (+/-) as observations, and the percent of ensemble members with biases that are small in comparison to the standard errors of the observed and simulated trends. The standard errors in the observed and simulated trends are calculated based on the equation in Hogg and Tanis (2009). (The detailed methodology and equations can be found in Appendix A).

MME trends may be influenced more strongly by some models that contribute more ensemble members than the others. For instance, there are models that provide as little as one simulation or as many as 25 different ensemble members. In the case of a model contributing a large number of ensemble members, this particular model will bear greater weight to the overall regional mean because each simulation is weighted equally when calculating MME (Appendix A, Equation 4). Considering the unequal number of simulations by the models, the standard error in MME's trend was estimated by randomly selecting N (Table 1) ensemble members trends with replacement (i.e. bootstrapping) and computing the mean of that selection. This sampling was repeated 1000 times to obtain standard deviation across the 1000 random ensemble trends for use as the standard error of the MME. The bias of the MME was then compared to the reference trend and the standard error (as calculated with the equation from Hogg and Tanis [2009]).

3. RESULTS

a. Near-surface air temperature

i. Evaluation of the multi-model ensemble trend

Figure 1b shows the MME and individual ensemble member temperature timeseries over the entire evaluation period along with the nClimGrid timeseries over the CONUS for JJA and DJF. The MME qualitiatively resembles the general observed trend with some exceptions, such as the late part of the record for DJF. The model spread is considerable, however, as represented by the gray lines in both figure panels. Temperature trends for the Reference (i.e. nClimGrid) and MME, and the trend biases of MME, for the three time periods during JJA are shown in Figure 2. The trend biases are marked with an asterisk if they are large relative to the standard errors of trends in nClimGrid and MME (discussed further below). During Period 1 (1895-1939), warming occurred in all seven regions over the CONUS with the largest positive trend in the northern Great Plains region (Figure 2a). The MME also has an average warming trend (~0.085 K/decade) over all of CONUS (Figure 2b), but with a lesser magnitude as indicated by the negative bias values in all but the Southeast (Figure 2c). Also, in the northern Great Plains, the difference between the MME and reference trends is larger than standard errors of the trends and is indicated with an asterisk (with the significance information based on results in Figure 3). During Period 2 (1940-1979), reference data shows a warming trend over the western half of the CONUS, whereas the eastern half shows cooling (Figure 2d). In particular, the Southeast region displays the strongest cooling trend (-0.24 K/decade). In MME, the cooling trend is weaker than the observed one in all four eastern half regions (~-0.03 K/decade). The negative (positive) biases in the western (eastern) half of CONUS (Figure 2f) indicate that MME underestimates both warming and cooling trends compared to nClimGrid. The difference between the observed and MME trends is largest in the Southeast region (0.22 K/decade). Kunkel et al. (2006) report the underestimated cooling trends in the Southeast region by GCMs and suggest that the large trend biases of models result from the North Atlantic sea surface temperature. Period 3 (1980-2005) is characterized by an overall warming trend except for the Southeast region where a cooling trend is observed (Figure 2g). The warming is strongest in the Northwest region (0.50 K/decade) with relatively weak warming trends in the three central regions. However, MME (Figure 2h) exhibits a strong warming trend of 0.4 K/decade averaged over CONUS, leading to a considerable positive trend bias in the Southeast (0.40 K/decade), and an overestimation of the warming trend over the remaining regions while still leaving an underestimation of the warming trend in the Northwest (Figure 2i). In the Southeast region, the trend bias is larger than standard errors thus indicated with an asterisk.

Figure 3 shows the temperature trends in nClimGrid (blue square) and MME (red circle), each plotted with its standard error as whiskers (i.e. trends ± standard errors). There is overlap between the red and blue whiskers in all regions, including at the CONUS scale, except for the northern Great Plains region during Period 1 (Figure 3a) and the Southeast region during Period 3 (Figure 3c). The non-overlapping whiskers indicate that the MME trends may not have much fidelity in representing the trends in these regions/periods with larger trend biases than the standard errors of trends in nClimGrid and MME.

Figure 4 and 5 illustrate the trends from nClimGrid and MME for winter (DJF). During Period 1, a warming trend dominated the central and eastern regions (Figure 4a). However, the

Northwest experienced cooling while there was no apparent trend in the Southwest. MME displays a weak warming trend (~0.08 K/decade) over all of CONUS (Figure 4b). In the Southeast region, the difference between the MME and reference trends is larger than standard errors of the trends (Figure 4c). During Period 2, reference data shows a cooling trend over all of CONUS except for the Northwest region that shows a warming trend (Figure 4d). In particular, the Midwest region displays a strong cooling trend. MME displays a cooling trend in the western half of CONUS while the eastern half shows a warming trend (Figure 4e). However, the northern Great Plains region did not exhibit a trend. In the Midwest region, the difference between the MME and reference trends is larger than standard errors of the trends (Figure 4f). The latest period, Period 3, is characterized by warming trends over all of CONUS (Figure 4g). MME exhibits a similar warming trend over all of CONUS (Figure 4h) but underestimates the magnitude as indicated by the bias map displaying negative values (Figure 4i). Figure 5 shows trends and standard errors from nClimGrid and MME for DJF. The trend differences between nClimGrid and MME are smaller than their standard errors except for the Southeast regions during Period 1 (Figure 5a) and the Midwest region from Period 2 (Figure 5b).

Overall, the MME reproduces the reference trends within their standard errors for both seasons over all regions except the four cases: Northern Great Plains in JJA during Period 1 (1895-1939), the Southeast in JJA during Period 3 (1980-2005), the Southeast in DJF during Period 1 (1895-1939), and the Midwest in DJF during Period 2 (1940-1979). The greater model departures from the trends in nClimGrid suggest the applied external forcing and/or the modeled processes in those regions might lack the same fidelity as for the other regions that tend to be represented better.

ii. Evaluation of the trends from ensemble members

In an effort to understand the performance of ensemble members that comprise MME, the trend distribution from 206 ensemble members for JJA are plotted as a box and whiskers with the lower and higher ends of the whiskers representing the 5% and 95% of the *N* ensemble member trend values (Figure 3). The lower and upper boundaries of the box represent the lower and upper quartile with the middle line indicating the median of the *N* trends.

For the JJA analysis (Figure 3), because MME's trends are almost the same as the median of all the ensemble member trend and lengths of red and black whiskers are similar to each other, many of the same findings and discussion of the results from Section 3.i apply here. Figure 3a shows that both the reference trend and at least 75% of the ensemble member trends show a warming over all of CONUS during Period 1. During Period 2 (Figure 3b), the boxplot shows that only about half of the ensemble members show a warming trend in the three western regions for Period 2 and a cooling trend for the other regions. This indicates that roughly half of the ensemble members reproduce the same warming or cooling trend as the reference for this time period. For Period 3 (Figure 3c), the boxplot of the ensemble members exhibit warming over all of CONUS, although in some cases with considerably greater magnitude, and thus are consistent with the reference in regards to the sign of the warming. However, in the Southeast, at least 95% of the ensemble members indicates a warming rather than the cooling trend found in the reference (i.e. the bottom of the black whisker has almost the same value as the top of the blue whisker).

The percentage of ensemble members whose mean temperature trends are not significantly different from the reference at the 90% confidence level are tabulated in Table 3 for JJA. For this table, a larger value indicates greater fidelity of the ensemble members to represent the reference trend and vice versa for a smaller value. A large confidence level (in this case 90%) indicates that the hypothesis (Appendix A, Equation 7), which states the reference trend equals the ensemble member trend, can be accepted as true. For example, for JJA during Period 1 (1895-1939), there are 184 ensemble members whose trend was compared to the reference trend. In the Southeast region, 11% of the simulation trends are not significantly different from the observed trend at the 90% confidence level. For the same region, during Period 2 (1940-1979), 8% of the ensemble members have temperature trends that are not significantly different from the observed trend while only 4% during Period 3 (1980-2005) are not significantly different at the 90% confidence level. This shows that for the Southeast region, the fidelity of the ensemble members to represent the observed temperature trends diminishes with each successive time block. The region and time block best represented by the ensemble members is the Northwest region during Period 3, where 30% of the simulated trends are not significantly different from the observed trend at the 90% confidence level. At the CONUS analysis level, 1% of the simulation trends during Period 1 (1895-1939) are not significantly different from the observed trend at the 90% confidence level while Period 2 (1940-1979) and Period 3 (1980-2005) produced higher percentages of 16% and 17.5% respectively. This shows that at the CONUS level, the fidelity of the ensemble members to represent the observed temperature trends increases with each successive time block.

For further analysis of the ensemble members, Figure 6(a-c) takes into account the percent of the GCMs that can correctly reproduce the same sign (+/-) of warming (cooling) as the temperature trends in nClimGrid. During Period 1 (Figure 6a), a large number of the ensemble members (76-83%) are capable of reproducing the same sign trend as nClimGrid. During Period 2 (Figure 6b), roughly half (39-59%) of the ensemble members reproduce the same sign trend, whereas an overwhelming number of ensemble members (91-97%) do so during Period 3 (Figure 6c) except for the Southeast region where only 3.9% of ensemble members produce the same sign trend. Figure 6(d-f) is used to demonstrate the percent of the ensemble members whose trend biases are small relative to standard errors of nClimGrid and simulated trends (i.e. their trend values are comparable within the range of the standard errors of each). During Period 1 (Figure 6d), about half (39-59%) of the ensemble members reproduce temperature trends similar to the observed trend except for the northern Great Plains (7.1%). During Period 2 (Figure 6e), only about a quarter (24-34%) of the ensemble members have smaller biases than standard errors. In the Southwest region, only 10% of the simulated trends show reasonable agreement with the observed trend. During Period 3 (Figure 6f), roughly half of the ensemble members over the Western regions along with the Northeast reproduce the reference trends (46-54%), whereas ensemble members perform poorly (10-22%) in the rest of the regions.

For the corresponding analysis of DJF during Period 1 (i.e. Table 4, Figure 5 and Figure 7), the reference mean trends in Figure 5 shows a period of warming in all regions except the Northwest whereas at least 75% of the ensemble member trends show a warming period over all of CONUS. Conversely, from 1940-1979, the reference data indicates a cooling trend in all regions except the Northwest. Only about half of the ensemble members show a cooling trend in all regions. The reference data between 1980-2005 show a warming trend for all regions and at least 90% of the ensemble members depict this warming trend.

The percentage of ensemble members whose mean temperature trends are not significantly different from the reference at the 90% confidence level are tabulated in Table 4 (see discussion of Table 3 above). As with Table 3, a large confidence level (in this case 90%) indicates that the hypothesis (Appendix A, Equation 7), which states the reference trend equals the ensemble member trend, can be accepted as true. The overall fidelity of the ensemble members to represent the observed temperature trend increases with each successive time block in all regions except the northern Great Plains and the Northeast where the observed temperature trend diminishes with each successive time block. Similar to the JJA analysis over CONUS, the DJF percentages of simulation trends with mean temperature trends that are not significantly different from the observed trend increased with each successive time block with 7%, 17%, and 26% of the simulations for Period 1 (1895-1939), Period 2 (1940-1979), and Period 3 (1980-2005) at the 90% confidence level.

Figure 7(a-c) illustrates the percent of the ensemble members that reproduce the same sign (+/-) as the reference temperature trend during winter. During 1895-1939 (Figure 7a), all regions except for the Northwest showed that more than half of the ensemble members (59-68%) can reproduce the same sign trend as the reference. From 1940-1979 (Figure 7b), roughly half (42-60%) of the ensemble members reproduce the same sign trend while a majority of ensemble members (72-81%) do so in 1980-2005 (Figure 7c). Figure 7(d-f) also displays the percent of the ensemble members that reproduce the values of temperature trend with overlapping standard error bars (with the overlap indicating that the difference between the reference and ensemble member trends are not statistically significant). During 1895-1939 (Figure 7d), a majority of the ensemble members (78-98%) in the western regions and a minority (7.1%) in the Southeast reproduce the reference trend with overlapping error bars. From 1940-1979 (Figure 7e), the ensemble members perform poorly (12-23%) over the Midwest and Southeastern region while performing better (>40%) in the other areas. However, between 1980-2005 (Figure 7f), the ensemble members show an outstanding reproduction of the reference trend (74%) over the

Southwest region while only about half of the ensemble members are able to reproduce the reference trend in the other regions.

b. Precipitation

The same regional trend analyses were performed using the same time periods and metrics for precipitation as described in Section 3.a for temperature. Graphs and maps analogous to Figures 3 and 6 were constructed for precipitation from this analysis for both JJA and DJF, and are included in the Supplementary Material for completeness. However, no statistically significant seasonal trends were found both in the reference and the MME for any region or time period for either season due to a large variability between the models. Despite this, a zoomed in version of Supplementary Figures 1 and 2 show that although the MME trends overlap completely with the reference trends, the large model variability reproduces the reference trends with no skill. However, while the results presented here are seasonal monthly mean values, it should be noted that other studies have found precipitation trends when examining extreme precipitation events in the U.S. (Janssen et al., 2016) where it was observed that the models overestimate the number of extreme events in the spring while underestimating in the summer. In examining the trend of extreme precipitation events, Karl et al., (1996) and Kunkel et al. (2003), amongst many others, report that frequencies of these events were high in the early twentieth century, followed by a period of low frequency in the 20-30s with a gradual increase in the extreme events thereafter.

