© 2018 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

Do Downscaled General Circulation Models Reliably Simulate Historical Climatic Conditions?

Andrew R. Bock^a

Colorado Water Science Center, U.S. Geological Survey, Lakewood, Colorado

Lauren E. Hay, Gregory J. McCabe, and Steven L. Markstrom

Water Mission Area, U.S. Geological Survey, Lakewood, Colorado

R. Dwight Atkinson

Office of Wetlands, Oceans, and Watersheds, U.S. Environmental Protection Agency, Washington, DC

Received 10 August 2017; in final form 23 January 2018

ABSTRACT: The accuracy of statistically downscaled (SD) general circulation model (GCM) simulations of monthly surface climate for historical conditions (1950–2005) was assessed for the conterminous United States (CONUS). The SD monthly precipitation (PPT) and temperature (TAVE) from 95 GCMs from phases 3 and 5 of the Coupled Model Intercomparison Project (CMIP3 and CMIP5) were used as inputs to a monthly water balance model (MWBm). Distributions of MWBM input (PPT and TAVE) and output [runoff (RUN)] variables derived from gridded station data (GSD) and historical SD climate were compared using the Kolmogorov–Smirnov (KS) test. For all three

^aCorresponding author: Andrew R. Bock, abock@usgs.gov

variables considered, the KS test results showed that variables simulated using CMIP5 generally are more reliable than those derived from CMIP3, likely due to improvements in PPT simulations. At most locations across the CONUS, the largest differences between GSD and SD PPT and RUN occurred in the lowest part of the distributions (i.e., low-flow RUN and low-magnitude PPT). Results indicate that for the majority of the CONUS, there are downscaled GCMs that can reliably simulate historical climatic conditions. But, in some geographic locations, none of the SD GCMs replicated historical conditions for two of the three variables (PPT and RUN) based on the KS test, with a significance level of 0.05. In these locations, improved GCM simulations of PPT are needed to more reliably estimate components of the hydrologic cycle. Simple metrics and statistical tests, such as those described here, can provide an initial set of criteria to help simplify GCM selection.

KEYWORDS: General circulation models; Hydrologic models; Precipitation; Runoff; Temperature

1. Introduction

In response to the SECURE Water Act of 2009 (House of Representatives 2009), the WaterSMART program (<http://water.usgs.gov/watercensus/WaterSMART.html>) was started by the U.S. Department of the Interior in February 2010. Under WaterSMART, the National Water Census was proposed as one of the U.S. Geological Survey's (USGS) key research directions, with a focus on developing new hydrologic tools and assessments. In October 2015, the USGS Water Availability and Use Science Program (<http://water.usgs.gov/wausp/>) was established. One of the major components of the program is to provide national estimates of water availability at a subwatershed resolution with the goal of determining 1) if the nation has enough fresh water to meet both human and ecological needs and 2) whether this water will be available to meet future needs.

Toward satisfying the first goal, the quantity of fresh water was simulated for the conterminous United States (CONUS) using a monthly water balance model (MWBM) (Bock et al. 2016a,b, 2017). To satisfy the second goal, estimates of future freshwater supplies are needed. This can be accomplished by using projected climate simulations from general circulation models (GCMs) to drive hydrologic models. Many atmospheric processes that have hydrologic consequences are not modeled adequately by GCMs (Liu et al. 2014; Papadimitriou et al. 2017); reliable hydrologic modeling requires climatological information on scales that are generally much finer than the typical grid size of even the highest-resolution GCMs (Hay et al. 2002). To study the hydrologic effects of climate change at the spatial scales required to estimate the amount of future water supplies, finer-resolution (downscaled) climate projections are required.

Wood et al. (2004) noted that reproducing accurate historical conditions is the minimum standard for any hydrologic application of downscaled climate simulations (Tebaldi and Knutti 2007). Hay et al. (2014) used the two-sample Kolmogorov–Smirnov (KS) test (Conover 1971) as a criterion to evaluate the accuracy of statistically downscaled (SD) GCM simulations as inputs for hydrologic and stream temperature simulations on a daily time step in the Apalachicola–Chattahoochee–Flint River basin, located in the southeastern CONUS. The KS test is a nonparametric test that determines if two samples or populations are from the same

distribution, but it does not provide an indication of negative or positive model bias and does not indicate the relative accuracy of extremes (i.e., droughts/floods, largest/smallest events, or sustained events). Results from Hay et al. (2014) indicate that many of the hydrologic model outputs simulated using downscaled GCM climate data may not be reliable for studies requiring analyses on daily or even weekly time scales. They concluded that until improved GCM simulations of daily precipitation are made available, estimates of future streamflow may be most appropriately evaluated on a weekly, or longer, time step.

This paper uses the KS test as a criterion for identifying SD GCMs that poorly replicate historical climatic conditions. For this study, the distributions of historical gridded station data (GSD) and GCM-simulated monthly precipitation (PPT) and temperature (TAVE) are compared for the 1950–2005 period for the CONUS. To provide more detail on the distributions of the variables, several comparative metrics were used to describe where in the distributions (i.e., tails or the middle) the largest differences in magnitude between the distributions of GSD and SD variables occur, and the direction of bias at this position.

2. Methods

The methods used in the study are described below in four sections, which correspond to Figure 1. The first two sections describe the climate data and hydrologic model used in the study (“Climate data” and “Hydrologic model” in Figure 1). Climate forcings (PPT and TAVE) from GSD and SD GCMs were summarized by hydrologic response unit (HRU; described in section 2.2) across the CONUS (Bock et al. 2016a, 2017). The MWBM, a monthly time step rainfall–runoff model (Figure 2), was used to simulate monthly runoff (RUN) at the HRU scale for historical conditions (1950–2005).

