Intermodel Variability and Mechanism Attribution of Central and Southeastern U.S. Anomalous Cooling in the Twentieth Century as Simulated by CMIP5 Models

ZAITAO PAN
Department of Earth and Atmospheric Sciences, Saint Louis University, St. Louis, Missouri

XIAODONG LIU
Institute of Earth Environment, Chinese Academy of Sciences, Xi’an, Shaanxi, China

SANJIV KUMAR
Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

ZHIQIU GAO
Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

JAMES KINTER
Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

(Manuscript received 31 July 2012, in final form 18 February 2013)

ABSTRACT

Some parts of the United States, especially the southeastern and central portion, cooled by up to 2°C during the twentieth century, while the global mean temperature rose by 0.6°C (0.76°C from 1901 to 2006). Studies have suggested that the Pacific decadal oscillation (PDO) and the Atlantic multidecadal oscillation (AMO) may be responsible for this cooling, termed the “warming hole” (WH), while other works reported that regional-scale processes such as the low-level jet and evapotranspiration contribute to the anomaly. In phase 3 of the Coupled Model Intercomparison Project (CMIP3), only a few of the 53 simulations could reproduce the cooling. This study analyzes newly available simulations in experiments from phase 5 of the Coupled Model Intercomparison Project (CMIP5) from 28 models, totaling 175 ensemble members. It was found that 1) only 19 out of 100 all-forcing historical ensemble members simulated negative temperature trend (cooling) over the southeast United States, with 99 members underpredicting the cooling rate in the region; 2) the missing of cooling in the models is likely due to the poor performance in simulating the spatial pattern of the cooling rather than the temporal variation, as indicated by a larger temporal correlation coefficient than spatial one between the observation and simulations; 3) the simulations with greenhouse gas (GHG) forcing only produced strong warming in the central United States that may have compensated the cooling; and 4) the all-forcing historical experiment compared with the natural-forcing-only experiment showed a well-defined WH in the central United States, suggesting that land surface processes, among others, could have contributed to the cooling in the twentieth century.

1. Introduction

Earth’s surface has experienced unprecedented warming since the Industrial Revolution began in the 1850s. The global mean surface air temperature over land rose 0.76°C during 1901–2006 (Solomon et al. 2007). This global warming has been neither spatially uniform nor persistent in time. The warming is faster in the high latitudes than in the tropics and greater in winter than summer, largely due to snowmelt–albedo feedbacks (Holland and Bitz 2003). It is also widely reported that the nighttime temperature rose more than the daytime
temperature because of cloud cover and other feedback processes (Karl et al. 1993). Furthermore, high mountain regions warmed more than low-lying regions (Liu and Chen 2000).

On regional scales, temperature changes often deviate from the above discussed general warming patterns. There are some special geographical regions where a lack of warming or even a cooling has occurred. The central and southeastern United States (CSE) actually cooled in the twentieth century, most notably during the second half of the century, while global mean temperature warmed at an increasing rate. The cooling or lack of warming regions is referred to as “warming holes” (WHs) (Pan et al. 2004; Kunkel et al. 2006). Attention has been paid to this abnormal cooling trend both observationally and in modeling (Tett et al. 2002; Portmann et al. 2009; Meehl et al. 2012). Kalnay and Cai (2003) have attributed this cooling trend to land surface processes by reconciling the temperature difference between upper-air and surface observations. Combining observations with a regional climate model’s results, Pan et al. (2004) suggested that regional hydrological processes coupled with the low-level jet contribute to the cooling. Other studies have attributed the cooling to the internal dynamics (Kunkel et al. 2006; Liang et al. 2006; Weaver 2013). A number of modeling studies have attributed the mechanisms for this abnormal trend to large-scale decadal oscillations such as the Pacific decadal oscillation (PDO) and the Atlantic multidecadal oscillation (AMO) (Robinson et al. 2002; Kunkel et al. 2006; Wang et al. 2009; Meehl et al. 2012).

While seeking reasons why only this part of the United States experienced cooling, Pan et al. (2009) found other similar WHs: one in south-central China and the other in central South America. Some features common to these WHs are their presence 1) on the eastern slope of major mountain ranges where the climatic warming gradient exists, 2) at low-level jet termini where warm moist air converges, and 3) in intense agricultural regions where the deep crop roots can extract soil moisture.

The midcontinental cooling goes against the common belief that the middle of continents, far from oceans, should warm faster than coastal regions. Also, it was a challenge for the great majority of models in phase 3 of the Coupled Model Intercomparison Project (CMIP3) to reproduce the WHs (Kunkel et al. 2006). It is of interest to see how well phase 5 of the Coupled Model Intercomparison Project (CMIP5; Taylor et al. 2012) models reproduce WHs in the twentieth-century simulations as well as how they predict the WH fate in the twenty-first century. The purposes of this paper are to 1) see how well CMIP5 models reproduce this abnormal cooling, 2) find the common features of models that simulated WHs well (or not), and 3) determine what mechanisms are responsible for the WHs as simulated in the new generation of the atmosphere–ocean coupled general circulation models (AOGCMs). A companion paper (Kumar et al. 2013a) investigates east–west gradient and multidecadal aspects of the WH trends in North America as in Meehl et al. (2012). More general results regarding North American climate in the CMIP5 models are reported in Sheffield et al. (2013a,b).