4. SUMMARIZING RESULTS

As a means of compactly summarizing aspects of the overall performance of the model suite for temperature, Figures 8a and 9a visualize simultaneously the bias in the MME temperature trends (Figure 3) and the percent of ensemble members reproducing the same sign as the nClimGrid trend (Figure 6a-c) for JJA (Figure 8a) and DJF (Figure 9a). Here, the different symbols, square (Period 1: 1895-1939), triangle (Period 2: 1940-1979), and circle (Period 3: 1980-2005) indicate the three time periods, with the colors representing the different regions where the color code is based on Figure 1a. In these figures, ideal model fidelity is exhibited by symbols that fall in the area close to the center horizontally and to the top vertically. The small cluster of squares around a bias of -0.1 (K/decade) and at 80% indicates that the model performance in the time block 1895-1939 (square symbols) is relatively good, for the bias is small and most of simulated trends have the same sign as the observational reference (i.e. nClimGrid). The values for Period 2 (triangles) are clustered between 40-60% with a wide spread in the bias between -0.15 and 0.22 (K/decade). Most of the values for Period 3 (circles) are in the ~90% range. Overall, the ensemble members qualitatively reproduce the observed trends. The exception is the Southeast region (gray circle) that shows a large bias (0.38 K/decade) and only 4% of the ensemble members reproduce the observed cooling trend. The Northern Great Plains region in Period 1 (green square) has a bias of -0.36 (K/decade) and 80% agreement, indicating that MME largely underestimates the trend, whereas 80% of ensemble members are capable of producing the same sign as the reference trend.

Another aspect of the performance of the CMIP5 model suite for these regional temperature trends can be summarized compactly by replacing the y-axis in Figures 8a and 9a with the percent of ensemble members whose trend and standard error overlap with the observed reference trend and its standard error (Figures 8b and 9b). Here, the green square, the Northern

Great Plains region from 1895-1939, has the same bias of -0.36 (K/decade) however only 7% of the ensemble members showed a trend that fell within the reference trend and standard error (Figure 8b).

Overall, for both Figure 8a and Figure 8b, the Northern Great Plains during Period 1, and the Northern and Southern Great Plains, Midwest, and the Southeast from Period 3 are outliers with relatively large warming and cooling biases, greater than 0.3 K/decade. However, the other ensembles during Period 1 generally replicate the sign (+/-) of the reference trend well, around 80%, and exhibit a 40-60% replication of the reference trend when the trend value is compared within standard error.

For DJF, Figure 9a shows that the values for Period 1 (squares) are largely scattered horizontally across the plot, exhibiting a large range in MME biases, between about -0.51 to 0.16 (K/decade), and with only 37-68% of the ensemble members able to reproduce the sign of the observed trend. The values for Period 2 (triangles) are well clustered between 40-50% but also exhibit a wide spread in the bias between -0.10 and 0.47 (K/decade). The values for Period 3 (circles) all exhibit agreements of 70-80% indicating that the sign reproduction of the ensemble members with the reference trend is relatively high compared to the two earlier periods. For this period, the MME biases still exhibits a large range of -0.37 to -0.02 (K/decade), and notably all underestimating the temperature trends.

In Figure 9b, the bias values for all periods each range over about 0.5 (K/decade), although Period 1 (squares) and Period 3 (circles) tend to be biased negative, while that for Period 2 (triangles) tends to be biased positive. However, in the case of the agreements exhibited with the observed temperature trends, Period 1 and 2 exhibit a wide variation of values (nearly over the whole range), while Period 3 exhibits a more consistent level of agreement across members

(~40-70%). Comparing DJF values (Figure 9b) to those for JJA (Figure 8b), the results suggest that MME exhibits greater biases during DJF (wide spread of the data along the bias axis) but ensemble members can better capture the reference trend within the standard error (i.e. there are more points in the upper half of the plot).

Overall, for both Figure 9a and Figure 9b, the Midwest from 1940-1979 (yellow triangle) represents a large outlier, with a large bias of 0.47 K/decade and only 12% of the ensemble members reproducing a trend that is within the standard error of the reference trend. Contrastingly, the Southwest region in 1895-1939 (red square) shows the best individual model performance (compared to the other periods and regions) with a small bias (0.05 K/decade) with 98% of the models producing a trend that is within the standard error of the reference trend.

5. CONCLUSIONS

The ability of CMIP5 GCMs to reproduce near-surface air temperature and precipitation trends over CONUS is quantified and characterized in this study. The novelty of this study, in contrast to similar past studies, is the utilization of nClimGrid, a new observational reference dataset that exhibits a number of improvements over other similar datasets in ways that are pertinent to this analysis (see Section 2a). In addition, this evaluation includes all the available historical simulations and thus has an element of comprehensiveness compared to past studies and the model contributions to CMIP5. The analysis framework involves comparing simulated trends, both as an MME and all 206 individual members, to the reference trends considering the values averaged over the 7 NCA regions and for three time periods (1895-1939, 1940-1979, 1980-2005). Most of the study's results are summarized compactly in Section 4 and in Figures 8 and 9. A few of the key results are highlighted below, with remarks and interpretations provided where possible.

For summer (JJA), MME mean trends in near-surface air temperature exhibit some degree of bias in most regions in all time periods relative to the observations (Period 1: 1895-1939, Period 2: 1940-1979, Period 3: 1980-2005). However, only the Northern Great Plains during Period 1 and Southeast during Period 3 (Figures 2, 3, and 8) exhibit significant biases. At the CONUS scale, the trend in nClimGrid is about 0.15 K/dec higher than MME for the earliest period while the two have roughly comparable trends for 1940-1979. The nClimGrid trend is about 0.2 K/dec lower than MME for 1980-2015. Differences are slightly larger for winter; in particular, the trend in nClimGrid is about 0.2 K/dec higher than MME for 1895-1939 and 1980-2015, and about 0.2 K/dec lower for 1940-1979. While a full explanation of the causes of such difference in the Southeast region is beyond the scope of this study, this is considered to be due to the fact that forcings may not be entirely accurate and/or the internal variability that contributes to the "warming hole" in this region is much larger than the simulations exhibit (Kunkel et al. 2006). The inability of the MME to capture the "warming hole" is corroborated with several previous studies (Kunkel et al., (2006), Meehl et al., (2012), Kumar et al., (2013)) and brings to attention the need for inquiry by the scientific and model development communities. Similarly, for the winter period, the MME shows varying degrees of bias, however only the Southeast during Period 1 and the Midwest during Period 2 (Figures 4, 5, and 9) have significant biases.

Considering the performance of the collective of ensemble members, for JJA, no region for any time block had more than about half of the ensemble members that are able to capture the value of the reference trend within the overlapping standard error (Figure 6d-f). For example, during Period 3, roughly half of the ensemble members over the western and eastern regions captured the observed trend whereas less than 22% of the ensemble members are capable of doing so over the central regions. For winter, the performance of the collective of ensemble members produced different results. For example, during Period 1, more than half of the ensemble members reproduced the observed trend (defined here as within the standard error of the reference) over the western regions, and as high as 98% of the ensemble members did so for the Southwest region (Figure 9b).

The results highlighted above, and outlined in more detail in Sections 3 and 4, show that the CMIP5 MME (Figures 3 and 5) can reproduce historical surface air temperature trends for both summer and winter seasons across all periods over the Northwest, Southwest, Southern Great Plains, and Northeast regions but not the Northern Great Plans, Midwest and Southeast. When considering the 206 individual CMIP5 model ensemble member simulations of the historical surface air temperature, at least 76% of them correctly reproduced a positive (warming) trend during Period 1 and Period 3 for both JJA and DJF over all regions except the Northwest and Southeast regions. In contrast, the fidelity of the model ensemble member simulations of the historical surface air temperature trends is not skillful for Period 2. It should also be noted that experimentation with a shift of plus or minus 5 years in the three time periods did not change the overall qualitative results. Finally, precipitation trends were not found to be skillfully replicated over either season, during any period, and over any region because the CMIP5 GCMs exhibit large variability and the reference trends for the mean monthly values examined here are relatively small or near zero.

The results of this study point to the specific regions over CONUS that warrant further investigation on the fidelity of the historical forcing data utilized for these simulations and/or on the fidelity of the model in representing the processes key to determining near surface temperature and precipitation. Focus on improving the performance of the simulations and realism of the forcing data such that model bias is reduced over these regions would be a productive investment in reducing uncertainty in future model projections.

Acknowledgements

We would like to acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Appendix B of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

The primary author would also like to acknowledge California State University, Los Angeles NASA DIRECT-STEM program and director, Dr. Hengchun Ye for funding and support.

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Support for this project was provided by NASA National Climate Assessment 11-NCA 11-0028. Kenneth Kunkel was supported by NOAA through the Cooperative Institute for Climate and Satellites — North Carolina under Cooperative Agreement NA14NES432003.

6. References

Baek, H.-J., Lee, J., Lee, H.-S., Hyun, Y.-K., Cho, C., Kwon, W.-T., … Byun, Y.-H. (2013). Climate change in the 21st century simulated by HadGEM2-AO under representative concentration pathways. *Asia-Pacific Journal of Atmospheric Sciences*, *49*(5), 603–618.

https://doi.org/10.1007/s13143-013-0053-7

Barnett, T. P., Pierce, D. W., Hidalgo, H. G., Bonfils, C., Santer, B. D., Das, T., … Dettinger, M. D. (2008). Human-Induced Changes in the Hydrology of the Western United States. *Science*, *319*(5866), 1080 LP-1083. Retrieved from http://science.sciencemag.org/content/319/5866/1080.abstract

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevåg, A., Seland, Ø., … Kristjánsson, J. E. (2013). The Norwegian Earth System Model, NorESM1-M – Part 1: Description and basic evaluation of the physical climate. *Geoscientific Model Development*, *6*(3), 687–720. https://doi.org/10.5194/gmd-6-687-2013

Bukovsky, M. S. (2012). Temperature trends in the NARCCAP regional climate models. *Journal of Climate*, *25*(11), 3985–3991. https://doi.org/10.1175/JCLI-D-11-00588.1

Cayan, D. R., Tyree, M., Kunkel, K. E., Castro, C., Gershunov, A., Barsugli, J., … Barlow, M. (2013). Future Climate: Projected Average. In G. Garfin, A. Jardine, R. Merideth, M. Black, & S. LeRoy (Eds.), *Assessment of Climate Change in the Southwest United States: A Report Prepared for the National Climate Assessment* (pp. 101–125). Washington, DC: Island Press/Center for Resource Economics. https://doi.org/10.5822/978-1-61091-484-0_6

Chylek, P., Li, J., Dubey, M. K., Wang, M., & Lesins, G. (2011). Observed and model simulated 20th century Arctic temperature variability: Canadian Earth System Model CanESM2. *Atmospheric Chemistry and Physics Discussions*, *11*(8), 22893–22907. https://doi.org/10.5194/acpd-11-22893-2011

Collier, M., & Uhe, P. (2012). *CMIP5 datasets from the ACCESS1.0 and ACCESS1.3 coupled climate models*. *CAWCR Technical Report No. 059*. https://doi.org/ISBN: 20978-1-922173-29-4

Collins, W. D., Rasch, P. J., Boville, B. A., Hack, J. J., Williamson, D. L., Kiehl, J. T., … Dai, Y. (2004). Description of the NCAR Community Atmosphere Model (CAM 3.0). *Ncar/Tn-464+Str*, (June), 214.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., … Woodward, S. (2011). Development and evaluation of an Earth-System model – HadGEM2. *Geoscientific Model Development*, *4*(4), 1051–1075. https://doi.org/10.5194/gmd-4-1051-2011

Crowley, T. J. (2000). Causes of Climate Change Over the Past 1000 Years. *Science*, *289*(5477), 270 LP-277. Retrieved from http://science.sciencemag.org/content/289/5477/270.abstract

Delworth, T. L., Broccoli, A. J., Rosati, A., Stouffer, R. J., Balaji, V., Beesley, J. A., … Zhang, R. (2006). GFDL's CM2 Global Coupled Climate Models. Part I: Formulation and Simulation Characteristics. *Journal of Climate*, *19*(5), 643–674. https://doi.org/10.1175/JCLI3629.1

Donner, L. J., Wyman, B. L., Hemler, R. S., Horowitz, L. W., Ming, Y., Zhao, M., … Zeng, F. (2011). The Dynamical Core, Physical Parameterizations, and Basic Simulation Characteristics of the Atmospheric Component AM3 of the GFDL Global Coupled Model CM3. *Journal of Climate*, *24*(13), 3484–3519. https://doi.org/10.1175/2011JCLI3955.1

Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., … Vuichard, N. (2013). Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Climate Dynamics*, *40*(9), 2123–2165. https://doi.org/10.1007/s00382-012-1636-1

Dunne, J. P., John, J. G., Shevliakova, E., Stouffer, R. J., Krasting, J. P., Malyshev, S. L., … Zadeh, N. (2013). GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models.

