Evaluations of high-resolution dynamically downscaled ensembles over the contiguous United States

Article *in* Climate Dynamics · February 2018

DOI: 10.1007/s00382-017-3645-6

CITATIONS
17

READS
108

4 authors, including:

Jiali Wang
Argonne National Laboratory
25 PUBLICATIONS **329** CITATIONS
SEE PROFILE

Donald J. Wuebbles
University of Illinois, Urbana-Champaign
197 PUBLICATIONS **8,375** CITATIONS
SEE PROFILE

V. R. Kotamarthi
Argonne National Laboratory
128 PUBLICATIONS **2,141** CITATIONS
SEE PROFILE

Clim Dyn
DOI 10.1007/s00382-017-3645-6

Evaluations of high-resolution dynamically downscaled ensembles over the contiguous United States

Zachary Zobel[1] · **Jiali Wang**[2] · **Donald J. Wuebbles**[1] · **V. Rao Kotamarthi**[2]

Received: 3 June 2016 / Accepted: 16 March 2017

Abstract This study uses Weather Research and Forecast (WRF) model to evaluate the performance of six dynamical downscaled decadal historical simulations with 12-km resolution for a large domain (7200×6180 km) that covers most of North America. The initial and boundary conditions are from three global climate models (GCMs) and one reanalysis data. The GCMs employed in this study are the Geophysical Fluid Dynamics Laboratory Earth System Model with Generalized Ocean Layer Dynamics component, Community Climate System Model, version 4, and the Hadley Centre Global Environment Model, version 2-Earth System. The reanalysis data is from the National Centers for Environmental Prediction-US. Department of Energy Reanalysis II. We analyze the effects of bias correcting, the lateral boundary conditions and the effects of spectral nudging. We evaluate the model performance for seven surface variables and four upper atmospheric variables based on their climatology and extremes for seven subregions across the United States. The results indicate that the simulation's performance depends on both location and the features/variable being tested. We find that the use of bias correction and/or nudging is beneficial in many situations, but employing these when running the RCM is not always an improvement when compared to the reference data. The use of an ensemble mean and median leads to a better performance in measuring the climatology, while it is significantly biased for the extremes, showing much larger differences than individual GCM driven model simulations from the reference data. This study provides a comprehensive evaluation of these historical model runs in order to make informed decisions when making future projections.

Keywords Dynamical downscaling · Statistical evaluation · Ensemble · Climate extremes · Regional climate models · Global climate models

1 Introduction

General circulation models (GCMs) show significant skill at the continental scale and incorporate a large proportion of the complexity of the global system (Sheffield et al. 2013a, b; IPCC AR5). However, they are restricted in their capacity for representing local or small scale features (e.g., topography) because of their coarse spatial resolution (Wigley et al. 1990; Carter et al. 1994; Wang et al. 2015). When considering the impacts of global climate change, the focus is primarily on the physical effects occurring at regional, local, and even smaller scales (e.g., a watershed). In order to provide well-informed projections of future changes in climate for planners and the public, analyses of these effects for spatial scales on the order of tens, not hundreds, of kilometers are necessary. This information is particularly needed in regions where global model's resolution cannot capture the orographic variations sufficiently enough to represent observed changes accurately and for an improved ability at modeling regional extremes associated with rare climate events.

Downscaling methods have been developed to overcome some of these issues (Wigley et al. 1990; Giorgi et al. 1991; Wilby and Wigley 1997). There are two common

✉ Zachary Zobel
zobel2@illinois.edu

[1] Department of Atmospheric Sciences, University of Illinois Champaign-Urbana, 105 South Gregory Street, Champaign, IL 61801, USA

[2] Division of Environmental Sciences, Argonne National Laboratory, Chicago, IL, USA