### 2. Model and data

The design of CMIP5 includes the new short-term decadal experiments hindcasting the interannual variability, emission-driven (versus concentration driven) earth system model (ESM) simulations exploring the sensitivity of the carbon cycle feedback, and time-evolving land use runs allowing for the dynamic vegetation feedback (Taylor et al. 2012). The core long-term CMIP5 simulations include historical and projection experiments. The historical experiments include all-forcing (historical), greenhouse gas (GHG) forcing only (historicalGHG), natural forcing only (historicalNat), and other specific forcing (such as aerosols). The projection experiments consist of four new representative concentration pathway (RCP) emission scenarios, RCP2.6 through RCP8.5, representing anthropogenic radiative forcing stabilizing at $2.6-8.5$ W m$^{-2}$ by 2100 (Moss et al. 2010). In this study, we focus on the all-forcing historical and RCP4.5 experiments, with limited exploration of the historicalGHG and historicalNat (natural forcing, such as volcanic eruption, only) runs. The historical runs are forced by observed atmospheric composition changes (reflecting both anthropogenic and natural sources). The historicalGHG and historicalNat runs are the same as the historical run except that they are forced by greenhouse gases alone or natural variability alone. The temporal span of the historical experiment covers 1850–2005, and thus is sometimes referred to as “20th century” simulations (Taylor et al. 2012). The RCP4.5 scenario assumes that anthropogenic forcing will essentially level off at 4.5 W m$^{-2}$ around the mid-twenty-first century and represents the intermediate range of the four scenarios. RCP4.5 runs cover 2005–2100 (some model groups extend it to 2300; see Thomson et al. 2011 for details).

In this study, we analyze all model ensemble members presently available in the historical and RCP4.5 experiments, totaling 28 models and 175 members. The historical experiment has 25 models available with 100 members and the RCP4.5 experiment encompasses 22 models with 63 members. The historicalNat and historicalGHG runs each have six single-member simulations. Fifteen out of 25 models in the historical experiment and 8 out
of 22 models in the RCP4.5 experiment have multiple ensemble members of 2–16 (Table 1). Monthly mean of daily maximum and minimum surface temperatures\(^1\) from all models were mapped to a \(1^\circ \times 1^\circ\) grid, the highest resolution of the models. The linear trends are computed based on a least squares regression. The horizontal resolution of the models ranges from 3.75\(^\circ\) to 1.0\(^\circ\).

The observed daily temperatures used in this study were obtained from the Global Historical Climatology Network (GHCN) as compiled into monthly means and interpolated onto regular latitude/longitude grids by the Climate Research Unit (CRU). The dataset includes monthly mean surface daily maximum and minimum temperatures on a \(0.5^\circ \times 0.5^\circ\) latitude–longitude grid over land for the period 1901–2009 (New et al. 2000; Vose et al. 2005). Since data stations before the 1950s were somewhat sparse (see New et al. 2000 for details), our analyses mainly focus on the temperature changes after the 1950s, although prior temperatures are also used to determine longer-term trends.

### 3. Observed cooling characteristics

Since temperature variations are not monotonic but rather fluctuate, the trend values will depend on the evaluation periods. While longer periods can give larger sample sizes, they may obscure underlying physical processes during different periods. For example, the second half of the twentieth century, an often-used period of recent studies (Wang et al. 2009), includes a period of global slight cooling before 1975 and a strong warming period after that. To reduce the effect of arbitrarily choosing the lengths of periods, we evaluated trends in three durations: 100 (1901–2000), 50 (1951–2000), and 25 yr (1951–75 and 1976–2000). The 100-yr duration represents the longest available dataset and the 50-yr period corresponds to the data-rich period. The separation of the second half of the century into two equal 25-yr periods is not chosen for simplicity, but is based on following considerations: 1) The year 1976 is around the turning point of two climate epochs when the PDO shifted from a negative to a positive phase (Miller et al. 1994); 2) the global temperature trend changed from a slight decrease to a strong increase around 1975 (Folland et al. 2002); and 3) the year 1979 was the beginning of the satellite era when the incorporation of satellite data introduces some discontinuity in datasets (Kalnay and Cai 2003).