Part II: Carbon System Formulation and Baseline Simulation Characteristics. *Journal of Climate*, *26*(7), 2247–2267. https://doi.org/10.1175/JCLI-D-12-00150.1

Easterling, D.R., K.E. Kunkel, J.R. Arnold, T. Knutson, A.N. LeGrande, L.R. Leung, R.S. Vose, D.E. Waliser, and M.F. Wehner, 2017: Precipitation change in the United States. In *Climate Science Special Report: Fourth National Climate Assessment, Volume I*. D.J. Wuebbles, D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock, Eds. U.S. Global Change Research Program, pp. 207-230, doi:10.7930/J0H993CC

Gordon, H., Farrell, S. O., Collier, M., Dix, M., Rotstayn, L., Kowalczyk, E., … Watterson, I. (2010). *The CSIRO Mk3 . 5 Climate Model*.

Hazeleger, W., Wang, X., Severijns, C., Ştefănescu, S., Bintanja, R., Sterl, A., … Van den Hurk, B. (2012). EC-Earth V2. 2: description and validation of a new seamless earth system prediction model. *Climate Dynamics*, *39*(11), 2611–2629.

Hogg, Robert V, and Elliot A. Tanis. (2009) *Probability and Statistical Inference*. Pearson Educational International.

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern- mental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Janssen, E., Sriver, R. L., Wuebbles, D. J., & Kunkel, K. E. (2016). Seasonal and regional variations in extreme precipitation event frequency using CMIP5. *Geophysical Research Letters*, *43*(10), 5385–5393. https://doi.org/10.1002/2016GL069151

Janssen, E., Wuebbles, D., & Kunkel, K. (2014). Observational and Model based Trends and Projections of Extreme Precipitation over the Contiguous United States. *Earth's Future*, 1–

15. https://doi.org/10.1002/2013EF000185.Received

Ji, D., Wang, L., Feng, J., Wu, Q., Cheng, H., Zhang, Q., … Zhou, M. (2014). Description and basic evaluation of Beijing Normal University Earth System Model (BNU-ESM) version 1. *Geoscientific Model Development*, *7*(5), 2039–2064. https://doi.org/10.5194/gmd-7-2039-2014

Karl, T. R., Knight, R. W., Easterling, D. R., & Quayle, R. G. (1996). Indices of Climate Change for the United States. *Bulletin of the American Meteorological Society*, *77*(2), 279–292. https://doi.org/10.1175/1520-0477(1996)077<0279:IOCCFT>2.0.CO;2

Knutson, T. R., Delworth, T. L., Dixon, K. W., Held, I. M., Lu, J., Ramaswamy, V., … Stouffer, R. J. (2006). Assessment of Twentieth-Century Regional Surface Temperature Trends Using the GFDL CM2 Coupled Models. *Journal of Climate*, *19*(9), 1624–1651. https://doi.org/10.1175/JCLI3709.1

Knutson, T. R., Zeng, F., & Wittenberg, A. T. (2013). Multimodel Assessment of Regional Surface Temperature Trends: CMIP3 and CMIP5 Twentieth-Century Simulations. *Journal of Climate*, *26*(22), 8709–8743. https://doi.org/10.1175/JCLI-D-12-00567.1

Kumar, S., III, J. K., Dirmeyer, P. A., Pan, Z., & Adams, J. (2013). Multidecadal Climate Variability and the “Warming Hole” in North America: Results from CMIP5 Twentieth- and Twenty-First-Century Climate Simulations. *Journal of Climate*, *26*(11), 3511–3527. https://doi.org/10.1175/JCLI-D-12-00535.1

Kumar, S., Merwade, V., Kinter, J. L., & Niyogi, D. (2013). Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 twentieth-century climate simulations. *Journal of Climate*, *26*(12), 4168–4185. https://doi.org/10.1175/JCLI-D-12-00259.1

Kunkel, K. ., Stevens, L. E., Stevens, S. E., Sun, L., Janssen, E., Wuebbles, D., & Dobson, J. G. (2013). Regional Climate Trends and Scenarios for the U . S . National Climate Assessment Part 9 . Climate of the Contiguous United States, (January), 77. Retrieved from http://www.nesdis.noaa.gov/technical_reports/149_Climate_Scenarios.html

Kunkel, K. E., Easterling, D. R., Redmond, K., & Hubbard, K. (2003). Temporal variations of extreme precipitation events in the United States: 1895–2000. *Geophysical Research Letters*, *30*(17), n/a-n/a. https://doi.org/10.1029/2003GL018052

Kunkel, K. E., Liang, X.-Z., Zhu, J., & Lin, Y. (2006). Can CGCMs simulate the twentieth-century “warming hole” in the central United States? *Journal of Climate*, *19*(17), 4137–4153. https://doi.org/10.1175/JCLI3848.1

Kunkel, K. E., & Liang, X. Z. (2005). GCM simulations of the climate in the central United States. *Journal of Climate*, *18*(7), 1016–1031. https://doi.org/10.1175/JCLI-3309.1

Kunkel, K. E., Vose, R. S., Stevens, L. E., & Knight, R. W. (2015). Is the monthly temperature climate of the United States becoming more extreme? *Geophysical Research Letters*, *42*(2), 629–636. https://doi.org/10.1002/2014GL062035

Li, L., Lin, P., Yu, Y., Wang, B., Zhou, T., Liu, L., … Qiao, F. (2013). The flexible global ocean-atmosphere-land system model, Grid-point Version 2: FGOALS-g2. *Advances in Atmospheric Sciences*, *30*(3), 543–560. https://doi.org/10.1007/s00376-012-2140-6

Marsh, D. R., Mills, M. J., Kinnison, D. E., Lamarque, J.-F., Calvo, N., & Polvani, L. M. (2013). Climate Change from 1850 to 2005 Simulated in CESM1(WACCM). *Journal of Climate*, *26*(19), 7372–7391. https://doi.org/10.1175/JCLI-D-12-00558.1

Meehl, G. A., Arblaster, J. M., & Branstator, G. (2012). Mechanisms Contributing to the Warming Hole and the Consequent U.S. East–West Differential of Heat Extremes. *Journal* *of Climate*, *25*(18), 6394–6408. https://doi.org/10.1175/JCLI-D-11-00655.1

Meehl, G. A., & Tebaldi, C. (2004). More Intense, More Frequent, and Longer Lasting Heat Waves in the 21st Century. *Science*, *305*(5686), 994 LP-997. Retrieved from http://science.sciencemag.org/content/305/5686/994.abstract

Melillo, J. M., Richmond, T. C., Yohe, G. W., & US National Climate Assessment. (2014). *Climate change impacts in the United States: the third national climate assessment. US Global change research program* (Vol. 841). https://doi.org/10.7930/j0z31WJ2

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, *29*(7), 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

Pope, V. D., Gallani, M. L., Rowntree, P. R., & Stratton, R. A. (2000). The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Climate Dynamics*, *16*(2), 123–146. https://doi.org/10.1007/s003820050009

Qiao, F., Song, Z., Bao, Y., Song, Y., Shu, Q., Huang, C., & Zhao, W. (2013). Development and evaluation of an Earth System Model with surface gravity waves. *Journal of Geophysical Research: Oceans*, *118*(9), 4514–4524.

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., … Zhang, J. (2014). Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, *6*(1), 141–184. https://doi.org/10.1002/2013MS000265

Scoccimarro, E., Gualdi, S., Bellucci, A., Sanna, A., Fogli, P. G., Manzini, E., … Navarra, A. (2011). Effects of Tropical Cyclones on Ocean Heat Transport in a High-Resolution Coupled General Circulation Model. *Journal of Climate*, *24*(16), 4368–4384.

https://doi.org/10.1175/2011JCLI4104.1

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., … Roeckner, E. (2013). Atmospheric component of the MPI-M Earth System Model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. https://doi.org/10.1002/jame.20015

T. SAKAMOTO, T., KOMURO, Y., NISHIMURA, T., ISHII, M., TATEBE, H., SHIOGAMA, H., … KIMOTO, M. (2012). MIROC4h-A New High-Resolution Atmosphere-Ocean Coupled General Circulation Model. *Journal of the Meteorological Society of Japan. Ser. II*, *90*(3), 325–359. https://doi.org/10.2151/jmsj.2012-301

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Voldoire, A., Sanchez-Gomez, E., y Mélia, D., Decharme, B., Cassou, C., Sénési, S., … Chauvin, F. (2013). The CNRM-CM5.1 global climate model: description and basic evaluation. *Climate Dynamics*, *40*(9), 2091–2121. https://doi.org/10.1007/s00382-011-1259-y

Volodin, E. M., Dianskii, N. A., & Gusev, A. V. (2010). Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations. *Izvestiya, Atmospheric and Oceanic Physics*, *46*(4), 414–431. https://doi.org/10.1134/S000143381004002X

Vose, R. S., Applequist, S., Squires, M., Durre, I., Menne, C. J., Williams, C. N., … Arndt, D. (2014). Improved historical temperature and precipitation time series for U.S. climate divisions. *Journal of Applied Meteorology and Climatology*, *53*(5), 1232–1251. https://doi.org/10.1175/JAMC-D-13-0248.1

Watanabe, M., Suzuki, T., O'ishi, R., Komuro, Y., Watanabe, S., Emori, S., … Kimoto, M. (2010). Improved Climate Simulation by MIROC5: Mean States, Variability, and Climate Sensitivity. *Journal of Climate*, *23*(23), 6312–6335. https://doi.org/10.1175/2010JCLI3679.1

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., … Kawamiya, M. (2011). MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development*, *4*(4), 845–872. https://doi.org/10.5194/gmd-4-845-2011

Wu, T., Song, L., Li, W., Wang, Z., Zhang, H., Xin, X., … Zhou, M. (2014). An overview of BCC climate system model development and application for climate change studies. *Journal of Meteorological Research*, *28*(1), 34–56. https://doi.org/10.1007/s13351-014-3041-7

Wuebbles, D. J., Kunkel, K., Wehner, M., & Zobel, Z. (2014). Severe Weather in United States Under a Changing Climate. *Eos, Transactions American Geophysical Union*, *95*(18), 149–150. https://doi.org/10.1002/2014EO180001

Wuebbles, D. J., D. W. Fahey, K. A. Hibbard, D. J. Dokken, B. C. Stewart, and T. K. Maycock (eds.). US Global Change Research Program, Washington, DC, USA, 470 pp, doi: 10.7930/J0J964J6.

Yukimoto, S. (2011). *Meteorological research institute earth system model version 1 (MRI-ESM1): model description*.

YUKIMOTO, S., ADACHI, Y., HOSAKA, M., SAKAMI, T., YOSHIMURA, H., HIRABARA, M., … KITOH, A. (2012). A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3 -Model Description and Basic Performance. *Journal of the*

Meteorological Society of Japan. Ser. II, *90A*, 23–64. https://doi.org/10.2151/jmsj.2012-A02

Appendix A

Methodology Equations

For each NCA region, the seasonal mean time series from the reference data is represented as:

$$x(t), (t = 1, 2, \ldots, m) \qquad (1)$$

where

t: the year from the starting year in each time block

m: number of years in the time block

The regionally-averaged time series from ensemble member *i* is defined as:

$$y_i(t), (t = 1, 2, \ldots, m) \qquad (2)$$

In addition, the regionally-averaged time series from the MME is represented as:

$$Y(t), (t = 1, 2, \ldots, m) \qquad (3)$$

where, the ensemble average of *N* simulations is calculated using the following equation:

$$Y(t) = \frac{1}{N}\sum_{k=1}^{N} y_i(t) \qquad (4)$$

See Table 2 for the number of simulations (*N*) used to calculate *Y(t)* in each time block.