The last two sections describe the statistical test and climate metrics used to evaluate the SD simulations (“Kolmogorov–Smirnov test” and “Distribution metrics” in Figure 1). The KS test was used to evaluate if the SD PPT, TAVE, and RUN replicated historical climatic conditions when compared to the variables derived from GSD simulations. Two additional metrics, D_X and D_B (described later), were calculated to further describe more detailed characteristics of the distributions of GSD and SD variables.

2.1. Climate data

Climate forcings (PPT and TAVE) for CONUS-extent gridded datasets (Maurer et al. 2002; Bureau of Reclamation 2011, 2013) were accessed through the USGS Geo Data Portal (<http://cida.usgs.gov/climate/gdp/>) (Blodgett 2013). The Geo Data Portal is an interactive online tool that utilizes a variety of processing algorithms to summarize spatial gridded datasets for customized features and areas of interest. The processing algorithms available on the portal were developed in part to enable efficient processing of gridded environmental time series, such as climate datasets, into machine-readable, model-ready formats. The area grid statistics (weighted) algorithm was used to summarize grids of PPT and TAVE from the GSD and SD into a single time series for each of the 109 951 HRUs across the CONUS. Details

Figure 1. Schematic flowchart of the methods used in the study.

Figure 2. Conceptual diagram of the MWBM (McCabe and Markstrom 2007).

of the HRUs used in this analysis area are described in more detail in [section 2.2](#) below.

2.1.1. Historical measured climate

The GSD produced by [Maurer et al. \(2002\)](#) was chosen to represent historical measured climatic conditions. The GSD was used as the training dataset for the downscaling procedure applied to phases 3 and 5 of the Coupled Model Inter-comparison Project (CMIP3 and CMIP5) ensembles used in this paper (described below; [Bureau of Reclamation 2011, 2013](#)) and has been used as the training dataset for other CONUS-scale downscaling efforts ([Stoner et al. 2013](#)). Daily PPT (mm) and minimum and maximum TAVE data ($^{\circ}\text{C}$) for the CONUS at a $1/8^{\circ}$ cell size (approximately 140 km^2) were summarized to HRUs using the Geo Data Portal (described in the previous paragraph) for calendar years 1950–2005. The data then were aggregated to a monthly time step for input to the MWBM, with daily time step PPT summarized for each month and daily time step maximum and minimum TAVE averaged for each day, and then averaged for each month.

2.1.2. Statistically downscaled GCM climate

The SD simulations from 95 GCMs were used in this analysis (38 from CMIP3 and 57 from CMIP5). These 95 SD GCM simulations were previously statistically downscaled using the bias-corrected, spatially disaggregated methodology ([Wood et al. 2004](#); [Maurer et al. 2007](#)) to $1/8^{\circ}$ grids across the CONUS ([Bureau of](#)

Reclamation 2011, 2013). Monthly PPT and TAVE from the 38 CMIP3 simulations and the 57 CMIP5 simulations were summarized to the HRUs using the Geo Data Portal. Because the GSD was used as the downscaling training dataset, it was considered as “truth” for this paper. For this study, the retrospective downscaling for the years 1950–2005 from the SD GCM simulations were compared with historical climatic conditions (GSD) for the CONUS. The SD climate simulations used in this analysis, including all ensemble members and their associated research institutions, are listed in appendixes 1 and 2 of Bock et al. (2017).

2.2. Hydrologic model

The MWBM (Figure 2) is a modular accounting system that provides monthly estimates of components of the hydrologic cycle (Wolock and McCabe 1999; McCabe and Markstrom 2007; McCabe and Wolock 2011a; Bock et al. 2016b). Monthly TAVE is used to compute potential evapotranspiration (PET) and to partition monthly PPT into rain and snow (Figure 2). PET was computed using a modified version of the Hamon equation (Hamon 1961). The Hamon equation incorporates a dimensionless coefficient in the computation of PET. In the CONUS-wide application of the MWBM, this coefficient was calibrated by matching PET estimates to mean monthly free-water surface evaporation (Farnsworth et al. 1982), creating spatially and mean monthly varying coefficients by HRU across the CONUS (McCabe et al. 2015). The PPT that occurs as snow is accumulated in a snowpack as snow storage; rainfall is used to compute direct runoff, actual evapotranspiration (AET), soil moisture storage, and surplus water, which eventually becomes runoff (Figure 2). When rainfall for a month is less than PET, AET is equal to the sum of rainfall, snowmelt, and the amount of moisture that can be removed from the soil. The fraction of soil moisture storage that can be removed as AET decreases linearly with decreasing soil moisture storage; that is, water becomes more difficult to remove from the soil as the soil becomes drier and less moisture is available for AET. When rainfall (and snowmelt) exceeds PET in a given month, AET is equal to PET; water in excess of PET replenishes soil moisture storage. When soil moisture storage reaches capacity during a given month, the excess water becomes surplus. A fraction of the accumulative surplus becomes runoff, while the remainder of the accumulative surplus is temporarily held in storage. The MWBM has been previously used to examine variability in runoff at both the CONUS extent (Wolock and McCabe 1999; Hay and McCabe 2002; McCabe and Wolock 2011a; Bock et al. 2016b) and the global extent (McCabe and Wolock 2011b).

The MWBM parameterization used in this study is described in detail in Bock et al. (2016b). Bock et al. (2016b) configured the MWBM for the CONUS to run for 109 951 HRUs from the Geospatial Fabric for National Hydrologic Modeling (Viger and Bock 2014), a national database of hydrologic features for national hydrologic modeling applications (Figure 3). This HRU derivation was based on an aggregation of the National Hydrography Dataset Plus dataset, version one (<http://www.horizon-systems.com/nhdplus/>) (U.S. EPA and USGS 2010), an integrated suite of geospatial data that incorporates features from the National Hydrography Dataset (<http://nhd.usgs.gov/>), the National Elevation Dataset (<http://ned.usgs.gov/>),

Figure 3. HRUs of the Geospatial Fabric for National Hydrologic Modeling overlaid by Hydrologic Region boundaries.

and the Watershed Boundary Dataset (<http://nhd.usgs.gov/wbd.html>). The median size of the HRUs is approximately 33 km².