Published online: 29 March 2017

downscaling methods, statistical downscaling (SD) and dynamical downscaling (DD). While the main concept of SD is to establish linear or non-linear relationships between subgrid-scale parameters and coarser resolution predictor (Wilby and Wigley 1997), the concept for DD is to embed a higher-resolution limited-area climate model within the GCM, using the GCM to define the time varying boundary conditions (Giorgi 1990; Mearns et al. 1995). The benefits and limitations of each downscaling method are well documented in the literature (e.g., von Storch et al. 1993; Wilby 1994; Leung et al. 2003). While there are acknowledged limitations of DD approach, for example, many regional DD utilize the parameterization developed for coarse resolution models without much modification, DD has the ability to simulate small-scale surface forcing, such as topography and vegetation type. If the model grid size is reduced to less than a few kilometers, the parameterization of the sub-grid scale process can be eliminated. Therefore, the model can simulate weather/climate with much less approximation (Hong and Kanamitsu 2014). Di Luca et al. (2012) showed that Regional Climate Models (RCMs) add high value for warm-season precipitation over short temporal scales, especially over regions of complex topography. Pryor et al. (2012) noted that an increase in RCM resolution from 50 to 6 km captures extreme wind speeds more realistically. Using a similar resolution to this study, Lee and Hong (2014) and Lee et al. (2014a, b) show that DD at higher resolutions perform better at capturing upper atmosphere dynamics and distributions, especially in areas of complex terrain. Vautard et al. (2013) found that heat extremes in Europe were generally better simulated in RCMs with resolution of 12 versus 50 km. Tripathi and Dominguez (2013) found that 10 km simulation captures individual extreme summer precipitation events better than 50 km simulation. Gao et al. (2012) also showed that the DD approach is more capable than the GCM input at capturing extreme events in the eastern part of the United States with improvements higher than 90% for heat wave duration and frequency as well as extreme precipitation. Wang et al. (2015) found that a RCM at 12 km spatial resolution captures significantly more details of the spatial and temporal variations of precipitation (especially over mountainous regions) than does the 2.5-degree National Centers for Environmental Prediction-US. Department of Energy Reanalysis II (NCEP-R2) (Kalnay et al. 1996), even when the RCM is aggregated to the grid resolution of NCEP-R2.

Uncertainties in the RCM simulations can come from a number of sources, including physics parameterization, model representation of internal variability, the choice of emission scenarios for projecting future changes in climate, and the differences in the global climate models used to drive the RCM (Giorgi and Bi 2000; Mearns et al. 2012). Development of multi-RCM ensembles have been identified as one of the current research needs (IPCC4, Doherty et al. 2009) to reduce uncertainties in projections of climate. For example, in order to explore the uncertainties due to various GCMs and RCMs applied to investigate regional climate change, the North American Regional Climate Change Assessment Program (NARCCAP) simulated 30 years in historical and future periods respectively by using six RCMs driven by different GCMs over North America (Mearns et al. 2012); the Prediction of Regional Scenarios and Uncertainties for Defining European Climate Change Risks and Effects (PRUDENCE) (Christensen et al. 2008) and ensembles (van der Linden and Mitchell 2009; Christensen et al. 2010) simulated historical and future periods over Europe. The coordinated regional climate downscaling experiment (CORDEX) has built a climate projection framework for different regions over the entire globe (Giorgi et al. 2009) and aims to explore the maximum extent of the contribution of different sources of uncertainty such as different boundary conditions from various GCMs, different initializations (internal variability), and different emission/concentration scenarios.

Regarding the uncertainty that is induced by different lateral boundary conditions, Knutti et al. (2013) found that there is no single GCM that stands out as being particularly better or worse across all analyzed variables over the entire model domain. Different simulations using the same RCM driven with multiple boundary conditions show varying skill over particular regions. For example, Halmstad et al. (2013) investigate the NARCCAP six simulations for the historical period over the Villamette River Basion, Oregon, and find that the weather research and forecasting (WRF) regional climate model performs better in extreme precipitation than its boundary conditions when driven by the Community Climate System Model, version 3, but performs worse than the boundary conditions when driven by the Canadian Climate Centre Coupled General Circulation Model version 3. Pryor et al. (2012) analyze NARCCAP simulations over six subregions of CONUS, and find that every single RCM's performance in wind climate (mean and extreme) can be very different when they are driven by varying boundary conditions.

Our study is the first analysis of the North American continent at a very high spatial resolution (12 km) using boundary conditions derived from a range of different coupled model intercomparison project phase 5 (CMIP5) (Taylor et al. 2012) models for decadal length scale simulations for diagnostic studies and projections of time slices in the future (Wang and Kotamarthi 2015). We choose three different GCMs, they are geophysical fluid dynamics laboratory earth system model with generalized ocean layer dynamics component (GFDL-ESM2G), community climate system model, version 4 (CCSM4), and the hadley centre global environment model, version 2-earth system