---

\(^1\) Most studies of GCM intercomparison use mean surface air temperature, masking the difference in maximum and minimum temperatures.
<table>
<thead>
<tr>
<th>Model/MemID</th>
<th>Full model name</th>
<th>Model abbreviation</th>
<th>Modeling center</th>
<th>Members (historical)</th>
<th>Members (RCP4.5)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/0</td>
<td>Australian Community Climate and Earth-System Simulator, version 1.0</td>
<td>ACCESS1.0</td>
<td>Commonwealth Scientific and Industrial Research Organisation (CSIRO) and Bureau of Meteorology (BOM), Australia</td>
<td>1</td>
<td>1</td>
<td><a href="http://wiki.csiro.au/confluence/display/ACCESS">http://wiki.csiro.au/confluence/display/ACCESS</a></td>
</tr>
<tr>
<td>2/1–3</td>
<td>Beijing Climate Center, Climate System Model, version 1.1</td>
<td>BCC CSM1.1</td>
<td>Beijing Climate Center (BCC), China Meteorological Administration, China</td>
<td>3</td>
<td>1</td>
<td>Wu et al. 2010</td>
</tr>
<tr>
<td>3/4–8</td>
<td>Second Generation Canadian Earth System Model</td>
<td>CanESM2</td>
<td>Canadian Centre for Climate Modeling and Analysis (CCCma), Canada</td>
<td>5</td>
<td>5</td>
<td>Arora et al. 2011</td>
</tr>
<tr>
<td>4/9–14</td>
<td>Community Climate System Model, version 4</td>
<td>CCSM4</td>
<td>National Center for Atmospheric Research (NCAR), United States</td>
<td>6</td>
<td>5</td>
<td>Gent et al. 2011</td>
</tr>
<tr>
<td>5/15-22</td>
<td>Centre National de Recherches Météorologiques Coupled Global Climate Model, version 5.1</td>
<td>CNRM-CM5.1</td>
<td>Centre National de Recherches Météorologiques (CNRM), France</td>
<td>8</td>
<td>1</td>
<td>Voldoire et al. 2013</td>
</tr>
<tr>
<td>6/23–32</td>
<td>Commonwealth Scientific and Industrial Research Organisation Mark, version 3.6.0</td>
<td>CSIRO-Mk3.6.0</td>
<td>CSIRO and Climate Change Centre of Excellence, Australia</td>
<td>10</td>
<td>10</td>
<td>Rotstayn et al. 2010</td>
</tr>
<tr>
<td>7/33–35</td>
<td>Flexible Global Ocean–Atmosphere–Land System Model gridpoint, second spectral version</td>
<td>FGOALS-s2</td>
<td>State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics (IAP), Chinese Academy of Sciences</td>
<td>3</td>
<td>—</td>
<td>Zhou et al. 2005</td>
</tr>
<tr>
<td>8/36–39</td>
<td>Geophysical Fluid Dynamics Laboratory Climate Model, version 3</td>
<td>GFDL CM3</td>
<td>National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL), United States</td>
<td>4</td>
<td>—</td>
<td>Donner et al. 2011</td>
</tr>
<tr>
<td>9/40</td>
<td>GFDL Earth System Model with Generalized Ocean Layer Dynamics (GOLD) component (ESM2G)</td>
<td>GFDL-ESM2G</td>
<td>As in GFDL-CM3</td>
<td>1</td>
<td>—</td>
<td>Dunne et al. 2012</td>
</tr>
<tr>
<td>10/41</td>
<td>GFDL Earth System Model with Modular Ocean Model 4 (MOM4) component (ESM2M)</td>
<td>GFDL-ESM2M</td>
<td>As in GFDL-CM3</td>
<td>1</td>
<td>1</td>
<td>As in GFDL-CM3</td>
</tr>
<tr>
<td>11/42–56</td>
<td>Goddard Institute for Space Studies Model E, coupled with the HYCOM ocean model</td>
<td>GISS-E2H</td>
<td>National Aeronautics and Space Administration (NASA)/Goddard Institute for Space Studies (GISS), United States</td>
<td>15</td>
<td>—</td>
<td>Schmidt et al. 2006</td>
</tr>
<tr>
<td>12/57–72</td>
<td>GISS Model E, coupled with the Russell ocean model</td>
<td>GISS-E2-R</td>
<td>As in GISS-E2-H</td>
<td>16</td>
<td>15</td>
<td>Schmidt et al. 2006</td>
</tr>
<tr>
<td>13/73</td>
<td>Hadley Centre Coupled Model, version 3</td>
<td>HadCM3</td>
<td>Met Office Hadley Centre, United Kingdom</td>
<td>1</td>
<td>—</td>
<td>Collins et al. 2001</td>
</tr>
<tr>
<td>14/74</td>
<td>Hadley Centre Global Environment Model, version 2–Atmosphere and Ocean</td>
<td>HadGEM2-AO</td>
<td>As in HadCM3</td>
<td>1</td>
<td>1</td>
<td>Jones et al. 2011</td>
</tr>
<tr>
<td>15/75</td>
<td>Hadley Centre Global Environment Model, version 2–Carbon Cycle</td>
<td>HadGEM2-CC</td>
<td>As in HadCM3</td>
<td>1</td>
<td>1</td>
<td>As in HadCM3</td>
</tr>
<tr>
<td>Model/ MemID</td>
<td>Full model name</td>
<td>Model abbreviation</td>
<td>Modeling center</td>
<td>Members (historical)</td>
<td>Members (RCP4.5)</td>
<td>Reference</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>--------------------</td>
<td>-----------------------------------------------------</td>
<td>----------------------</td>
<td>------------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>16/76</td>
<td>Hadley Centre Global Environment Model, version 2–Earth System</td>
<td>HadGEM2-ES</td>
<td>As in HadCM3</td>
<td>1</td>
<td>4</td>
<td>As in HadCM3</td>
</tr>
<tr>
<td>17/77-81</td>
<td>L’Institut Pierre-Simon Laplace Coupled Model, version 5, coupled with NEMO, low resolution</td>
<td>IPSL-CM5A-LR</td>
<td>Institut Pierre Simon Laplace (IPSL), France</td>
<td>5</td>
<td>4</td>
<td>Hourdin et al. (2006)</td>
</tr>
<tr>
<td>18/82-83</td>
<td>IPSL Coupled Model, version 5, coupled with NEMO, medium resolution</td>
<td>IPSL-CM5A-MR</td>
<td>As in IPSL-CM5A-LR</td>
<td>2</td>
<td>1</td>
<td>As in IPSL-CM5A-LR</td>
</tr>
<tr>
<td>19/84-7</td>
<td>Model for Interdisciplinary Research on Climate, version 5</td>
<td>MIROC5</td>
<td>Atmosphere and Ocean Research Institute and Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Japan</td>
<td>4</td>
<td>3</td>
<td>Watanabe et al. 2010</td>
</tr>
<tr>
<td>20/88</td>
<td>Model for Interdisciplinary Research on Climate, Earth System Model</td>
<td>MIROC-ESM</td>
<td>As in MIROC5</td>
<td>1</td>
<td>1</td>
<td>As in MIROC5</td>
</tr>
<tr>
<td>21/89</td>
<td>Model for Interdisciplinary Research on Climate, Earth System Model, Chemistry Coupled</td>
<td>MIROC-ESM-CHEM</td>
<td>As in MIROC5</td>
<td>1</td>
<td>1</td>
<td>As in MIROC5</td>
</tr>
<tr>
<td>22/90-92</td>
<td>Max Planck Institute Earth System Model, low resolution</td>
<td>MPI-ESM-LR</td>
<td>Max Planck Institute for Meteorology, Germany</td>
<td>3</td>
<td>3</td>
<td>Raddatz et al. 2007</td>
</tr>
<tr>
<td>23/93-94</td>
<td>Max Planck Institute Earth System Model, paleo</td>
<td>MPI-ESM-P</td>
<td>As in MPI-ESM-LR</td>
<td>2</td>
<td>—</td>
<td>As in MPI-ESM-LR</td>
</tr>
<tr>
<td>24/95-98</td>
<td>Meteorological Research Institute Coupled Atmosphere–Ocean General Circulation Model, version 3</td>
<td>MRI-CGCM3</td>
<td>Meteorological Research Institute, Japan</td>
<td>4</td>
<td>1</td>
<td>Yukimoto et al. 2011</td>
</tr>
<tr>
<td>25/99</td>
<td>Norwegian Earth System Model, version 1 (intermediate resolution)</td>
<td>NorESM1-M</td>
<td>Norwegian Climate Center, Norway</td>
<td>1</td>
<td>1</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Vose et al. (2005). Global annual Tmin over land increased 0.20°C decade$^{-1}$ while Tmax increased 0.14°C decade$^{-1}$ from 1950 to 2004, resulting in a DTR decrease (−0.07°C decade$^{-1}$; Vose et al. 2005). During the same period over North American land (15°–75°N, 175°–60°W), summer Tmax and Tmin increased by 0.07 and 0.12°C decade$^{-1}$, respectively, resulting in −0.05°C decade$^{-1}$ in DTR change. A similar decrease
(−0.06°C decade⁻¹) occurred in winter. Over the central and southeast United States (30°−45°N, 105°−80°W), summer Tmax actually decreased sharply (−0.13°C decade⁻¹), while Tmin increased slightly (0.05°C decade⁻¹), yielding a DTR decrease of 0.18°C decade⁻¹. Winter DTR also decreased by 0.13°C decade⁻¹.