Inputs to the MWBM by HRU are time series of monthly PPT (mm) and monthly mean TAVE (°C) (derived in this study using the Geo Data Portal), latitude of the site [decimal degrees (°)], soil moisture storage capacity (mm), and a monthly varying coefficient (dimensionless) for the computation of PET (i.e., Hamon PET) (McCabe et al. 2015). Latitude for each HRU was computed as the centroid of each HRU. Soil moisture storage capacity was calculated for each HRU using State Soil Geographic Database (STATSGO) data on a 1-km²-grid resolution, assuming a 1-m rooting depth (Wolock 1997). The MWBM was calibrated across the CONUS using measured streamflow from 1575 streamgages and modeled snow water equivalent (Bock et al. 2016b), and the calibrated parameters were regionalized for all 109 951 HRUs. On average, the model performed adequately across the CONUS (section 4.2 in Bock et al. 2016b). The largest simulated streamflow residuals occurred in regions with 1) data scarcity (sparser networks of streamgages), 2) greater spatial and temporal variability of precipitation, and 3) runoff processes that occur at finer time steps than monthly, such as the arid intermontane, desert, and steppe regions of the western CONUS (section 5.3 and Figures 9, 11–14 in Bock et al. 2016b).

2.3. Kolmogorov–Smirnov (KS) test

The goal of this study was to evaluate the ability of the SD climate simulations to reproduce 1) historical TAVE and PPT and 2) corresponding MWBM RUN output. The KS test (Figure 4) was used to determine if PPT and TAVE from the 95 SD

Figure 4. Setup of the two datasets GSD and SD for the KS test and the two additional metrics used to evaluate the SD datasets— D , the KS test statistic, the greatest vertical difference between the distributions of the GSD variable and the SD variable; D_X , the x-axis value along the distribution (x axis) at which the KS test statistic occurs (D); and D_B , the direction of bias between the distributions of the GSD variable and SD variable at the D_X .

datasets replicated historical PPT and TAVE represented by GSD. The KS test also was used to evaluate if SD PPT and TAVE can be used with the MWBM to reliably simulate RUN for historical climatic conditions across the CONUS (compared to RUN, computed using GSD PPT and TAVE as inputs to the MWBM).

The nonparametric two-sample KS test from the R Core packages (Conover 1971; R Core Team 2015) was used to determine if variables (PPT, TAVE, and RUN) from the GSD and SD simulations had different distributions for historical conditions (1950–2005). The KS test can be used to test the agreement between two cumulative distributions and was developed by Smirnov (1939), based upon previous work by Kolmogorov (1933, 1941). The KS test finds the maximum distance between two empirical cumulative distribution functions (the KS test statistic, or D in Figure 4) and is sensitive to differences in both central tendency and distribution shape. It has the disadvantage of being more sensitive to deviations near the center of the distribution than at the tails. The null hypothesis of the KS test (H_0 ; the two datasets are from the same population) is rejected if the KS test p value calculated for a pair of samples (the distribution of GSD and SD variables) is below a specified significance level or alpha (α) selected. If the calculated p value is greater than the specified significance level, the KS test does not reject the null hypothesis, and the datasets are from the same population.

In this study, cumulative distribution functions for SD PPT, TAVE, and RUN at each HRU across the CONUS are compared with cumulative distribution functions for GSD PPT, TAVE, and RUN using the KS test (Figure 4). If the null hypothesis is rejected (when the KS test p value is smaller than the significance level selected),

then the two populations tested may differ in median, variability, and the shape of the distribution. If the null hypothesis is not rejected, then the SD variable has a similar distribution to the GSD-based variable for current climatic conditions and has “passed” the KS test. With this configuration, the KS test can be used for subsetting SD simulations to those that replicate historical conditions by excluding SD simulations for which the null hypothesis is rejected. In this paper, the KS test is applied with the significance value of 0.05, which was also used with the KS test significance level in Hay et al. (2014). To supplement the binary pass/fail functionality of using a single significance level for the KS test, the magnitude and variability of p values across the CONUS and the performance of the significance test over a range of significance values are summarized.

2.4. Distribution metrics

Two additional metrics were computed to provide more information on the maximum difference and direction of the bias between the distributions of SD and GSD variables, D_X , and D_B (Figure 4). The D_X is the x -axis value (variable distribution) at which the KS test statistic occurs (D ; Figure 4). The D_X was converted to a percentile of the GSD range with the following formula [Equation (1)]:

$$D_X = \text{SD}_X - \text{GSD}_{\min} / \text{abs}(\text{GSD}_{\max} - \text{GSD}_{\min}), \quad (1)$$

where SD_X is the SD position of D for an SD variable, and $\text{GSD}_{\min}$ and $\text{GSD}_{\max}$ are the minimum and maximum bounds of the range for the GSD variable. This normalization allows for the D_X of SD variables to be scaled on a 0 to 1 scale relative to the range of the GSD variable. For a single summary statistic across the CONUS, the median of all D_X values (computed for the 95 model simulations) for each of the three variables was calculated for each HRU.

The second metric, D_B (Figure 4), is the percent of 95 SD simulations for a variable that underestimate the GSD at the x -axis value (D_X), where the KS test statistic occurs (D ; Figure 4):

$$B = \begin{cases} 1 & \text{if SD-derived variable} < \text{GSD-derived variable at } D_X \\ 0 & \text{if SD-derived variable} > \text{GSD-derived variable at } D_X \end{cases},$$

$$D_B = \sum_{n=1}^{95} B / 95, \quad (2)$$

where D_B is the percentage of SD datasets where the SD variable underestimates the GSD variable at D_X . This metric summarizes the probability that the 95 SD GCMs underestimate the GSD data where the two distributions differ in the greatest amount.