(HadGEM2-ES). As presented by Sherwood et al. (2014), the ultimate change in global mean temperature in response to a doubling of atmospheric CO2 in CMIP5 models, span roughly 2.1 to 4.6 °C. Among more than 30 GCMs, GFDL-ESM2G is one of those models which have very low response to the increase of CO2, with global mean temperature increasing by 2.38 °C, while HadGEM2-ES is one out of two models that have the highest response to the doubling of CO2, with global mean temperature increasing by 4.55 °C. CCSM4 shows a response which is in between above two models, with global mean temperature increasing by 2.92 °C. Therefore, these three GCMs capture the range of responses of the climate models to RCP scenarios without doing every GCM in between. This study aims to rank the 12-km model performance based on different meteorological fields through an evaluation of a six-member ensemble of RCMs over seven subregions of CONUS. We seek to address questions that are raised by the user community, for example, which model has the smallest bias over an area of interest? We provide a possible path for a user by ranking the various sets of model runs that are performed at a high spatial resolution. The laborious process followed to carefully perform this task will give the user community an option of carefully selecting the outputs for specific uses. There has been extensive analysis on the effectiveness of downscaling to evaluate regional climate (e.g. Fowler et al. 2007; Wang et al. 2016; Xue et al. 2014), but this study's use of a 12-km and multi-GCMs over a large domain is unique among other downscaling research. This project uses a similar domain as the NARCCAP projections covering most of North American continent, but with approximately a four times higher resolution. The increased resolution by this factor allows for more applications over small spatial scale such as watershed scale. The value of using three separate GCMs to create an ensemble over North America will allow for an improved ability to capture the future uncertainties. Sect. 2 describes the regional climate model and GCMs applied in this study as well as the reference data. Sect. 3 ranks the model performance in terms of relative error and extremes. Discussion and summary follow in Sect. 4.

2 Model description and reference data

2.1 Regional climate models

The WRF model version 3.3.1 (Skamarock et al. 2008) is applied at a horizontal resolution of 12 km, with 600 west-east × 515 south-north grid points over most of North America (Fig. 1 in Wang and Kotamarthi 2014) (WK14 hereafter). The lateral boundary conditions are specified in two different ways. As shown in Table 1, in the first set of the experiments, the WRF model is driven by the reanalysis of the NCEP-R2 (Kalnay et al. 1996) over the period 1980–2010. In the second to the sixth sets of the experiments, the WRF models are driven by datasets from three fully coupled GCMs. The evaluation of model performance focuses not only on the climatology, but also on the extreme climate events. The atmospheric fields of interest include near surface variables as well as above surface (e.g., the most common studied 850, 500, and 200 hPa) fields. The second to sixth sets of experiments span three different time periods: 11 years over historical period (1994–2004). The name of each GCM dataset is listed in Table 1.

The six WRF model runs listed in Table 1 are the same in horizontal resolution. They are also the same in most of the physical parameterizations, which includes the Grell-Devenyi convective parameterization (Grell and Devenyi 2002), the Yonsei University planetary boundary layer scheme (Hong et al. (2006); Noh et al. 2003), the Noah land surface model (Chen and Dudhia 2001), as well as the longwave and shortwave radiative schemes of the Rapid Radiation Transfer Model for GCM applications (http://rtweb.aer.com) (Iacono et al. 2008). However, as shown in Table 1, the first WRF run, driven by NCEP-R2, use WSM6 (Hong and Lim 2006) microphysics and it applies spectral nudging with a nudging coefficient of 3×10^{-4} s^{-1} (the strength of nudging) Moreover, it only allows 1 day as spin-up time and is re-initialized every year. Among the six WRF simulations, the NCEP-R2 driven run was conducted first, along with which we conducted sensitivity experiments considering different nudging strength, microphysics, convective parameterizations and spin-up time. The details are referred to WK14. The sensitivity study showed that using weaker nudging, Morrison microphysics, and longer spin-up time helps reducing the model bias in several different aspects, respectively. Therefore, we adjust these settings for the GCM driven runs using the Morrison microphysics scheme (Morrison et al. 2009) and 1-year spin-up time for each 10-yr continuous run. For those runs who apply spectral nudging, the nudging strength is 3×10^{-5} s^{-1}. We apply weak spectral nudging to air temperature, geopotential height, and wind for levels above 850 hPa. See Wang and Kotamarthi (2013) for more details about spectral nudging applied in this study. It is worth mentioning that, the comparison between WRF simulation driven by NCEP-R2 and that driven by GCMs does not aim to investigate the effect of any one of the different model setup or physics (e.g., microphysics). As we stated in the introduction, one goal of this study is to provide the users a possible path to rank the various sets of model runs that fit their unique needs. While the study by Wang and Kotamarthi (2015) compare the impacts of bias correction using CCSM4 to drive the WRF model, they only focus on precipitation over different regions of North America. This study compares