The seasonal mean trend for the reference data, α_{ref} [K year^{-1}], is defined as the least square fit for a linear regression model:

$$x = \alpha_{ref} \times t + \beta_{ref} \qquad (5)$$

The linear trend, α_{yi} [K year^{-1}], for ensemble members, $y_i(t)$, and the ensemble linear trend, α_Y, for MME, $Y(t)$, is calculated in the same manner for three time blocks (1895-1939, 1940-1979, 1980-2005). The choice of the time blocks is based on the observed warming and cooling trends, and closely mimics those in Kunkel et. al (2006).

The performance metric of the simulated trends in each region are:

a) trend bias of the MME, $\alpha_Y - \alpha_{ref}$,
b) trend biases of ensemble members, $\alpha_{yi} - \alpha_{ref}$,
c) percentage of the ensemble members reproducing the same sign (+/-) trend as the observed trend, and
d) percentage of the ensemble members whose trend biases are small relative to standard errors of the observed and simulated trends

For a) and b), the following null hypothesis is tested per time block per region.

$$H_o: \alpha_{ref} = \alpha_Y \text{ for a).} \tag{6}$$

$$H_o: \alpha_{ref} = \alpha_{yi} \text{ for b).} \tag{7}$$

For the reference, linear trend calculation, the standard error of α_{ref} is defined as (Hogg and Tanis, 2009)):

$$s_{ref} = \sqrt{\frac{\sum_{k=1}^{m}(x_k - \hat{x}_k)^2/(m-2)}{\sum_{k=1}^{m}(t_k - \bar{t})^2}} \tag{8}$$

where

$$\hat{x}_k = \alpha_{ref} \times k + \beta_{ref} \tag{9}$$

and

$$\bar{t} = \frac{1}{m}\sum_{k=1}^{m} k \tag{10}$$

It should be noted that a), the trend bias of the MME, has high dependence on some models that contribute many ensemble members. For instance, there are models that contribute as little as one simulation or as many as 25 different ensemble members. In the case of a model contributing large ensemble simulations, this particular model will bear greater weight to the overall regional mean because each simulation is weighted equally when calculating a MME (Equation 4). Considering the unequal weights of models in α_Y, the standard error of α_Y was computed by randomly selecting *N* individual model trends with replacement (called bootstrapping) and computing the mean of that selection. We repeated this sampling 1000 times to obtain standard deviation across the 1000 random ensemble trends and use it as α_Y's standard error (s_Y). We compared $Y's$ bias ($\alpha_Y - \alpha_{ref}$) with s_{ref} and s_{ref} to test the null hypothesis (6).

To assess the statistical significance of b), the trend bias for simulation *i*, ($\alpha_{yi} - \alpha_{ref}$), it is reasonable to assume that $\alpha_{ref}$and α_{yi} likely have unequal variances. Therefore, the Welch's t-test statistic (T_i) is used to estimate the statistical significance of ($\alpha_{yi} - \alpha_{ref}$). T_i is defined as (Hogg and Tanis 2009):

$$T_i = \frac{\alpha_{yi} - \alpha_{ref}}{\sqrt{\frac{s_{ref}^2 + s_{yi}^2}{m-2}}} \tag{11}$$

Using the Welch-Satterthwaite equation, the degrees of freedom, f_i, for T_i can be approximated by:

$$f_i \approx \frac{\left(\frac{s_{ref}^2 + s_{yi}^2}{m-2}\right)^2}{\frac{s_{ref}^4 + s_{yi}^4}{(m-2)^2(m-3)^2}} \tag{12}$$

Let C^{f_i} be the cumulative density function of a student's t-distribution with f_i be the number of degrees of freedom. Then,

$$p_i = C^{f_i}(T_i) \tag{12}$$

and using p_i, the confidence level (d_i) of $\alpha_{yi} - \alpha_{ref}$ can be calculated and used as a metric:

$$\text{when } p_i < 0.5,\ d_i = (1 - 2 \times p_i) * 100[\%] \quad (13)$$

$$\text{when } p_i > 0.5,\ d_i = (2 \times p_i - 1) * 100[\%] \quad (14)$$

$$\text{when } p_i = 0.5,\ \alpha_{ref} = \alpha_{yi},\ \text{therefore } d_i = 0\% \quad (15)$$

The null hypothesis (H_o) is rejected if T_i and p_i are too small (indicating $\alpha_{yi} \ll \alpha_{ref}$), or too large (indicating $\alpha_{yi} \gg \alpha_{ref}$). In this case, α_{yi} is statistically different from α_{ref} at a confidence level of d_i. We calculated d_i of the 206 simulations for each period and region, and show a fraction of simulations whose trend biases are not statistically significant with 90% confidence level. In other words, the fraction represents how many simulations reproduce observed trends considering standard errors of the trends.

Part c) calculated the total percentage of *N* simulations in which the α_{yi} and α_{ref} have the same sign. If the product of the two trends are greater than 0, than the two carry the same warming (cooling) trend. The total tally count is divided by the number of simulations and multiplied by 100 to produce a percentage as follows:

$$f = \frac{\sum_{i=1}^{N} \mathrm{X}_i}{N} * 100 \quad \text{where } \mathrm{X}_i = \begin{cases} 1 \text{ if } \alpha_{yi} \cdot \alpha_{ref} \geq 0 \\ 0 \text{ if } \alpha_{yi} \cdot \alpha_{ref} < 0 \end{cases} \quad (16)$$

In a similar manner, part d) also produces a fraction examines the magnitude of the warming (cooling) trend. If a trend of a given simulation and its standard deviation ranges intersects with the reference and its standard error (as calculated with the equation from Hogg and Tanis [2009]), a tally is given. The total tally count is divided by the number of simulations and multiplied by 100 to produce a percentage as follows:

$$f = \frac{\sum_{i=1}^{N} \mathrm{X}_i}{N} * 100 \quad \text{where } \mathrm{X}_i = \begin{cases} 0 \text{ if } (\alpha_{yi} \pm 1\sigma) \cap (\alpha_{ref} \pm SE) = \emptyset \\ 1 \; otherwise \end{cases} \tag{17}$$

Appendix B

Summary of CMIP5 Historical Simulation Dataset

Modeling Center (Country)	**Model**	**Simulations**	**Reference**
Commonwealth Scientific and Industrial Research Organization Bureau of Meteorology (Australia)	ACCESS1-0	r1i1p1 r2i1p1	(Collier and Uhe 2012)
	ACCESS1-3	r1i1p1 r2i1p1 r3i1p1	
Beijing Climate Center (China)	bcc-csm1-1	r1i1p1 r2i1p1 r3i1p1	(Wu et al. 2014)
	bcc-csm1-1-m	r1i1p1 r2i1p1 r3i1p1	
Beijing Normal University (China)	BNU-ESM	r1i1p1	(Ji et al. 2014)
Canadian Center for Climate Modeling and Analysis (Canada)	CanCM4	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1 r6i1p1 r7i1p1 r8i1p1 r9i1p1 r10i1p1	(Chylek et al. 2011)
	CanESM2	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1	
National Center for Atmospheric Research (USA)	CCSM4	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1 r6i1p1	(Collins et al. 2004)
	CESM1-BGC	r1i1p1	
	CESM1-CAM5	r1i1p1 r2i1p1 r3i1p1	
	CESM1-FASTCHEM	r1i1p1 r2i1p1 r3i1p1	
	CESM1-WACCM	r1i1p1 r2i1p1 r3i1p1 r4i1p1	(Marsh et al. 2013)
Centro Euro-Mediterraneo sui Cambiamenti Climatici (Italy)	CMCC-CESM	r1i1p1	(Folgi and Iovino 2014)
	CMCC-CM	r1i1p1	(Scoccimarro et al., 2011)
	CMCC-CMS	r1i1p1	
Center National de Recherches Meteorologiques Center Europeen de Recherche et de Formation Avancee en Calcul Scientifique (France)	CNRM-CM5	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1 r6i1p1 r7i1p1	(Voldoire et al., 2013)

		r8i1p1 r9i1p1 r10i1p1	
	CNRM-CM5-2	r1i1p1	
Commonwealth Scientific and Industrial Research Organization Queensland Climate Change Center of Excellence (Australia)	CSIRO-Mk3-6-0	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1 r6i1p1 r7i1p1 r8i1p1 r9i1p1 r10i1p1	(Gordon et al., 2010)
EC-EARTH Consortium published at Irish Center of High-End Computing (Netherlands/Ireland)	EC-EARTH	r1i1p1 r2i1p1 r6i1p1 r7i1p1 r8i1p1 r9i1p1 r11i1p1 r12i1p1 r13i1p1 r14i1p1	(Hazeleger et al., 2012)
Institute of Atmospheric Physics Chinese Academy of Sciences (China)	FGOALS-g2	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1	(Li et al., 2013)
The First Institute of Oceanography, SOA (China)	FIO	r1i1p1 r2i1p1 r3i1p1	(Qiao et al., 2013)
Geophysical Fluid Dynamics Laboratory (USA)	GFDL-CM2p1	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1 r6i1p1 r7i1p1 r8i1p1 r9i1p1 r10i1p1	(Delworth et al., 2006)
	GFDL-CM3	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1	(Donner et al., 2011)
	GFDL-ESM2G	r1i1p1	(Dunne et al., 2013)
	GFDL-ESM2M	r1i1p1	
NASA/GISS (USA)	GISS-E2-H	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1 r6i1p1	(Schmidt et al., 2014)
	GISS-E2-H-CC	r1i1p1	
	GISS-E2-R	r1i1p1 r2i1p1 r3i1p1 r4i1p1 r5i1p1 r6i1p1 r1i1p2 r2i1p2 r3i1p2 r4i1p2 r5i1p2	

		r6i1p2	
		r1i1p3	
		r2i1p3	
		r3i1p3	
		r4i1p3	
		r5i1p3	
		r6i1p3	
		r1i1p121	
		r1i1p122	
NASA/GISS (USA)	GISS-E2-R	r1i1p124	(Schmidt et al., 2014)
		r1i1p125	
		r1i1p126	
		r1i1p127	
		r1i1p128	
	GISS-E2-R-CC	r1i1p2	
Met Office Hadley Center (UK)	HadCM3	r1i1p1	(Pope, Gallani, Rowntree, & Stratton, 2000)
		r2i1p1	
		r3i1p1	
		r4i1p1	
		r5i1p1	
		r6i1p1	
		r7i1p1	
		r8i1p1	
		r9i1p1	
		r10i1p1	
	HadGEM2-CC	r1i1p1	(W. J. Collins et al., 2011)
	HadGEM2-ES	r1i1p1	
		r2i1p1	
		r3i1p1	
		r4i1p1	
		r5i1p1	
National Institute of Meteorological Research Korea Meteorological Administration (South Korea)	HadGEM2-AO	r1i1p1	(Baek et al., 2013)
Russian Academy of Sciences Institute of Numerical Mathematics (Russia)	inmcm4	r1i1p1	(Volodin et al. 2010)
Institut Pierre Simon Laplace (France)	IPSL-CM5A-LR	r1i1p1	(Dufresne et al., 2013)
		r2i1p1	
		r3i1p1	
		r4i1p1	
		r5i1p1	
		r6i1p1	
	IPSL-CM5A-MR	r1i1p1	
		r2i1p1	
		r3i1p1	
	IPSL-CM5B-LR	r1i1p1	
Atmosphere and Ocean Research Institute (The University of Tokyo) National Institute for Environmental Studies Japan Agency for Marine-Earth Science and Technology (Japan)	MIROC-ESM	r1i1p1	(Watanabe et al. 2011)
		r2i1p1	
		r3i1p1	
	MIROC-ESM-CHEM	r1i1p1	
	MIROC4h	r1i1p1	(Sakamoto et al. 2012)
		r2i1p1	
		r3i1p1	
	MIROC5	r1i1p1	(Watanabe et al. 2010)
		r2i1p1	
		r3i1p1	
		r4i1p1	
		r5i1p1	
Max Planck Institute for Meteorology (Germany)	MPI-ESM-LR	r1i1p1	(Stevens et al., 2013)
		r2i1p1	
		r3i1p1	
	MPI-ESM-MR	r1i1p1	
		r2i1p1	
		r3i1p1	
	MPI-ESM-P	r1i1p1	
		r2i1p1	
Meteorological Research Institute (Japan)	MRI-CGCM3	r1i1p1	(Yukimoto et al. 2012)
		r2i1p1	

		r3i1p1	
	MRI-ESM1	r4i1p2 r5i1p2	(Yukimoto et al. 2011)
Bjerknes Center for Climate Research Norwegian Meteorological Institute (Norway)	NorESM1-M	r1i1p1 r2i1p1 r3i1p1	(Bentsen et al., 2013)
	NorESM1-ME	r1i1p1	

Figures

Figure 1a. Regions of analysis in this study, adapted from Janssen et al. 2014.