3. Results

Results from the KS test are presented first using null hypothesis significance testing to map the percentage of SD GCMs for each variable that replicate historical

Figure 5. Percent of the 95 SD GCM simulations used in this study that replicate historical climatic conditions for (a)–(c) CMIP3 and (d)–(f) CMIP5 at HRUs across the CONUS using a significance level of 0.05. (a),(d) PPT; (b),(e) TAVE; (c),(f) RUN.

conditions by HRU with a selected significance level ($\alpha = 0.05$). The results are shown by variable for each of the two ensembles (CMIP3 and CMIP5; Figures 5–7). These maps are accompanied by maps showing the median and interquartile range (IQR) of the p values for each HRU for PPT and RUN by ensemble (Figures 8, 9). Classification matrices (Tables 1, 2) are used to show percentages of the variables that replicate historical conditions based on the significance level. Figures 10–12 show the results for D_X and D_B .

Figure 5 shows the percent of SD GCMs that replicate historical conditions from null hypothesis significance testing with a significance value of 0.05. The results are mapped for the CONUS by variable (PPT, TAVE, and RUN) for every HRU in

Figure 6. Percent of the 95 SD GCM simulations used in this study that replicate historical conditions for all three variables (PPT,TAVE, and RUN) at HRUs across the CONUS using a KS test significance level of 0.05 for (a) CMIP3 and (b) CMIP5.

the Geospatial Fabric (Figure 3). The results from the CMIP3 and CMIP5 SD variables are shown side by side for comparison in Figure 5. Rose-colored HRUs indicate that none of the SD simulations for the given variable passed the KS test (SD variable p value < 0.05), and cyan-colored HRUs indicate that 100% of the SD

Figure 7. Percent area of the CONUS at which percentages of the 95 SD GCM simulations used in this study replicate historical conditions for SD PPT, TAVE, and RUN for CMIP3 and CMIP5 using a KS test with significance level of 0.05.

Figure 8. The median of p values of the SD PPT, RUN from the 95 SD GCMs for all HRUs in the CONUS for (a),(c) CMIP3 and (b),(d) CMIP5.

simulations for the given variable passed the KS test (SD variable p value > 0.05). For the remaining HRUs, the percent of SD variables that passed the KS test is indicated using a green color ramp, with dark green indicating HRUs with high percentages of SD simulations that passed the KS test. Evident from Figure 5, for a specific HRU, GCM reliability can vary substantially for PPT and RUN.

For large areas of the western CONUS (especially California), southern Texas, and the Appalachians/Piedmont, no CMIP3 simulations replicate historical PPT conditions (rose HRUs in Figure 5a). The results for CMIP5-based PPT simulations show a higher number of SD GCMs that reliably replicate PPT, compared with the CMIP3 PPT simulations (Figure 5d). Sheffield et al. (2014) compared CMIP3 and CMIP5 GCM differences (not downscaled) and reported that the multimodel GCM ensemble mean performance did not improve substantially from CMIP3 to CMIP5 for PPT over North America. In this study, results indicate a clear improvement in the downscaled CMIP5 PPT simulations (Figure 5d) relative to CMIP3 PPT simulations (Figure 5a). The pattern of differences between the GSD and SD PPT for CMIP3 (e.g., striations in Figure 5a) is likely an artifact of the original GCM resolution and subsequent effects from downscaling. In Figures 5b and 5e, results for TAVE are nearly identical; most of the CMIP3 and CMIP5 simulations of TAVE passed the KS test (cyan HRUs in Figures 5b,e), indicating that the SD TAVE for almost all the HRUs (too few to be visible in Figures 5b,e) reliably replicates the median, variability, and shape of monthly TAVE for historical climatic conditions. The results of the KS tests for RUN in the western CONUS (Figures 5c,f) are similar to those for PPT (Figures 5a,d).

Figure 9. As in Figure 8, but for the IQR of p values.

It is interesting to note that there are some HRUs where no SD PPT simulation replicates historical PPT conditions (rose HRUs in Figures 5a,d), yet for these same HRUs, historical RUN is reliably replicated (Figures 5c,f). Tebaldi and Knutti (2007) note that errors in different components of a single model can cancel, potentially giving the “right answer for the wrong reason.” Figure 6 shows the percent of SD GCM simulations that replicate historical conditions for all three MWBM variables examined in the study (PPT, TAVE, and RUN) at HRUs across the CONUS using the KS test significance value of 0.05. Figure 7 summarizes the CMIP3 and CMIP5 spatial results in Figure 6 with a histogram, further emphasizing the greater accuracy of CMIP5, compared to CMIP3 simulations. Note that Figures 6 and 7 present an accounting of the independent test results of the KS test for the three variables, rather than results of the KS test for the simultaneous reliability of model simulations for PPT, TAVE, and RUN.

The median and IQR of the 95 p values produced from the KS test for each HRU (Figure 3) were calculated for the PPT and RUN. These results are shown in Figure 8 (median) and Figure 9 (IQR). KS test results for TAVE are excluded from this set of plots because of the high median p value (0.99) and small variability across the CONUS, compared to KS test results for PPT and RUN (Figure 5). In Figure 8, HRUs with high (low) median p values match locations of HRUs with high (low) percentages of SD GCMs that replicate historical conditions in Figures 5 and 6. In many parts of the CONUS, the IQR of RUN is wider than PPT and is noticeably higher in areas of the Midwest (between -90° and -100° longitude) and along the southeast coast (25° to 35° latitude, -75° to -90° longitude). The HRUs with low

Table 1. Classification matrix that shows the percentages of paired PPT and TAVE p values from the same SD GCM that meet conditions defined for each category based on the significance levels on the top and left-hand sides of the matrix.