Figure 3 shows the time series of surface temperature over CN and SE regions delineated in Fig. 2. The U.S. temperature (both CN and SE) of the twentieth century is characterized by the hot Dust Bowl period in 1930s when the Southern Plains experienced extreme droughts and thus dusty soil for eight years (Schubert et al. 2004), followed by slight cooling until the mid-1970s, and fast warming after that. This overall pattern is similar to the global mean, but with large fluctuations. Separating temperatures into summer and winter as well as Tmax and Tmin shows that the U.S. temperatures deviate from the global ones notably. In terms of summer daytime Tmax, the 1930s is still over 1°C warmer than other decades, including the globally warmest 2010s. On the other hand, winter Tmin was greatest in the later 1990s and early 2000s. Globally, the temporal variations of...
Tmax and Tmin during winter and summer follow similar patterns (dotted blue lines), while those of the U.S. (both CN and SE) deviate from each other notably. Winter Tmin trends follow the global mean quite well (lower right panel, Fig. 3). However, the Tmax trends deviate from the global mean (top left). During 1985–95, when global warming was quite fast, the U.S. Tmax decreased sharply, indicative of the complex forcing in daytime during summer.

4. CMIP5 model-simulated twentieth-century cooling—Historical experiment

Here we present the model simulated temperature variations from 100 members of 25 models in the historical experiment. In this section, presentation is based on individual members (not model based), meaning that ensemble means are contributed by each member equally. This way we can avoid the mutual cancellation of temperature variations among ensemble members within a model.

a. Model ensemble means

On the 100-yr scale, the 100-member mean seems to mimic the pattern of the SE cooling by showing slightly less warming than surroundings in winter (top panels, Fig. 4). The region warms by <0.1°C decade⁻¹, which is less than the warming in the rest of country. The presence of a WH is less evident in summer than winter. On the 50-yr scale (1951–2000), the models showed no cooling during summer in the central and southeast United States (middle panels). The winter pattern shows a clear north–south (N–S) warming gradient as observed. On the quarter-century scale (1976–2000), the models simulated more extensive warming in both summer and winter (bottom panels). The trend variance among the 100 members generally follows the trend magnitude itself (contours in Fig. 4). Larger variance tends to coincide with larger warming or cooling values themselves and vice versa. This seems to suggest that the weak warming or slight cooling over WH region is unlikely due to the cancellations of strong warming and cooling among the members, offering some confidence in the simulations.

b. Intermodel variability

The model ensemble mean represents all individual members that simulated diverse temperature patterns. As an example, Fig. 5 depicts the 25 individual Tmax trend during the 50-yr period (1951–2000) simulated by the first ensemble member of each model. About half of the models simulated a variant of the WH pattern (less warming) although simulated WH locations vary across models. For example, CanESM2 shows a cooling trend in the central United States, consistent with the observation, whereas GISS-E2-H model simulates a cooling trend in the western United States (expansions of all model names can be found in Table 1). A couple of models (e.g., HadGEM3-AO and IPSL-CM5A-LR)
FIG. 4. Twenty-five-model mean (100 members) of linear trends (°C decade\(^{-1}\)) of mean surface air temperature on (top) 100-, (middle) 50-, and (bottom) 25-yr time scales of the twentieth century for (left) summer and (right) winter. Each model equally contributes to the ensemble mean regardless of whether it has multiple or single ensemble members. The contours are the variance (°C decade\(^{-1}\)) among all the model ensemble members. The H and L denote the high and low centers of the variance, respectively.
show an excessive warming maximum around the observed WH region. As expected, within individual model families, the trend patterns are more similar. For example, the HadGEM2 family consisting of three models-versions (-AO, -CC, and -ES) generally simulated mid-Atlantic warming and Pacific Northwest cooling. The only exception to the model family similarity is GFDL-ESM2 (-G and -M), whose two versions simulated opposite trends in the southeastern coastal region.

To quantify the model skill in reproducing the WH phenomena, Fig. 6 shows the trends of 25 models’ first ensemble member in summer and winter for Tmax and Tmin averaged over the SE WH region. On the century scale in summer (top left panel), the observed cooling only occurred in Tmax (rightmost red bar denoted by an O on the x axis). Eight out of 25 models simulated negative trends ranging from 0.005°C to 0.06°C decade^{-1} in summer. The remaining models simulated warming trends from 0.001°C to 0.20°C decade^{-1}. The all-model mean is +0.06°C decade^{-1}. The observed winter temperatures in the SE WH region warmed during the century by 0.01°C–0.02°C decade^{-1} (top right panel). The model means simulated slightly stronger warming (0.03°C–0.04°C decade^{-1}) although a number of models simulated larger positive and negative trends.