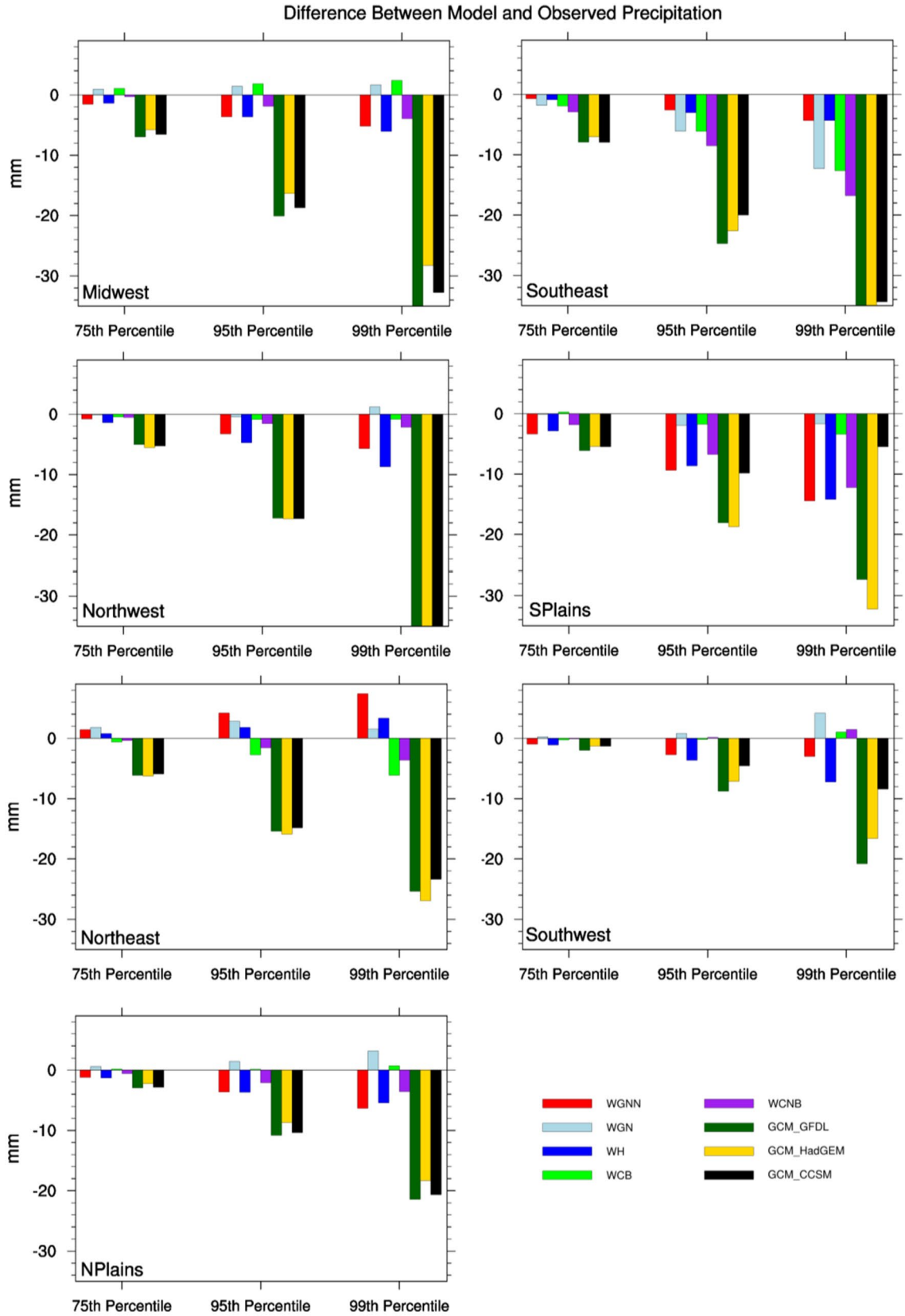

Fig. 1 Difference between model and observed precipitation percentiles (model-observed) in the PDF curve for WRF driven GCM simulations and GCM raw data. Regions shown from *top* to *bottom* and left to right are Midwest, Southeast, Northwest, and SPlains (SGP)

Table 1 The lateral boundary conditions that drive the six WRF model runs as well as information about microphysics, spin-up time, spectral nudging, and bias correction that are applied