		CMIP3 precipitation		CMIP5 precipitation	
		p value < 0.05	p value > 0.05	p value < 0.05	p value > 0.05
Temperature	p value < 0.05	0.01	0	< 0.01	0
	p value > 0.05	30.36	69.63	7.95	92.04

median p values and low IQRs generally coincide with the rose-colored locations in Figures 5 and 6, where SD simulated variables fail to replicate historical conditions for both PPT and RUN.

Results shown in Figures 5–9 raise the question of how dependent the reliability of RUN simulations is on the reliability of simulations of PPT and TAVE. The p values derived from the KS tests at all HRUs were used to derive a classification matrix (Table 1) to quantify the percentage of cases (pairs of SD PPT and TAVE p values from the same GCM at an HRU) that meet four different conditions based on a single significance level ($\alpha = 0.05$). For both model ensembles, the majority of cases fall within the bottom two categories (p values for PPT and TAVE > 0.05; Table 1). There is a large difference in percentages of cases where both variables replicate historical conditions (lower-right categories for both ensembles), which is 69.63% for CMIP3 and 92.04% for CMIP5. This difference between these two values is almost entirely due to the reduced reliability of CMIP3 PPT to replicate historical conditions, compared to CMIP5, which is highlighted in the earlier figures (Figures 5–9).

While Table 1 focuses on PPT and TAVE, Table 2 expands the analysis to include RUN as a variable, and shows the percentages across three significance levels. The categories presented in this table focus only on cases where TAVE and RUN both replicate historical conditions for a given significance level (top row, TAVE and RUN p values > each α), and on cases where all three variables replicate historical conditions (bottom row, p values from all three variables > each α). The results here demonstrate that there is a much larger percentage of cases from CMIP3 SD GCMs where RUN replicates historical conditions without PPT replicating historical conditions (26%, 18.8%, and 9.1% across the three significance levels), as compared to CMIP5 (5.7%, 4.15%, and 2.17%). This difference between CMIP3 and CMIP5 also holds across the three different significance levels shown in the table.

Table 2. The probability of cases across the CONUS where PPT (P), RUN, and TAVE (T) replicates (top) historical conditions for a given significance level ($\alpha = 0.10, 0.05$, and 0.01), and all PPT, TAVE, and RUN replicate historical conditions for several different significance levels.

		CMIP3 runoff			CMIP5 runoff		
		p value < 0.10	p value > 0.05	p value > 0.01	p value > 0.10	p value > 0.05	p value > 0.01
PPT/TAVE	T p value > α	26.0	18.80	9.10	5.7	4.15	2.71
	T and P	53.41	66.67	83.09	85.59	90.24	94.51
	p value > α

Figure 10. Percent area of the CONUS at which D_X values (x-axis position of the KS test statistic) occur for SD PPT, TAVE, and RUN.

The distribution of D_X and D_B values for all 95 SD datasets (Figure 4) for PPT, TAVE, and RUN for CMIP3 and CMIP5 are shown in Figures 10 and 11, and maps of these two metrics by HRU across the CONUS are shown in Figure 12. Figures 10, 12a, and 12c illustrate that the D_X statistic occurs most frequently in the lowest quarter of the distribution for both SD PPT and RUN. The distribution of D_X for SD TAVE across HRUs shows a bimodal distribution with peaks in both the lower and upper halves of the distribution (Figure 10). The distributional characteristics of PPT and TAVE D_X , as they relate to RUN D_X , indicate that low RUN can be related to both low PPT and high TAVE (high climatic water demand) in many parts of the CONUS, especially the portions of western CONUS identified earlier. However, considering the results from Figures 5 and 6, the ability of SD RUN to replicate historical conditions appears to be more dependent on the ability of SD PPT to replicate historical conditions than on the reliability of TAVE simulations, especially in water-limited environments in the western CONUS. This is illustrated in Figures 12a–c, which show that much of the southwestern region of CONUS has a similar range of D_X for both SD PPT and RUN.

The results for the D_B statistic (Figures 11, 12d–f) show that for most HRUs across the CONUS, the value of the GSD distribution is greater than the SD distribution at the location of the KS test statistic on the x axis (D_X) for all three variables. This indicates that at the part of the distribution where the largest vertical differences between the GSD and SD variables (D ; Figure 4) occur, the SD variables underestimate historical conditions represented by the GSD for most SD PPT, TAVE, and RUN simulations at most HRUs across the country. In Figures 12d–f, a

Figure 11. Percent area of the CONUS at which D_B values (direction of bias between GSD and SD variable at D_X) occur for SD PPT, TAVE, and RUN.

clear distinction can be seen between the eastern CONUS (humid areas), where D_B appears less directly tied to PPT or RUN, and the western CONUS (arid areas), where the positive bias for SD RUN appears directly tied to SD PPT. One note is that for both CMIP3 and CMIP5 simulations used here, D_X and D_B showed similar patterns and magnitudes across the CONUS.

4. Discussion

Results presented in Figures 5–9 and Tables 1 and 2 show that for most of the CONUS, SD PPT, TAVE, and RUN have similar statistical distributions to those represented by historical climatic conditions. However, results presented in Figure 6 and Table 2 emphasize that the reliability of RUN is closely tied to reliability in PPT. While the KS test results indicated that the skill of SD PPT is greater for the CMIP5 when compared to the CMIP3, both CMIPs still shared common areas where no SD PPT simulations replicated historical PPT (Figures 5 and 6).