On the 50-yr scale (bottom panels), the observed cooling on Tmax reached −0.17°C decade^{-1} in summer. The majority of models simulated warming on both Tmax (and Tmin) with an all-model mean of +0.13°C decade^{-1} (marked by an M on the x axis). Only six models produced
negative trends of Tmax (bottom left). In winter, eight models simulated sizeable negative trends of temperatures.

The intermodel variability is further quantified using a box-and-whisker plot (Fig. 7). All the model medians are positive for both periods in winter and summer. Note that Tmax is slightly more dispersed than Tmin, especially on the 100-yr scale in summer. On the 100-yr scale, the summer middle 50 percentile spans about 0.04°C decade⁻¹ for Tmin and 0.08°C decade⁻¹ for Tmax. On the 50-yr scale, middle half summer Tmax warming is around 50% more dispersed than Tmin. The skews are generally small.

c. Contrast between the best and worst ensemble members

To evaluate the individual skills of all members, we rank all 100 members according to difference metrics. The first metric in measuring model skill is trend bias \( b(j) \) defined as the mean difference between each individual member and observation over the WHs:

\[
b(j) = \bar{TRm(j)} - TRo, \quad j = 0-99 \text{ members.} \tag{1}
\]

Here TRo and TRm are, respectively, the observed and simulated temperature trends during the 1951–2000 or 1901–2000 period. The overbar is the average over the WH regions defined in Fig. 2. Since almost all members (99 out of 100; see below) underestimate the observed cooling rate (99 members simulated less cooling than the observed or warming in the southeast WH), we will simply use the trend value itself, denoted TR, in lieu of bias hereafter. Doing so can avoid the need to distinguish between positive and negative biases that complicates the ranking and allow for the direct conclusion: the smaller the TR value is, the smaller the bias is, except for the one member that overpredicts the cooling rate in the southeast WH region (Table 2, first paired columns). Therefore we will use the terms bias \( b \) and cooling rate TR in the WH regions interchangeably.

The second metric is the temporal correlation \( r \) between the simulated and observed WH temperature anomaly:
Here, $N = 50$ or 100 yr; $T_o$ and $T_m$ are the observed and simulated temperature anomaly respectively. The quantities $S_{T_o}$ and $S_{T_m}$ are sample standard deviation of observed and simulated temperature anomaly respectively. The third parameter is the pattern or spatial correlation ($r$) of temperature trends between simulations and observation:

$$r(j) = \frac{1}{N-1} \sum_{n=1}^{N} \left( \frac{T_{o_n} - T_o}{S_{T_o}} \right) \left[ \frac{T_{m_n}(j) - T_{m}(j)}{S_{T_m}(j)} \right],$$

$j = 0–99$. (2)

Here, $T_{o}$ and $T_{m}$ are the observed and modeled temperature trends, respectively; $\text{cov}[]$ is the covariance. The quantity $\rho(j)$ is computed over a larger land region ($30^\circ–50^\circ$N, $110^\circ–75^\circ$W) in order to reflect spatial pattern around the cooling region. Since the cooling in the SE WH is more persistent than the CN WH over the whole twentieth century (Fig. 1, top panels) and cooling is more pronounced for $T_{max}$ in summer, we will focus our attention on $T_{max}$ over the SE WH during 1951–2000 summer when cooling is most evident.

Figure 8 contrasts model skills between the top and bottom quartiles of 100 members based on the three metrics defined in (1)–(3). As expected, the best-bias ($b$ or TR) members collectively simulated a well-defined cooling region in the south-central United States, matching observed WHs quite well (Fig. 8a). The worst-TR 25 members, however, simulated a clear warming in the region. While the sharp parity in bias performance between the two quartiles is somewhat expected, this demonstrates that quite a portion of members can reproduce the WH phenomenon, which at least allows for the opportunity to diagnose what caused the two quartiles to differ. The best temporal correlation ($r$) members did not reproduce the resemblance of the observed WH (Fig. 8b), indicating the secondary role of $r$ in simulating the WH. On the other hand, the best spatial correlation ($\rho$) members produced well the southeast WH relative cooling (i.e., lack of warming, Fig. 8c), suggesting that capturing the correct $\rho$ is the key for the WH simulation.

It could be argued that it should be apparent that $\rho$ is more important than $r$ in determining the WH pattern, so we also compared its relative role in determining the temporal variations of WH temperature. Figure 9 contrasts the best and worst model performance measured by $r$. The shaded area represents the spread (maximum minus minimum) of the top 25 best-$r$ members and the eight lines are the best eight members. [We chose eight best (worst) members mainly for plot clarity.] The members captured the temporal variation of the WH $T_{max}$ very well, especially during the latter half of the twentieth century, as indicated by both the shading envelope and eight-member means (thick solid red line in Fig. 9a). On the other hand, the 25 worst-$r$ members simulated decadal variations that are largely out of
phase with the observations (Fig. 9b). The eight best-TR members simulated mainly smoothed decadal variations with a gentle cooling during the second half of the twentieth century (Fig. 9c). Finally, the eight best-member captured the decadal trend and sharp cooling in the 1950–60s reasonably well (Fig. 9d).

Table 2 ranks the top quartile members sorted by various metrics. The first three paired columns are sorted by the SE WH Tmax trend bias measured by the trend (TR; °C decade \(^{-1}\)), U.S. Tmax pattern correlation (\(\rho\)), and the WH Tmax temporal correlation (\(r\)), respectively. The last two paired columns are ranked by temporal correlations of observed and modeled temperature over the PDO (20°–70°N, 110°E–80°W) and AMO (0°–60°N, 60°–10°W) regions, \(r_{\text{PDO}}\) and \(r_{\text{AMO}}\), respectively. The following can be seen from Table 2:

1) Among the first three paired columns (TR, \(\rho\), and \(r\)) that directly measure temperature simulation skill in or around the WH region, only four members (0, 4, 29, and 88) made all the three top-25 lists and the orders of the first three columns bear little resemblance, implying that the ensemble members hardly get all three metrics in a consistent manner. They may simulate one metric well, but for incorrect reasons. For example, correct bias or TR (i.e., cooling trend) might be caused by a model’s cold bias globally.