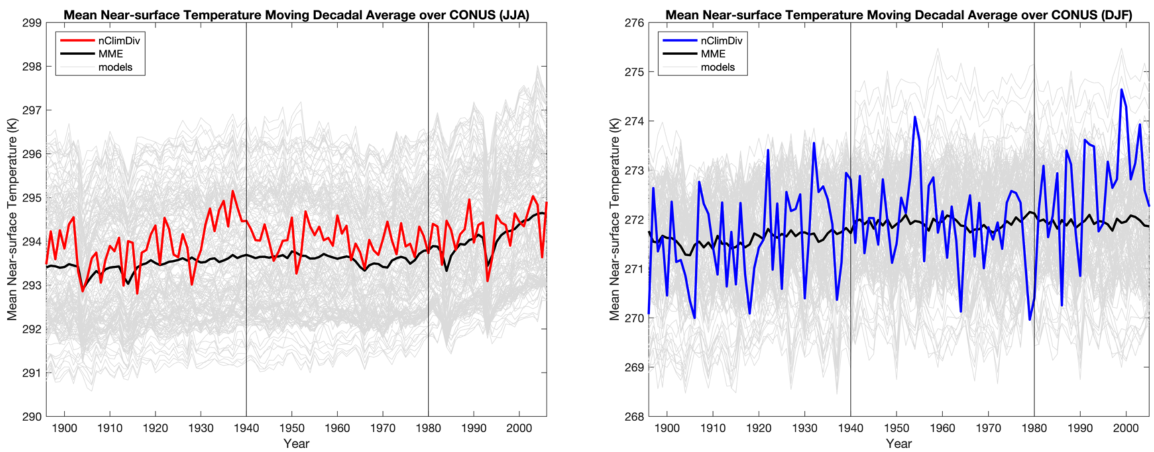


Figure 1b. Decadal moving average time series for mean near-surface air temperature of nClimGrid, MME, and all model ensembles for JJA (left) and DJF (right) from 1895-2005 separated by three distinct time periods (1895-1939, 1940-1979, 1980-2005) by a black vertical line.

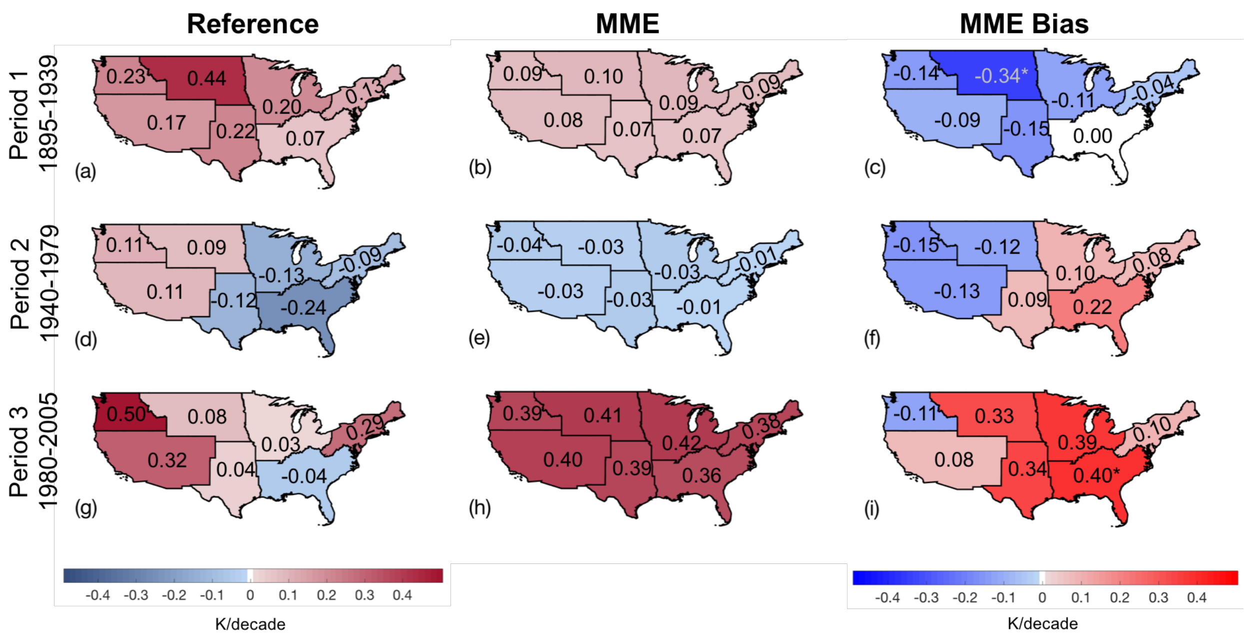


Figure 2. JJA mean near-surface air temperature decadal trends for reference (left column), MME (center column), and bias (right column) for 1895-1939 (top row), 1940-1979 (middle row), and 1980-2005 (bottom row) in K/decade. In the bias column, an asterisk in a region indicates that the difference between the MME and the observed trends is larger than their errors (See Figure 3).

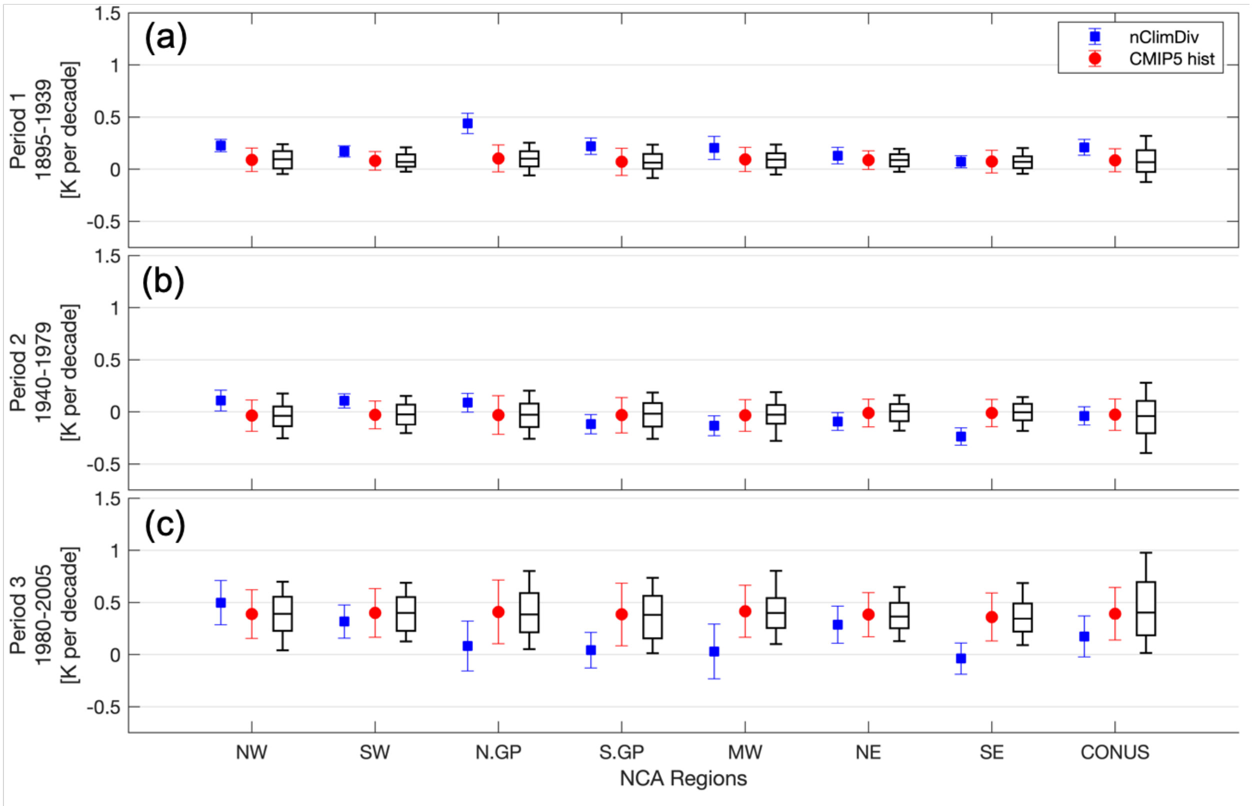


Figure 3. JJA mean near-surface air temperature decadal trend and standard error of reference (blue square), bootstrap multi-model ensemble (red circle) and standard error, and box plot of individual model simulation decadal mean trend by regions and over CONUS for 1895-1939 (top row), 1940-1979 (middle row), and 1980-2005 (bottom row) in K/decade. The line in the box represents the median ensemble member trend, the lower and upper boundary represents the 25^{th} and 75^{th} percentiles while the whiskers are the 5^{th} and 95^{th} percentiles.

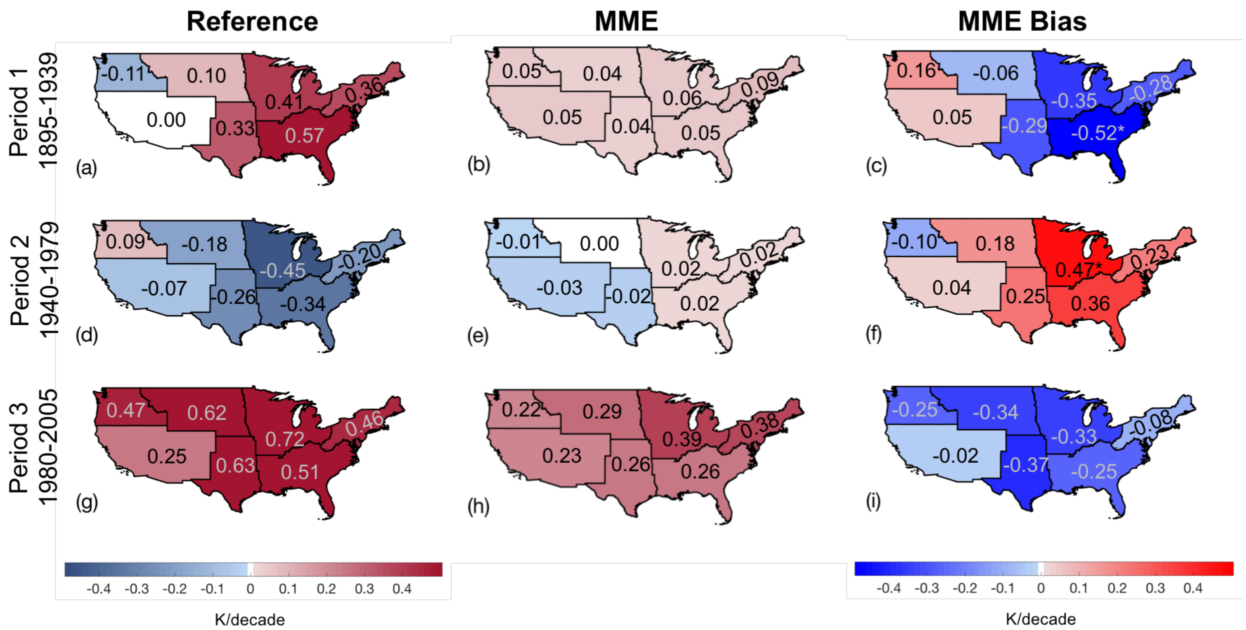


Figure 4. As in Figure 2, but for DJF.