The hydrologic model used in this study (MWBM) uses a monthly time step accounting system, with limited/simplistic representation of groundwater reservoirs, actual evapotranspiration, and surface runoff processes (Bock et al. 2016b). Many of the HRUs in Figures 5c and 5f, where SD GCMs from both ensembles do not replicate historical RUN area, are in arid or semiarid regions. Data scarcity is an important limitation to modeling these areas. Many of the rose-colored HRUs in Figures 5c and 5f fall within hydrologic regions (regions 13, 15, and 16 in Figure 3) that have the lowest cumulative percentage of the region that is gaged for all

Figure 12. (a)–(c) The x-axis value of the KS test statistic (D_X) for the lower half of the distribution ($D_X < 0.50$) among SD variables for 95 SD GCM simulations used in this study from both CMIP datasets across the CONUS. HRUs with D_X for the specific SD variable in the upper half of the distribution ($D_X > 0.50$) are symbolized as white (number of HRUs where $D_X > 0.50$: 27 for PPT, 49 985 for TAVE, and 21 for RUN). (d)–(f) Percent of SD simulations that underestimate the GSD simulation at the D_X (D_B) for the SD variables from both CMIP datasets across the CONUS.

hydrologic regions in the CONUS (Table 1; Kiang et al. 2013). In addition, these areas where SD RUN does not replicate historical RUN coincide with areas where other hydrologic models applied at both monthly and daily time steps have struggled to simulate RUN with a high degree of skill (Martinez and Gupta 2010; Newman et al. 2015). This indicates the need for better representation in PPT and the use of subdaily models to improve estimates of RUN in these areas.

While the significance level applied in this study (0.05; Figures 5–7, Table 1) is a commonly used level of significance and maintains consistency with previous work (Hay et al. 2014), a growing body of literature (Cohn and Lins 2005; Johnson 2013;

Figure 13. Percent area of the CONUS at which percentages of the (a) 38 SD GCM simulations for CMIP3 and (b) 57 GCM simulations for CMIP5 used in this study replicate historical conditions for SD PPT, TAVE, and RUN for CMIP3 and CMIP5 using a KS test across a range of significance levels.

Hirsch et al. 2015) has cited the limitations of null-hypothesis significance testing, especially as it relates to the “arbitrary nature of the selected value of α , the significance level” (Hirsch et al. 2015). This criticism is focused on the use of null hypothesis significance testing using a single significance value to reject or not reject the null hypothesis, as it focuses on type I errors at the expense of type II errors (Cohn and Lins 2005; Johnson 2013; Hirsch et al. 2015).

Figure 13 shows the percentage of the CONUS where all SD GCMs replicate historical conditions across a range of significance levels, from 0.50 on the left x axis to 0.005 on the right for CMIP3 (left) and CMIP5 (right). There is a large difference among the sensitivities of the three variables in replicating historical conditions (p value $> \alpha$) based on the significance level chosen for the null hypothesis. While the choice of significance level has no discernible effect on the percentages of TAVE for both CMIP3 and CMIP5, there are very noticeable differences in the percentages of PPT and RUN as the significance value decreases. While the choice of significant levels applied to the KS test can be arbitrary, sensitivity to the level of significance should be a consideration for the application of the method to “subset” SD GCM selection for a given location or variable.

While the detection of significant differences is a key component of this study, the parts of the distribution where the differences (Figures 10, 12a–c; D_X) between the “simulated” and “truth” are the greatest and the general direction of bias between the two (Figures 11, 12d–f; D_B) were also identified. For the HRUs in the western CONUS where SD PPT and RUN failed to replicate historical conditions, the low values of D_X for PPT and RUN indicate that the largest difference between SD and GSD occurs in the low tail of the distribution. For these same HRUs, PPT and RUN from SD simulations underestimated the PPT and RUN from GSD. These limitations are important to consider, as many researchers utilize SD data for studies relating to environmental events, such as drought frequency, ecological flows, and

soil moisture deficits, where the highest and lowest ranges of a variable have more importance than the median/mean characteristics of a variable's distribution.

Evaluation of SD GCM simulations for replicating historical climatic conditions suggests some guidelines for use of projections of climate. While it seems reasonable to place greater confidence in SD GCMs that reliably simulate historical climatic conditions, this does not necessarily mean the simulations of future climatic conditions from these same GCMs will be accurate. There can be problems with stationarity (Milly et al. 2008), and model parameters calibrated to historical climatic conditions might not be relevant for future climatic conditions (Charles et al. 1999; Blöschl and Montanari 2010; Milly and Dunne 2017). Results from this study indicate no “best” overall GCM. Different downscaling procedures and hydrologic models of varying complexity give different results. However, this type of analysis can provide an indication of the usefulness of SD GCM simulations for specific studies and geographic locations.

Results from this study indicate that for the majority of the CONUS, SD GCM simulations reliably represent historical climatic conditions (Figure 5). Results show that SD distributions of PPT, TAVE, and RUN replicate distributions of historical PPT, TAVE, and RUN for approximately 70%, 99%, and 82% of the CONUS, respectively. For some locations in the CONUS, SD GCM simulations from CMIP3 and CMIP5 ensembles may not be reliable for studies or decision-making at monthly time scales. This draws into question conclusions from studies that have used SD GCM simulations in these areas. In these locations, if the monthly simulations are not similar in basic distributional properties to historical data (i.e., GSD), then daily downscaled data are highly suspect. California is an example where improvement in SD GCM simulations are needed. California experiences high year-to-year variability in PPT (Dettinger and Cayan 2014), and large areas of California indicate that none of the CMIP3 or CMIP5 simulations are able to replicate historical climatic conditions of PPT, TAVE, or RUN for California (Figure 6). Until improved SD GCM simulations of PPT are available, estimates of future climatic conditions should be used with caution.