2) The temporal correlation varies from 0.82 to 0.53 among the 25 members, higher than the spatial correlation (0.59–0.24), suggesting that the models captured the temporal variation better.

3) Model members simulated temperature PDO (AMO) regions quite well with temporal correlation of 0.87–0.67 (0.82–0.67). However, member ID rank lists of PDO and AMO temperature bear no resemblance to the other three columns, indicative of disconnection between WH temperature and PDO/AMO temperature in these coupled CMIP5 models. This is in disagreement with some uncoupled modeling studies that show strong association between WH temperatures and the PDO index (Robinson et al. 2002; Wang et al. 2009).

To further partition error characteristics, we quantify the top and bottom quartile members’ performance collectively. Table 3 lists the mean of the three metrics
FIG. 8. Mean linear trends of Tmax (°C) among the (left) best and (right) worst quartile members during 1951–2000 period sorted by (a) bias $b$, (b) temporal correction $r$, and (c) spatial correlation $\rho$ as defined in (1)–(3).
averaged over the best and worst 25 members of various sorting categories. The lists also include temperatures over the PDO and AMO regions ranking. Each row represents a sorting according to a parameter and columns are the average of the top and bottom quartile members. For example, sorted by TR (first two rows), the average spatial correlation of the top and bottom 25 models are +0.19 and −0.13, respectively, while the temporal correlation values just differ slightly (0.42 vs 0.31), indicating that the separation of the best and worst members measured by trend bias is more caused by spatial than temporal correlations. Similarly, if sorted by the spatial correlation (third and fourth rows), the bias difference between the top and bottom quintiles is still large, but the temporal correlation has no difference. Finally, if sorted by temporal correlation, the separation in bias is smaller while the spatial correlation is essentially the same. This leads us to believe that 1) the spatial and temporal correlations do not collaborate well in the models and 2) temporal variation, compared to spatial, is less important in contributing to the model bias. To further substantiate this, we also computed temperature over the PDO and AMO regions and their temporal correlation with observed ones. The PDO–AMO sorting does separate the spatial correlation somewhat, but they reverse parity of the temporal correlations ($r$).

TABLE 3. Contrast of various metrics between means of the top and bottom quartile members in the first and second row for a given entry, respectively. The ranking and thus quartiles are sorted by individual metrics; TR is WH Tmax trend, $\rho$ is U.S. Tmax spatial correlation, and $r$ is WH Tmax temporal correlation; $r_{\text{PDO}}$ and $r_{\text{AMO}}$ are temporal correlations of PDO and AMO temperatures, respectively. See text for details.

<table>
<thead>
<tr>
<th></th>
<th>TR</th>
<th>$\rho$</th>
<th>$r$</th>
<th>$r_{\text{PDO}}$</th>
<th>$r_{\text{AMO}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR</td>
<td>−0.06</td>
<td>0.19</td>
<td>0.42</td>
<td>0.32</td>
<td>0.53</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.07</td>
<td>0.38</td>
<td>0.37</td>
<td>0.42</td>
<td>0.51</td>
</tr>
<tr>
<td>$r$</td>
<td>0.06</td>
<td>0.03</td>
<td>0.62</td>
<td>0.26</td>
<td>0.57</td>
</tr>
<tr>
<td>$r_{\text{PDO}}$</td>
<td>0.13</td>
<td>0.04</td>
<td>0.01</td>
<td>0.55</td>
<td>0.52</td>
</tr>
<tr>
<td>$r_{\text{AMO}}$</td>
<td>0.10</td>
<td>0.01</td>
<td>0.28</td>
<td>0.78</td>
<td>0.56</td>
</tr>
</tbody>
</table>

FIG. 9. Time series of SE WH maximum surface temperature during summer. (a) Best and (b) worst members measured by temporal correlation. Eight best members based on (c) bias and (d) on spatial correlation. Shading in (a) and (b) indicates the spread of the best and worst 25 members, respectively. The model mean is the average over the eight members.
d. Intramodel variability—Internal dynamics

This subsection will compare results within individual models that have multiple member realizations. Since the model’s integration physics (and numerics) and external forcing remain the same, different initial conditions only represent internal variability or model noise. In the CMIP5 experimental design, individual members are named rN MiMpL where the triad (N, M, and L) denotes ways in which each initial condition is formed: N denotes different starting times from the same realistic time series; M, the initializing method; and L, the way of physical perturbation. All models but two have only varying N (i.e., changing only in starting times). GISS-E2-H and GISS-E2-R runs have members with two ways of varying initial conditions (N and L). Figure 10 shows the Tmax trends during 1951–2000 simulated by all 15 ensemble members in GISS-E2-H. The three panels in a given row (e.g., r1i1p1-p3 on the top row) represent same initial time and initializing method, but with three different physical perturbations. Similarly, the five rows represent different initial times. The 15 members vary significantly, but some patterns are still identifiable. The physical perturbation method has larger impacts than the starting time.\(^2\) The middle column (L = 2) tends to have sharp cooling comparable to the observed WH extent, but located too far west as compared with the observations. On the other hand, in the right column strong warming occurred over the observed cooling region. The box-and-whisker plots demonstrate that the intramodel variability of GISS-E2-H (along with the other five models that have the largest ensemble members) is similar to that of the 100 members, but with less spread (not shown).

e. External forcing

Globally, it is very likely that the climatic warming observed over the past decades is attributable to human influences, primarily to an increase in concentrations of well-mixed greenhouse gases (Meehl et al. 2004; Solomon et al. 2007). The anthropogenic signal was detected in each of 14 regions of the globe except for one in central North America, although the results were more uncertain when anthropogenic and natural signals were considered together (Taylor et al. 2012). To attribute observed climate change to particular causes, it is essential to perform simulations of the historical period with only a subset of known forcing.