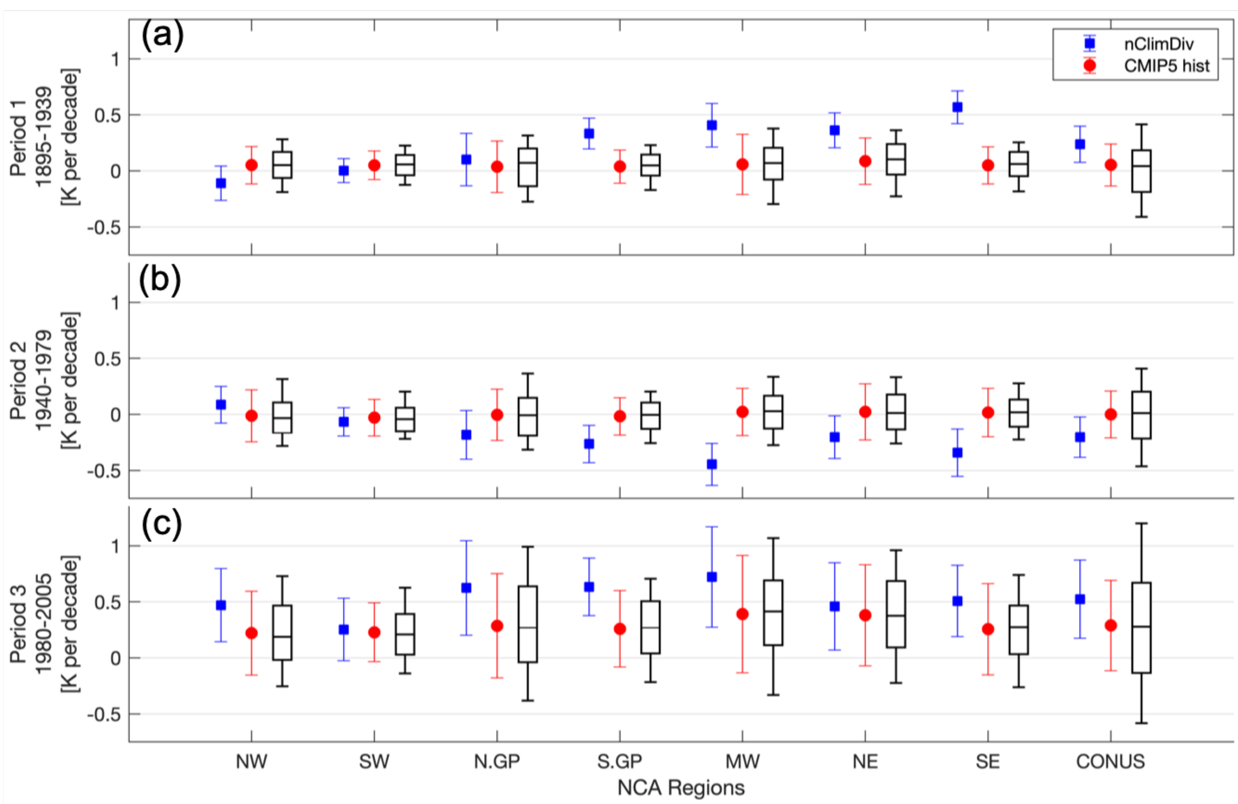


Figure 5. As in Fig. 3, but for DJF.

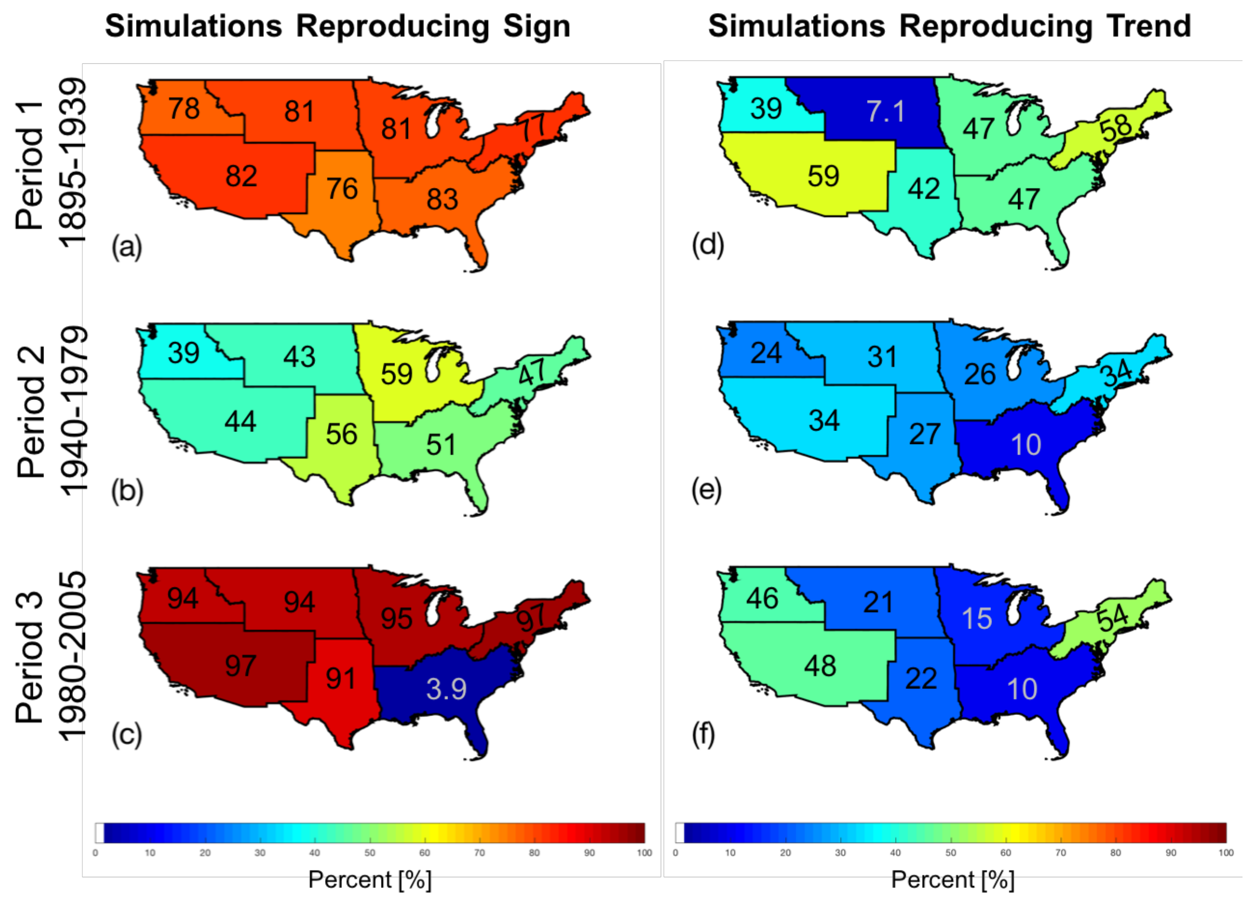


Figure 6. Percentage of JJA CMIP5 models reproducing the same sign (+/-) of mean temperature trends as the reference (left column), and percentage of JJA CMIP5 models reproducing mean temperature trend values with overlap of reference error (right column) for 1895-1939 (top row), 1940-1979 (middle row), and 1980-2005 (bottom row).

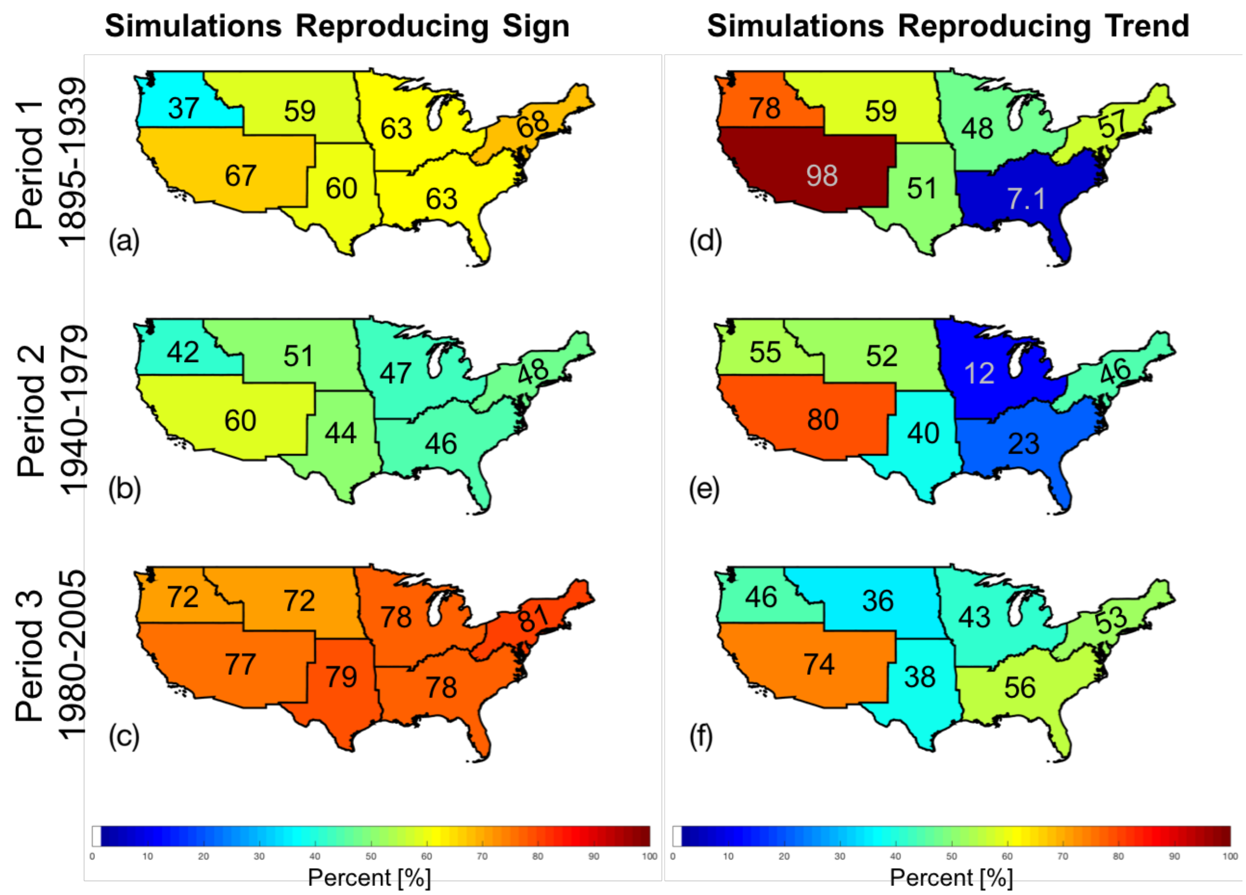


Figure 7. As in Fig. 6, but for DJF.