5. Conclusions

The accuracy of SD GCM simulations of monthly PPT and TAVE for historical climatic conditions (1950–2005) and the implications when these climate simulations are used to drive an MWBM to estimate RUN were assessed for the CONUS. The distributional similarities among SD PPT, TAVE, and RUN for historical conditions with measured PPT, TAVE, and RUN were examined using the KS test and two additional metrics (D_X and D_B). For the majority of the CONUS, SD GCM simulations from both CMIP3 and CMIP5 reliably simulate historical climatic conditions based on the KS test, with a significance level of 0.05. However, SD GCM simulations from CMIP3 showed much less skill, compared to CMIP5. There are some geographic locations where SD GCM simulations of PPT and RUN are not reliable for studies or decision-making at monthly time scales based on the KS test results presented here. For PPT and RUN at these locations, the SD simulations were least accurate in replicating historical conditions in the low tails of the cumulative distributions (low D_X values) and consistently underestimated historical conditions represented by the

GSD (high D_B values). The results shown in this study may vary in other applications by GCM, variable, and geographic location, as well as by downscaling procedure. In any application that uses SD GCMs, a simple KS test can be applied for guidance on determining the most reliable or most appropriate set of SD GCMs to use in a climate change study.

Acknowledgments. This research was supported by the U.S. Environmental Protection Agency Office of Water, the U.S. Department of the Interior South Central Climate Science Center, the U.S. Geological Survey (USGS) Water Availability and Use Science Program, and the USGS Community for Data Integration. Further technical support was provided by the USGS Core Science Systems Mission Area. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. government.

REFERENCES

- Blodgett, D. L., 2013: The U.S. Geological Survey Climate Geo Data Portal: An integrated broker for climate and geospatial data. U.S. Geological Survey Fact Sheet 2013-3019, 2 pp.
- Blöschl, G., and A. Montanari, 2010: Climate change impacts—Throwing the dice? *Hydrol. Processes*, **24**, 374–381, <https://doi.org/10.1002/hyp.7574>.
- Bock, A. R., L. E. Hay, S. L. Markstrom, and R. D. Atkinson, 2016a: Monthly water balance model futures: U.S. Geological Survey data release. U.S. Geological Survey, accessed 15 June 2016, <https://doi.org/10.5066/F7VD6WJQ>.
- , —, G. J. McCabe, S. L. Markstrom, and R. D. Atkinson, 2016b: Parameter regionalization of a monthly water balance model for the conterminous United States. *Hydrol. Earth Syst. Sci.*, **20**, 2861–2876, <https://doi.org/10.5194/hess-20-2861-2016>.
- , —, S. L. Markstrom, C. E. Emmerich, and M. Talbert, 2017: The U.S. Geological Survey monthly water balance model futures portal. U.S. Geological Survey Rep. 2016-1212, 32 pp., <https://pubs.usgs.gov/of/2016/1212/ofr20161212.pdf>.
- Bureau of Reclamation, 2011: West-Wide climate risk assessments: Bias-corrected and spatially downscaled surface water projections. U.S. Department of the Interior/Bureau of Reclamation Tech. Memo. 86-68210-2011-01, 138 pp, <https://www.usbr.gov/watersmart/docs/west-wide-climate-risk-assessments.pdf>.
- , 2013: Downscaled CMIP3 and CMIP5 climate projections. Bureau of Reclamations Rep., 110 pp, http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/techmemo/downscaled_climate.pdf.
- Charles, S. P., B. C. Bates, P. H. Whetton, and J. P. Hughes, 1999: Validation of downscaling models for changed climate conditions: Case study of southwestern Australia. *Climate Res.*, **12**, 1–14, <https://doi.org/10.3354/cr012001>.
- Cohn, T. A., and H. F. Lins, 2005: Nature's style: Naturally trendy. *Geophys. Res. Lett.*, **32**, L23402, <https://doi.org/10.1029/2005GL024476>.
- Conover, W. J., 1971: *Practical Nonparametric Statistics*. John Wiley & Sons, 462 pp.
- Dettinger, M. D., and D. R. Cayan, 2014: Drought and the California delta—A matter of extremes. *San Francisco Estuary Watershed Sci.*, **12** (2), <https://doi.org/10.15447/sfew.2014v12iss2art4>.
- Farnsworth, R. K., E. S. Thompson, and E. L. Peck, 1982: Evaporation atlas for the contiguous 48 United States. NOAA Tech. Rep. NWS 33, 41 pp., http://www.nws.noaa.gov/oh/hdsc/Technical_reports/TR33.pdf.
- Hamon, W. R., 1961: Estimating potential evapotranspiration. *J. Hydraul. Div. ASCE*, **87**, 107–120.
- Hay, L. E., and G. J. McCabe, 2002: Spatial variability in water-balance model performance in the conterminous United States. *J. Amer. Water Resour. Assoc.*, **38**, 847–860, <https://doi.org/10.1111/j.1752-1688.2002.tb01001.x>.