Like in CMIP3, CMIP5 designed the so-called attribution and detection experiments consisting of historicalNat and historicalGHG, among others. The historical (all forcing) runs presented so far impose changing conditions (consistent with observations), which include atmospheric composition (including CO2), due to both anthropogenic and volcanic influences, solar forcing, emissions or concentrations of short-lived species and natural and anthropogenic aerosols or their precursors, and land use change (Taylor et al. 2012). The natural forcing only experiment imposes natural variations (e.g., volcanoes and solar variability) evolving as in the historical run. Correspondingly, the GHG forcing only experiment includes greenhouse gas forcing alone evolving as in the historical run.

Whether the abnormal cooling in the central and southeastern United States is caused by internal variability of the climate system or forced by external forcing is a long-standing issue (Wang et al. 2009; Meehl et al. 2012). If the cooling is transient internal, the WH regions would become warmer when the transient masking mechanism is gone in the future and the WH regions will “catch up” the missed warming (Kunkel et al. 2006). If, on the other hand, they are a response to the global warming forced by an external forcing, they will likely continue to exist in the future. Several studies suggest the WHs are related to the PDO and AMO indices, an internal variation of the atmosphere–ocean coupled system (Kunkel et al. 2006; Wang et al. 2009; Meehl et al. 2012). Others suggest that land surface processes and regional hydrological processes contribute to the cooling (Kalnay and Cai, 2003; Pan et al. 2004; Liang et al. 2006). Here we analyze the CMIP5 attribution experiments that include historicalNat and historicalGHG experiments.

Fewer models carried out these attribution experiments with fewer ensemble members compared with the historical and RCP4.5 runs. We evaluated six models with a single member (r1i1p1): CCSM4, GFDL-ESM2M, GISS-E2-H, GISS-E2-R, MRI-CGCM3, and NorESM1-M. Figures 11a and 11b show that on the century scale, natural forcing has a cooling effect only in the central and northern United States in summer. The position matches quite well the observed CN WH. In the second half of the twentieth century, the northern tier of the United States cooled considerably. Conversely, GHG forcing only would make the central United States warmer, particularly during the latter half of the century in summer (Figs. 11c,d). This suggests that GHGs would counteract the WH, rather than causing or enhancing it. The forcing difference between the historical and historicalNat should reflect land use evolution, among other factors such as aerosols. Interestingly, the difference showed a clear WH feature, especially in summer (Figs. 11e,f). On the century scale, a large area of 0°–0.05°C decade\(^{-1}\) cooling over the southeastern-central

---

\(^2\)The extent of the member spread may have been underestimated in CMIP5 experiments since the great majority of the models used varying start time only, not the perturbation method.
FIG. 10. Linear trends of summer Tmax (°C decade⁻¹) simulated by the 15 individual members of the GISS-E2-H models listed during 1951–2000. The triad of integers in ensemble member (each panel) name rNiMpL (N, M, and L) denote initial time, initiation method, and perturbation physics, respectively.
FIG. 11. Six-model ensemble mean of linear trends of summer Tmean (°C decade⁻¹) during (left) 1901–2000 and (right) 1951–2000 simulated in the (a),(b) historicalNat and (c),(d) historicalGHG experiments; and (e),(f) the difference between the historical and historicalNat experiments.
United States resembles the observed central WH very well. On the 50-yr scale, the cooling extent retreated to the southeast coast. The larger cooling difference between the two experiments in the whole twentieth century, rather than the latter half-century, perhaps reflects that major land use change in the region occurred during the earlier decades of the twentieth century (Kumar et al. 2013b). The larger impacts in summer rather than winter (not show) may be due to the larger roles that land surface processes play in the warm season due to greater evapotranspiration than in winter. If this is the case, the summer WH in the central and southeastern United States in the twentieth century is at least influenced by local/regional land surface processes, consistent with previous studies (Kalnay and Cai 2003; Pan et al. 2004; Liang et al. 2006), disregarding forcing from other factors such as aerosols (Leibensperger et al. 2011)

Figure 12 compares trends under different scenarios over different periods and seasons. The GHG forcing only has strong warming effects (0.12°–0.22°C decade⁻¹) that may have partly compensated for cooling effects from the natural forcing in the all-forcing historical experiment. The historical experiment that incorporates both natural and anthropogenic forcing resulted in moderate warming as seen in the historical experiment.

5. Fate of warming holes in the twenty-first century as simulated in RCP4.5

This section discusses the RCP4.5 simulations from 22 models with 63 ensemble members. These 22 models include all the historical models except for FGOALS-S2.0, GFDL-CM3, GFDL-ESM2G, GISS-E2-H, HadCM3, and MPI-ESM-P. The list also includes BNU-ESM, INM-CM4, and IPSL-CM5B-LR that did not appear in historical run, each of which has one member (Volodin et al. 2010). One-third (8) of the models had multiple members, ranging from 3 to 15 (Table 1). For simplicity, in this section the analysis is based on individual models (not members), meaning that each model contributes equally to the ensemble mean. Figure 13 shows the model ensemble mean of the projected Tmean over the first half of the twenty-first century (2006–55). During summer, the northern section of the United States warmed more than both lower and higher latitudes. The largest warming is located in the Great Lakes across to the mountain region where Tmean would warm by 0.4–0.5°C decade⁻¹. In winter, the strongest warming is over the northern United States all the way to the north with a magnitude of more than 0.5°C decade⁻¹. This warming pattern both in winter and summer is very similar to the simulated trend distributions for 1976–2000 (Fig. 4), the peak global warming period of the twentieth century.