	Lateral boundary conditions	WRF Simulation	Microphysics	Spin-up time	Spectral nudging (nudging strength)	Bias correction
1	NCEP-R2	WN	WSM6	1 day	Yes (3×10^{-4} s^{-1})	No
2	CCSM4	WCNB	Morrison	1 year	Yes (3×10^{-5} s^{-1})	No
3	CCSM4	WCB	Morrison	1 year	Yes (3×10^{-5} s^{-1})	Yes
4	GFDL-ESM2G	WGNN	Morrison	1 year	No	Yes
5	GFDL-ESM2G	WGN	Morrison	1 year	Yes (3×10^{-5} s^{-1})	Yes
6	HadGEM2-ES	WH	Morrison	1 year	No	No

not only the effect of bias correction, but also the effect of spectral nudging and different lateral boundary conditions on the model performance. The GCMs include CCSM4 developed by National Center for Atmospheric Research, United States (Gent et al. 2011), GFDL-ESG2G developed by NOAA/Geophysical Fluid Dynamics Laboratory, United States (Donner et al. 2011), and HadGEM2-ES developed by Met Office Hadley Centre, United Kingdom (Jones et al. 2011). To explore the impacts of spectral nudging on model performance when bias correction is applied, we conducted two WRF runs driven by GFDL-ESG2G, with spectral nudging turned on in one of the simulations and turned off in the other simulation. In addition to these six simulations, two more individual datasets are incorporated into this ensemble—the mean and median of the six simulations at each grid point for the 10-year period.

2.2 Reference data

We employ North American Regional Reanalysis (NARR) data (Mesinger et al. 2006; Bukovsky and Karoly 2007) to evaluate model performance in near surface relative humidity, wind, and high level fields, such as geopotential height, humidity, and wind. The NARR is on a spatial resolution of 32 km and covers more than 30 years from 1979 to present. The NARR assimilates observed information from multiple sources (aircraft, satellite, stations, etc.) (Table 1 and 2 in Mesinger et al. 2006), and has been used widely as reference data by the climate downscaling community (e.g., Bowden et al. 2016; Otte et al. 2012; Liu et al. 2011; Loikith et al. 2013), although inaccuracies remain in some regions. For example, Bukovsky and Karoly (2007) found that, while the NARR provides a fairly good representation of observed precipitation over much of CONUS, some inaccuracies are found over Canada because of the relatively poor data quality that NARR assimilates. Wang et al. (2016) found that NARR overestimates (underestimates) the warming trend of January temperature over southeastern CONUS (over most of western CONUS).

For other near surface fields, such as daily maximum and minimum temperature and precipitation, we use a gridded dataset based on observations (Maurer et al. 2002). This gridded dataset is on a spatial resolution of 1/8 degree and covers 66 years from 1950 to 2015. It has been applied extensively as meteorological references for evaluating dynamical and/or statistical downscaled results (e.g., Wood et al. 2004; Christensen et al. 2004; Maurer and Hidalgo 2008; Gutowski et al. 2010; Wehner 2013). The gridded precipitation within the CONUS is from the National Oceanic and Atmospheric Administration Cooperative Observer (Co-op) stations. The precipitation gauge data are first gridded to the one-eighth degree resolution using the synergraphic mapping system algorithm of Shepard (1984) as implemented by Widmann and Bretherton (2000). The gridded daily precipitation data are then scaled to match the long-term average of the parameter-elevation regressions on independent slopes model (PRISM) precipitation climatology (Daly et al. 1994, 1997), which is a comprehensive dataset that is statistically adjusted to capture local variations due to complex terrain. The minimum and maximum daily temperature data over CONUS, also obtained from Co-op stations, are gridded using the same algorithm as for precipitation, and are lapsed to the grid cell mean elevation. We also use PRISM monthly precipitation data set as the reference data to evaluate the model and understanding the uncertainty of model's performance to different reference data.

For our analysis, the CONUS is broken into seven subregions that are consistent with those used in the US. National Climate Assessment (Melillo et al. 2014). They are Northwest (NW), North Great Plains (NGP), South Great Plains (SGP), Midwest (MW), Northeast (NE), Southwest (SW), and Southeast (SE) (see Fig. 2 in Janssen et al. 2014).