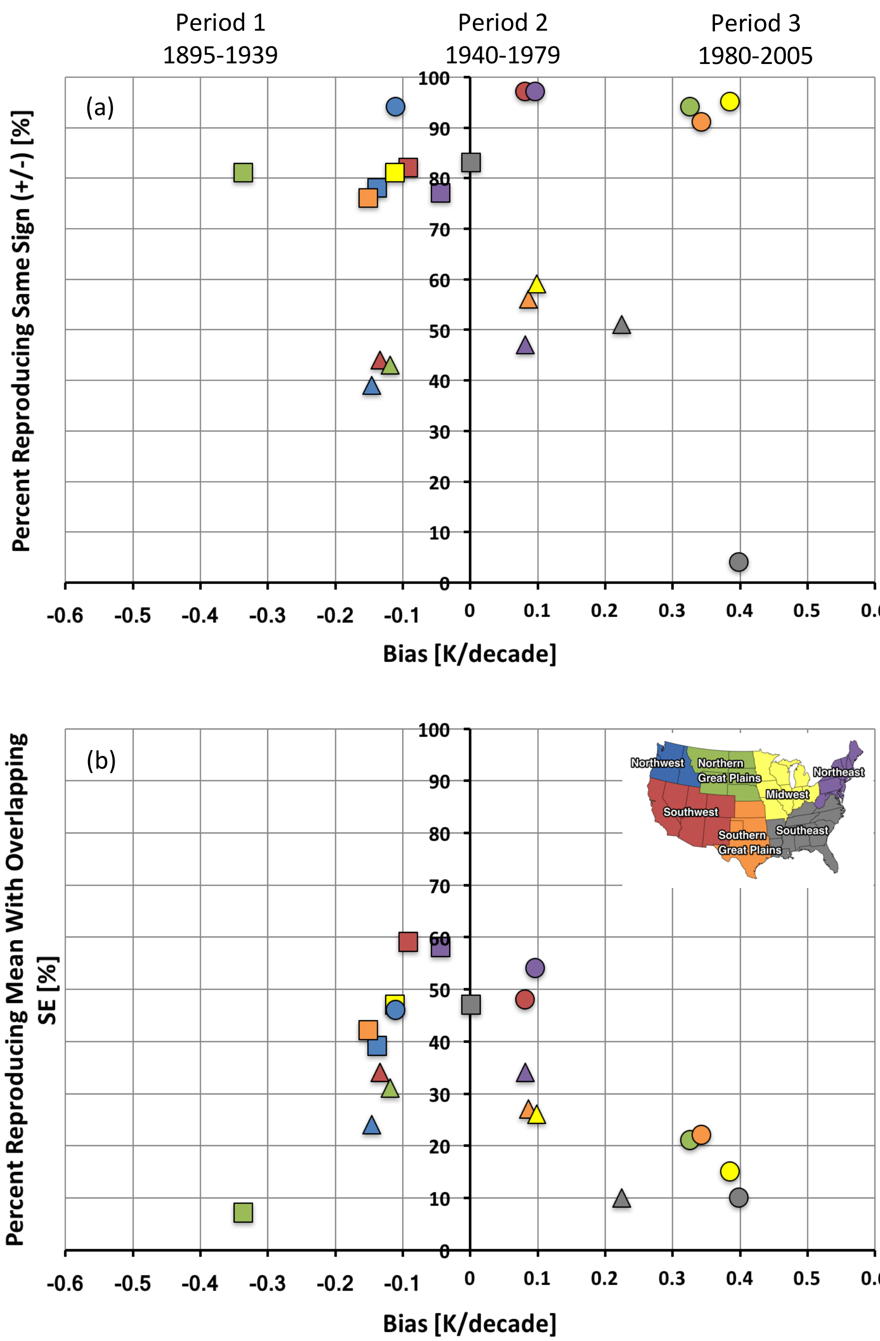


Figure 8. The top plot (Figure 8a) shows JJA mean near-surface air temperature decadal trend reference and MME bias [K/decade] (horizontal axis) and percent of CMIP5 models reproducing the same sign (+/-) of mean near-surface air temperature trends as the reference (vertical axis) for time periods 1895-1939 (square), 1940-1979 (triangle), and 1980-2005 (circle) of NCA-defined regions: Northwest (blue), Southwest (red), Northern Great Plains (green), Southern Great Plains (orange), Midwest (yellow), Northeast (purple), and Southeast (gray). The plot on the bottom (Figure 8b) is the same as the top except that the vertical axis is the percent of CMIP 5

models reproducing the mean near-surface air temperature decadal trend with overlapping standard error. The color of the data points correlates to regions represented in Figure 1a.

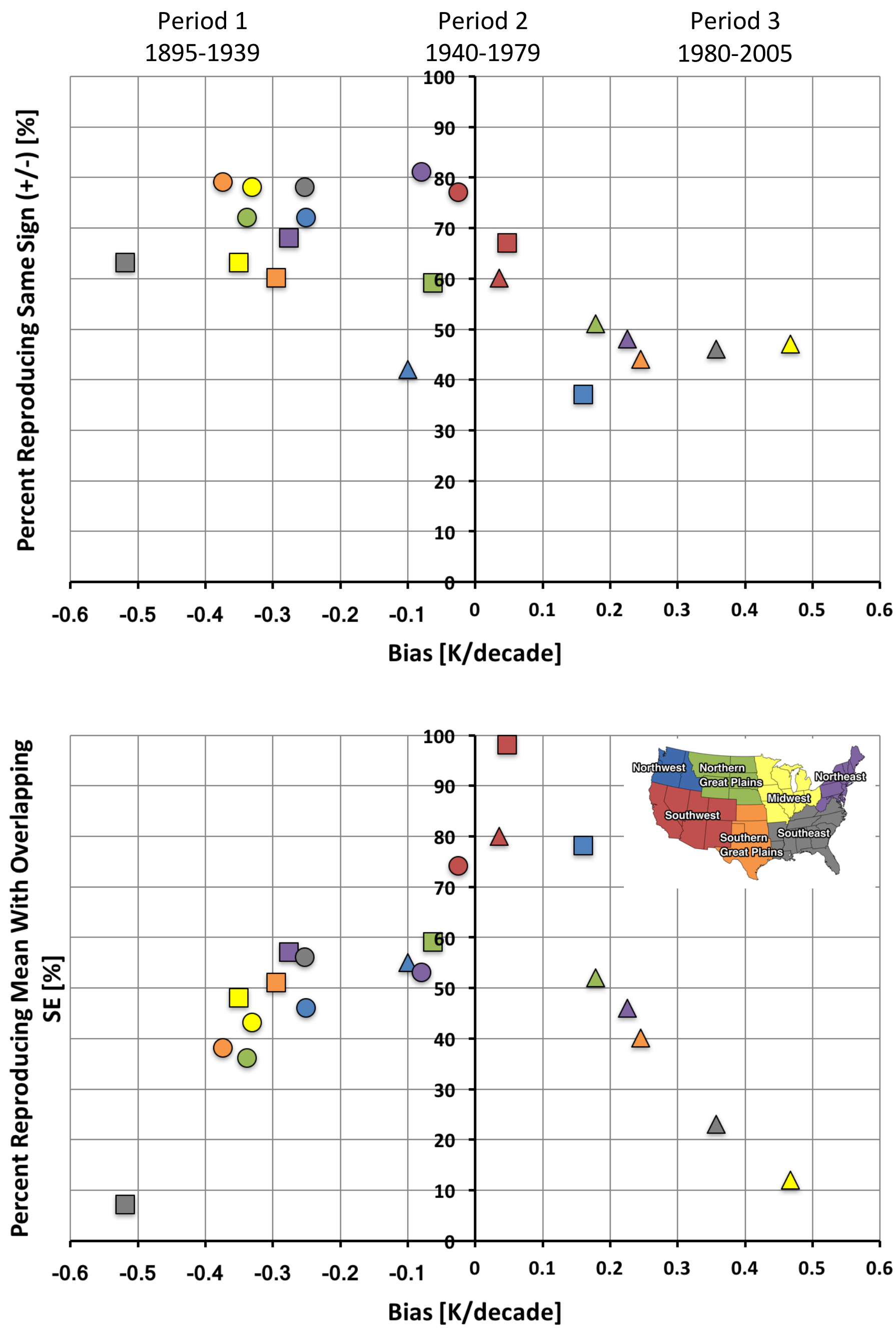


Figure 9. As in Fig. 8, but for DJF.

TABLE 1. CMIP5 analysis parameters

Variable	Time Block	*m*	Season	*N*	Variable	Time Block	*m*	Season	*N*
tas	1895-1939	45	JJA	184	pr	1895-1939	45	JJA	185
			DJF	184				DJF	185
tas	1940-1979	40	JJA	186	pr	1940-1979	40	JJA	186
			DJF	186				DJF	186
tas	1980-2005	26	JJA	206	pr	1980-2005	26	JJA	206
		25	DJF	203			25	DJF	203

Note: *m* denotes the number of years while *N* indicates the number of ensemble members

TABLE 2. Regions of analysis

Region	States
Northwest	Idaho, Oregon, Washington
Southwest	Arizona, California, Colorado, Nevada, New Mexico, Utah
Northern Great Plains	Montana, Nebraska, North Dakota, South Dakota, Wyoming
Southern Great Plains	Kansas, Oklahoma, Texas
Midwest	Illinois, Indiana, Iowa, Michigan, Minnesota, Missouri, Ohio, Wisconsin
Northeast	Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, West Virginia, District of Columbia
Southeast	Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia

TABLE 3. Percentage of JJA ensemble members whose temperature trends are not significantly different from the observed one at the 90% confidence level

Region	Period 1 1895-1939	Period 2 1940-1979	Period 3 1980-2005
Northwest	9	10	30
Southwest	14	8	27
Northern Great Plains	0	11	17
Southern Great Plains	14	8	27
Midwest	17	15	13
Northeast	0	11	17
Southeast	11	8	4
CONUS	1	16	17.5

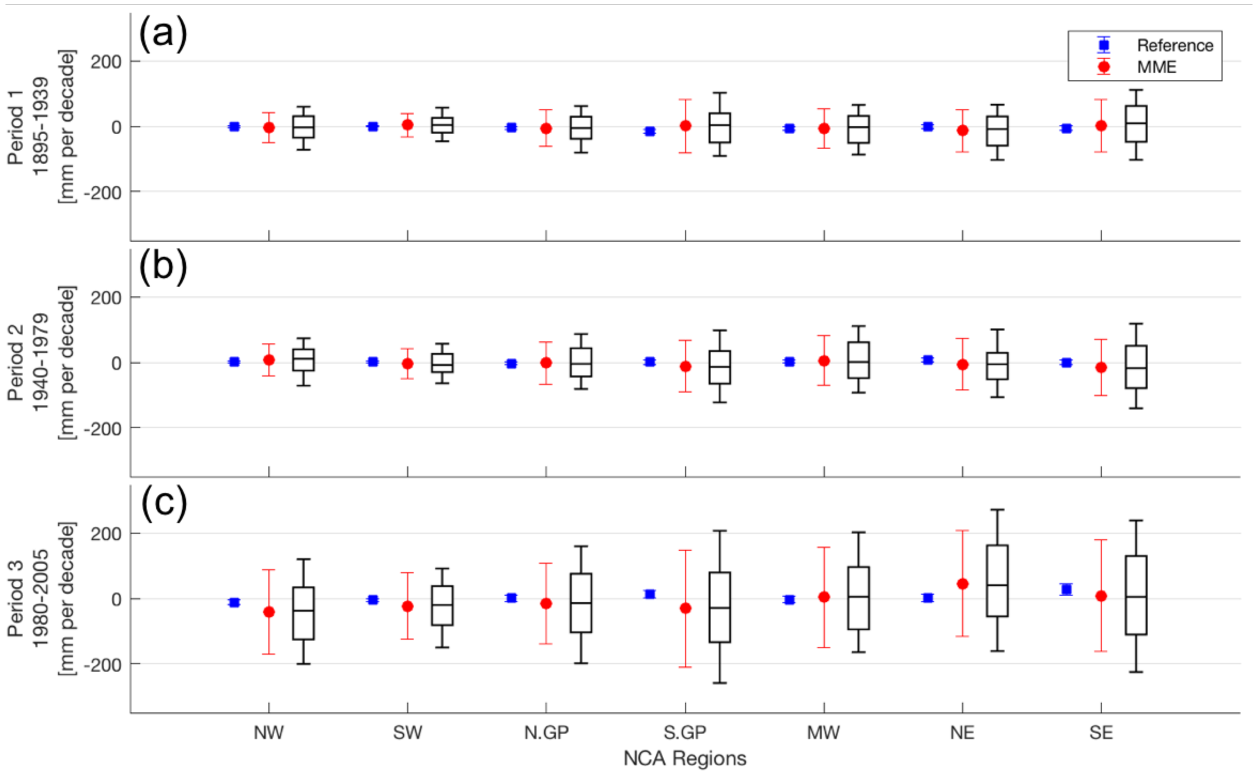
TABLE 4. Percentage of DJF ensemble members whose temperature trends are not significantly different from the observed one at the 90% confidence level

Region	Period 1 1895-1939	Period 2 1940-1979	Period 3 1980-2005
Northwest	13.5	16	25
Southwest	17	16.5	24
Northern Great Plains	35.5	29	22
Southern Great Plains	17	16.5	24
Midwest	4	4	19
Northeast	35.5	29	22
Southeast	0	10	22
CONUS	7	17	26