- , M. P. Clark, R. L. Wilby, W. J. Gutowski, G. H. Leavesley, Z. Pan, R. W. Arritt, and E. S. Takle, 2002: Use of regional climate model output for hydrologic simulations. *J. Hydrometeor.*, **3**, 571–590, [https://doi.org/10.1175/1525-7541\(2002\)003<0571:UORCMO>2.0.CO;2](https://doi.org/10.1175/1525-7541(2002)003<0571:UORCMO>2.0.CO;2).
- , J. H. LaFontaine, and S. L. Markstrom, 2014: Evaluation of statistically downscaled GCM output as input for hydrological and stream temperature simulation in the Apalachicola–Chattahoochee–Flint River basin (1961–99). *Earth Interact.*, **18**, 1–32, <https://doi.org/10.1175/2013EI000554.1>.
- Hirsch, R. M., S. A. Archfield, and L. A. De Cicco, 2015: A bootstrap method for estimating uncertainty of water quality trends. *Environ. Modell. Software*, **73**, 148–166, <https://doi.org/10.1016/j.envsoft.2015.07.017>.
- House of Representatives, 2009: Concurrent Resolution 146.
- Johnson, V. E., 2013: Revised standards for statistical evidence. *Proc. Natl. Acad. Sci. USA*, **110**, 19 313–19 317, <https://doi.org/10.1073/pnas.1313476110>.
- Kiang, J. E., D. W. Stewart, S. A. Archfield, E. B. Osborne, and K. Eng, 2013: A national streamflow gap analysis. U.S. Geological Survey Scientific Investigations Rep. 2013-5013, 82 pp., <https://pubs.usgs.gov/sir/2013/5013/pdf/sir2013-5013.pdf>.
- Kolmogorov, A. N., 1933: On the empirical determination of a distribution function. (in Italian). *Giornale dell’Istituto Italiano degli Attuari*, **4**, 83–91.
- , 1941: Confidence limits for an unknown distribution function. *Ann. Math. Stat.*, **12**, 461–463, <https://doi.org/10.1214/aoms/1177731684>.
- Liu, M., and Coauthors, 2014: What is the importance of climate model bias when projecting the impacts of climate change on land surface processes? *Biogeosciences*, **11**, 2601–2622, <https://doi.org/10.5194/bg-11-2601-2014>.
- Martinez, G., and H. V. Gupta, 2010: Toward improved identification of hydrologic models: A diagnostic evaluation of the “abcd” monthly water balance model for the conterminous United States. *Water Resour. Res.*, **46**, W08507, <https://doi.org/10.1029/2009WR008294>.
- Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen, 2002: A long-term hydrologically based dataset of land surface fluxes for the conterminous United States. *J. Climate*, **15**, 3237–3251, [https://doi.org/10.1175/1520-0442\(2002\)015<3237:ALTHBD>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2).
- , L. Brekke, T. Pruitt, and P. B. Duffy, 2007: Fine-resolution climate projections enhance regional climate change impact studies. *Eos, Trans. Amer. Geophys. Union*, **88**, 504, <https://doi.org/10.1029/2007EO470006>.
- McCabe, G. J., and S. L. Markstrom, 2007: A monthly water-balance model driven by a graphical user interface. U.S. Geological Survey Open-File Rep. 2007-1008, 12 pp., https://pubs.usgs.gov/of/2007/1088/pdf/of07-1088_508.pdf.
- , and D. M. Wolock, 2011a: Century-scale variability in global annual runoff examined using a water balance model. *Int. J. Climatol.*, **31**, 1739–1748, <https://doi.org/10.1002/joc.2198>.
- , and —, 2011b: Independent effects of temperature and precipitation on modeled runoff in the conterminous United States. *Water Resour. Res.*, **47**, W11522, <https://doi.org/10.1029/2011WR010630>.
- , L. E. Hay, A. Bock, S. L. Markstrom, and D. R. Atkinson, 2015: Inter-annual and spatial variability of Hamon potential evapotranspiration model coefficients. *J. Hydrol.*, **521**, 389–394, <https://doi.org/10.1016/j.jhydrol.2014.12.006>.
- Milly, P. C. D., and K. A. Dunne, 2017: A hydrologic drying bias in water-resource impact analyses of anthropogenic climate change. *J. Amer. Water Resour. Assoc.*, **53**, 822–838, <https://doi.org/10.1111/1752-1688.12538>.
- , J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer, 2008: Stationarity is dead: Whither water management? *Climatic Change*, **319**, 573–574, <https://doi.org/10.1126/science.1154915>.
- Newman, A. J., and Coauthors, 2015: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of

- regional variability in hydrologic model performance. *Hydrol. Earth Syst. Sci.*, **19**, 209–223, <https://doi.org/10.5194/hess-19-209-2015>.
- Papadimitriou, L. V., A. G. Koutroulis, M. G. Grillakis, and I. K. Tsanis, 2017: The effect of GCM biases on global runoff simulations of a land surface model. *Hydrol. Earth Syst. Sci.*, **21**, 4379–4401, <https://doi.org/10.5194/hess-21-4379-2017>.
- R Core Team, 2015: R: A language and environment for statistical computing. R Foundation for Statistical Computing, accessed 1 March 2015, <https://www.R-project.org/>.
- Sheffield, J., and Coauthors, 2014: Regional climate processes and projections for North America: CMIP3/CMIP5 differences, attribution and outstanding issues. NOAA Tech. Rep. OAR CPO-2, 48 pp.
- Smirnov, N. V., 1939: On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Moscow Univ.*, **2**, 3–14.
- Stoner, A. M. K., K. Hayhoe, X. Yang, and D. J. Wuebbles, 2013: An asynchronous regional regression model for the statistical downscaling of daily climate variables. *Int. J. Climatol.*, **33**, 2473–2494, <https://doi.org/10.1002/joc.3603>.
- Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc. London*, **365A**, 2053–2075, <https://doi.org/10.1098/rsta.2007.2076>.
- U.S. EPA and USGS, 2010: NHDPlus version 1 user guide. U.S. Environmental Protection Agency and U.S. Geological Survey Rep., 126 pp., ftp://ftp.horizon-systems.com/nhdplus/nhdplusv1/documentation/nhdplusv1_userguide.pdf.
- Viger, R., and A. R. Bock, 2014: GIS features of the Geospatial Fabric for National Hydrologic Modeling. USGS, accessed October 2014, <https://doi.org/doi:10.5066/F7542KMD>.
- Wolock, D. M., 1997: STATSGO soil characteristics for the conterminous United States. U.S. Geological Survey Open-File Rep. 1997-656, <http://water.usgs.gov/GIS/metadata/usgswrd/XML/muid.xml>.
- , and G. J. McCabe, 1999: Explaining spatial variability in mean annual runoff in the conterminous United States. *Climate Res.*, **11**, 149–159, <https://doi.org/10.3354/cr011149>.
- Wood, A. W., L. R. Leung, V. Sridhar, and D. P. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, **62**, 189–216, <https://doi.org/10.1023/B:CLIM.0000013685.99609.9e>.