If we view intermodel variance as a measure of the uncertainty in the projection, generally the areas of large uncertainty tend to coincide with large trends themselves. Perhaps the ratio of trend to intermodel variance would be a better measure of model uncertainty. The highest ratios (or confidence) are over high latitudes in winter with large trends and slightly large spread (right panel) and the lowest confidences in summer over the U.S.–Mexico border are likely related to the complex topography in the region.
The warming is faster during the first half of the twenty-first century and then slows down after about the 2050s (not shown), consistent with the leveling off of the atmospheric CO₂ concentration under the RCP4.5 scenario. The diminishing warming and even cooling during the latter periods of the twenty-first century suggest the likely return of the WH, considering the models’ underestimation of WHs in the twentieth century in the historical simulation. However, given the large intermodel variability and models’ limited capability in reproducing the observed WH in the twentieth-century simulations, cautions are warranted in interpreting the fate of WH in the twenty-first century.

During the first 50 years of the twenty-first century, the model ensemble mean showed a 0.3°C–0.4°C decade⁻¹ warming over SE region, ranging from −0.01°C to 0.9°C decade⁻¹ (Fig. 14); Tmax and Tmin warm at a similar rate, which differs from the twentieth-century simulations where Tmax rose slower than Tmin (Fig. 6). In fact, the summer Tmax rises faster than Tmin, a phenomenon rarely observed.

6. Conclusions and discussion

A total of 175 ensemble members of long-term simulations from 28 AOGCMs available in the historical suite and RCP4.5 experiments are analyzed to examine the models’ skill in reproducing the twentieth-century observed U.S. temperature trends. The focus is on the southeast and central regions where abnormal cooling occurred despite the fact that global warming accelerated during the twentieth century. To the author’s knowledge, this work is one of the studies that analyzed the largest number of AOGCM members to evaluate the collective skills of the climate models. Unlike most climate studies with this kind of model evaluation, we evaluated maximum and minimum surface air temperatures (Tmax and Tmin) separately rather than the mean of the two, although observational studies often separate the two temperatures. Surface temperature is affected by many factors, some of which affect Tmax and Tmin more symmetrically than others (Durre and Wallace 2001; Zhou et al. 2007). For example, longwave radiation tends to affect equally Tmax and Tmin, whereas evapotranspiration and sensible heat flux mainly affect Tmax (Dai et al. 1999; Zhou et al. 2007). With the separation, more detailed physical and dynamical processes can be revealed. We did not examine these forcing contrasts between Tmax and Tmin because of the unavailability of those data associated with these processes.

We evaluated model skills in three periods, 1901–2000, 1951–2000, and 1976–2000, corresponding to the whole twentieth century, the data-rich period, and the peak global warming period, respectively. The 100-ensemble-member mean in the historical experiment hardly shows regional cooling in the central and southeast United States for any of the three periods. Focusing on Tmax over the southeast United States in summer during 1951–2000, 19 out of 100 members did simulate a negative trend (cooling), with 99 members under-predicting the cooling amplitude. If measured by one of the three metrics (bias, spatial correlation, and temporal correlation), at least the top quartile of the 100 ensemble members were able to capture reasonably well the
anomalous cooling trends in the southeastern United States. For example, the top quartile based on bias sorting or spatial correlation sorting can reproduce a warming hole–like feature quite well. The two rank lists, however, consist of almost two entirely different sets of member sets. In fact, there are only four members fall into all the three top-quartile lists sorted by bias, spatial pattern correlation, and temporal correlation, respectively. This seems to imply that a model’s high skill in one metric is not achieved in a consistent manner. For example, a smaller bias in cooling is not due to the model’s better performance in capturing spatial pattern or temporal variation. In other words, a single correct metric could be achieved for wrong reasons. For instance, one member simulates cooling rate better than others in the warming hole region might be due to its global cooling bias (less warming) rather than to its correct simulation of regional temporal correlation or spatial pattern. Thus it would be preferable to have multiple metrics in measuring model performance even though one deals with a cooling trend in a specific region such as the warming hole (WH) area.

The model members had more difficulty in producing the spatial trend pattern than (decadal) temporal variations, as indicated by larger parity in the former among the top and bottom quartile members measured by bias and by higher temporal correlation than spatial one. The model members simulated decadal variation represented by temperature in PDO and AMO regions quite well with the best temporal correlation higher than 0.87 despite the fact that only a small portion of members (19/100) simulated trend sign correctly in the southeast region. This disconnection between the WH region temperature and PDO–AMO region temperature in the coupled AOGCMs differs from some uncoupled modeling studies (Robinson et al. 2002; Wang et al. 2009) that showed strong correlation between the U.S. warming hole and PDO index. Further studies are needed to reconcile the difference, particularly for individual models in these studies.

Whether the abnormal cooling is due to the atmospheric internal variability or external forcing is the focus of a number of studies (Robinson et al. 2002; Wang et al. 2009; Meehl et al. 2012). The historical suite experiments in CMIP5 consisting of all forcing (historical), GHG forcing only (historicalGHG), and natural forcing only (historicalNat) runs provide an opportunity to look into this issue. The GHG forcing has a warming effect in the central United States, implying that the WH is not due to the GHG forcing. The difference between the all forcing and natural forcing only runs showed a well-defined cooling region resembling the WH location, implying that local and regional surface processes may contribute to the WH. The fate of central and southeast WHs would likely depend on the relative magnitudes of GHG forcing that contributes to warming and the natural forcing that contributes to cooling. If the GHG forcing is strong enough, the WHs will likely disappear in the future.

Acknowledgments. We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI), and the WCRP’s Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP5 multimodel dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy. The U.S. authors acknowledge the support of NOAA Climate Program Office “Modeling, Analysis, Predictions and Projections” (MAPP Grant NA11OAR4310094) Program as part of the CMIP5 Task Force. This research is also supported partly by the CAS/SAFEA International Partnership Program for Creative Research Teams (KZZD-EW-TZ-03) and WKC foundation. We are thankful to two anonymous
reviewers for their constructive and insightful suggestions. We are thankful to the two anonymous reviewers whose constructive comments strengthened this paper.

REFERENCES