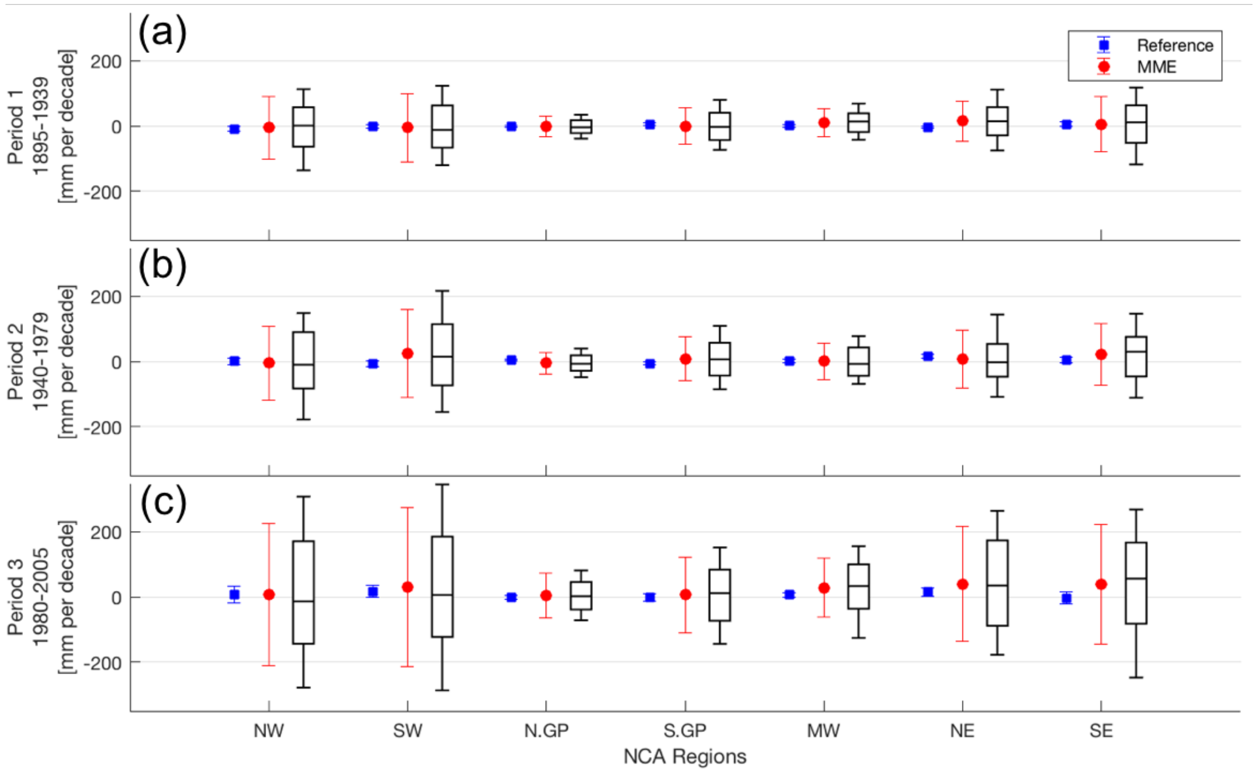
Supplementary Material

Supplementary Table 1. List of ensemble members that capture the Southeast "warming hole" during Period 2 (1940-1979) for JJA and DJF.

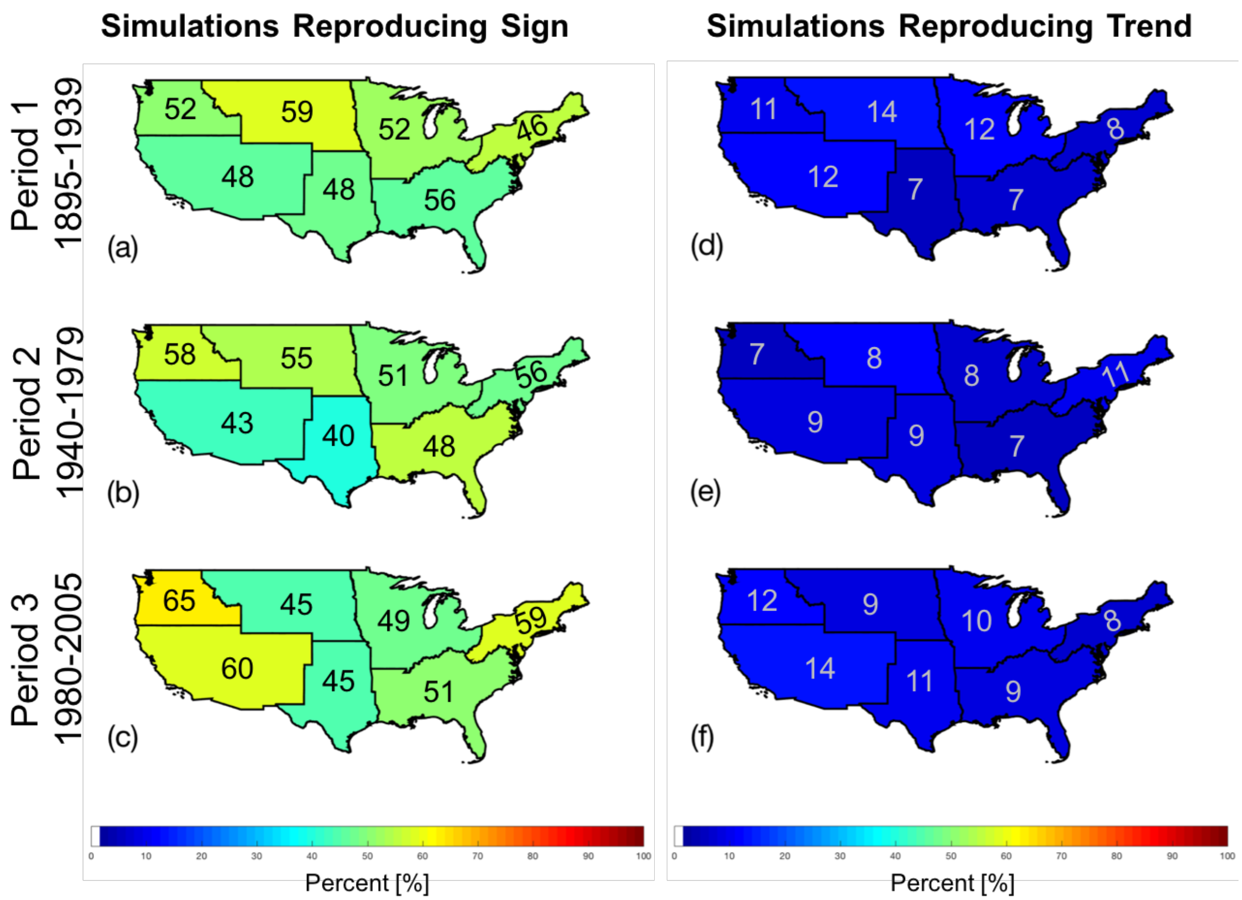
Model	Simulation	JJA	DJF
ACCESS1-0	r2i1p1	✓	✓
ACCESS1-3	r2i1p1	✓	✓
bcc-csm1-1-m	r3i1p1	✓	
CanESM2	r1i1p1		✓
	r3i1p1		✓
CCSM4	r1i1p1		✓
	r5i1p1		✓
CESM1-FASTCHEM	r1i1p1		✓
CMCC-CM	r1i1p1		✓
CNRM-CM5	r1i1p1		✓
	r2i1p1		✓
	r5i1p1		✓
	r9i1p1		✓
	r10i1p1		✓
CSIRO-Mk3-6-0	r1i1p1		✓
	r2i1p1		✓
	r3i1p1	✓	
	r6i1p1		✓
	r8i1p1	✓	✓
FGOALS-g2	r5i1p1	✓	✓
GFDL-CM2p1	r2i1p1		✓
	r7i1p1		✓
	r8i1p1		✓
GFDL-CM3	r1i1p1	✓	
	r4i1p1	✓	
	r5i1p1	✓	
GFDL-ESM2M	r1i1p1		✓
GISS-E2-H	r1i1p1		✓
	r2i1p1		✓
	r5i1p3	✓	✓
	r6i1p3	✓	
GISS-E2-R	r1i1p1		✓
	r2i1p1		✓
	r2i1p3		✓
	r4i1p1		✓
	r4i1p2		✓
	r6i1p1		✓
	r6i1p2	✓	✓
HadCM3	r3i1p1		✓
	r4i1p1		✓
HadGEM2-ES	r1i1p1	✓	✓
	r2i1p1	✓	
	r3i1p1		✓
	r5i1p1	✓	✓
IPSL-CM5A-MR	r1i1p1		✓
	r3i1p1	✓	
MIROC-ESM	r3i1p1		✓
MIROC-ESM-CHEM	r1i1p1	✓	✓
MIROC5	r2i1p1	✓	
	r5i1p1	✓	
MRI-CGCM3	r4i1p2		✓
NorESM1-M	r2i1p1		✓



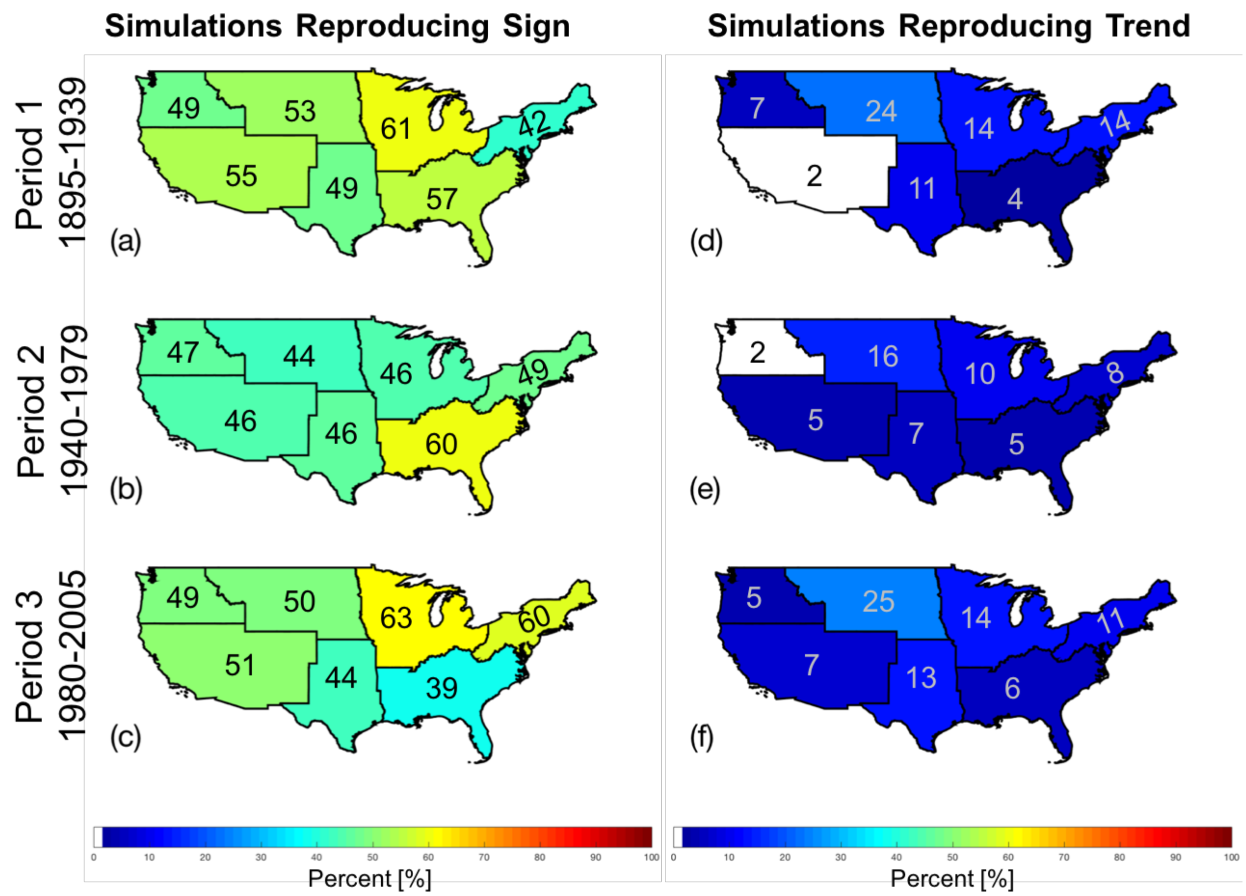
Supplementary Figure 1. JJA mean precipitation decadal trend and standard error of reference (blue square), bootstrap multi-model ensemble (red circle) and standard error, and box plot of individual model simulation decadal mean trend by regions for 1895-1939 (top row), 1940-1979 (middle row), and 1980-2005 (bottom row) in K/decade. The line in the box represents the median ensemble member trend, the lower and upper boundary represents the 25^{th} and 75^{th} percentiles while the whiskers are the 5^{th} and 95^{th} percentiles.



Supplementary Figure 2. As in Supplementary Figure 1, but for DJF.



Supplementary Figure 3. Percentage of JJA CMIP5 models reproducing the same sign (+/-) of mean precipitation trends as the reference (left column), and percentage of JJA CMIP5 models reproducing mean precipitation trend values with overlap of reference error (right column) for 1895-1939 (top row), 1940-1979 (middle row), and 1980-2005 (bottom row).



Supplementary Figure 4. As in Supplementary Figure 2, but for DJF.