3 Results

3.1 Added value by dynamical downscaling

One of the key questions in downscaling research is whether the high-resolution simulation adds value against the driving GCM data. Wang et al. (2015) developed

spatial and spatiotemporal correlations considering dynamics features of precipitation. They found the improvements are apparent not only at resolutions finer than that of GCMs, but also when the RCM and observational data are aggregated to the resolution of GCMs (not illustrated). In this study, we calculate the probability density functions (PDF) of precipitation and compare the bias of GCM and RCM at the tails of the PDF distribution to investigate the potential value added by downscaling. Figure 1 shows differences in certain percentiles (75, 95, and 99th percentile) between the model and observed PDFs over seven subregions. The 10-year precipitation PDF is calculated taking only grid points where daily precipitation is greater than 1 mm. We find that there is clear advantage by using the downscaled simulations over the raw GCM counterparts, especially in mountainous and convection dependent regions and for higher percentiles in the PDF distribution. This is an important aspect of these downscaled simulations because the ability to forecast these events has major economical and societal impacts. While the RCM data still has some shortcomings at forecasting the frequency and intensity of high impact precipitation (see Sect. 3.3.3 for more details), Fig. 1 shows that these simulations are a significant improvement over using raw GCM data, except the raw CCSM4 data shows slightly smaller bias than the WCNB run for the 99th percentile over Southern Plains region. Overall, the improvements seen in typical DD simulation over raw GCM data is higher than 90% for most of the regions, which is similar to the results described in Gao et al. (2012).

3.2 Relative error

In this and the next section, we evaluate model performance based on metrics that describes relative error of daily mean and PDF that drawing distribution tails. To describe relative error, we employ the performance metrics developed by Gleckler et al. (2008). To begin, root-mean-square-error (RMSE) is calculated for each variable and NCA subregion for all six model runs as well as their mean and median. The reference data set depends on the type of variable being analyzed. NARR is used to evaluate above surface variables (e.g., Liu et al. 2012) while the gridded observations are used to evaluate the appropriate surface variables (e.g., temperature and precipitation). Once we determine the RMSE values, we calculate relative error (i.e. error relative to the median of the members of this ensemble) for each variable. As shown in Eq. (1), to calculate relative error for a field f and model m (E'_{mf}), we define a typical model error ($\overline{E_f}$), which is the median of the eight RMSE values (six simulations plus median and mean) for that region and variable. We use median of RMSEs rather than mean as the typical model error to prevent models with unusually large errors from influencing the results (Gleckler et al. 2008). E_{mf} is the RMSE of one particular simulation out of six simulations plus the mean and median. The relative error (E'_{mf}) is a measure of how well a particular model performs compared to the typical model error in the ensemble. For example, if a model has a negative E'_{mf} this means it has a lower RMSE than the simulations with positive E'_{mf}.

$$E'_{mf} = \frac{E_{mf} - \overline{E_f}}{\overline{E_f}} \tag{1}$$

Figure 2 shows the relative error for daily precipitation (upper left), mean temperature (upper right), and daily maximum/minimum temperature (lower left and lower right) over seven NCA regions and CONUS from the WRF simulations comparing with the gridded observation data set described in Sect. 2.2. In general, the WRF simulations driven by GCMs score worse for all four variables than the ensemble mean and median. For precipitation, the WH and WGNN show less RMSE than other WRF simulations driven by GCMs in the MW and that includes the NCEP driven simulation for the NGP and SGP regions. The WN and WCNB predict lower RMSE than other WRF simulations driven by GCMs in NE and SE. There are noticeable differences between the models with and without bias correction. The relative error between WCNB and WCB has the greatest differences for precipitation in NGP, SGP, and MW. Using bias correction for these regions caused larger error than when no bias correction is applied to the boundary conditions. A similar trend is observed for models with and without spectral nudging. For example, WRF_GFDLN shows larger error in precipitation relative error than does WRF_GFDLNN run for all regions except for the NE.

It is worth mentioning that over the Great Plains, the WN shows positive relative errors, which means it has larger RMSEs than the typical model error. This is because, although we are using the "perfect" boundary conditions, the physics and the model setup are somewhat different from the other WRF simulations driven by GCMs. WN run is the first run that we have conducted for the project, aiming to understand the model performance and the model sensitivity to different physics and setup. In this run, we only allowed 1 day as spin-up time, and we re-initialized the model every year. These are two of the reasons that model shows wet bias over Great Plains. In addition, the microphysics scheme that is applied for the run also induces wet bias over Great Plain in cold seasons (WK14). Thus, we modify the model setup and microphysics for WRF simulations driven by GCMs to reduce the bias that generated by those factors.

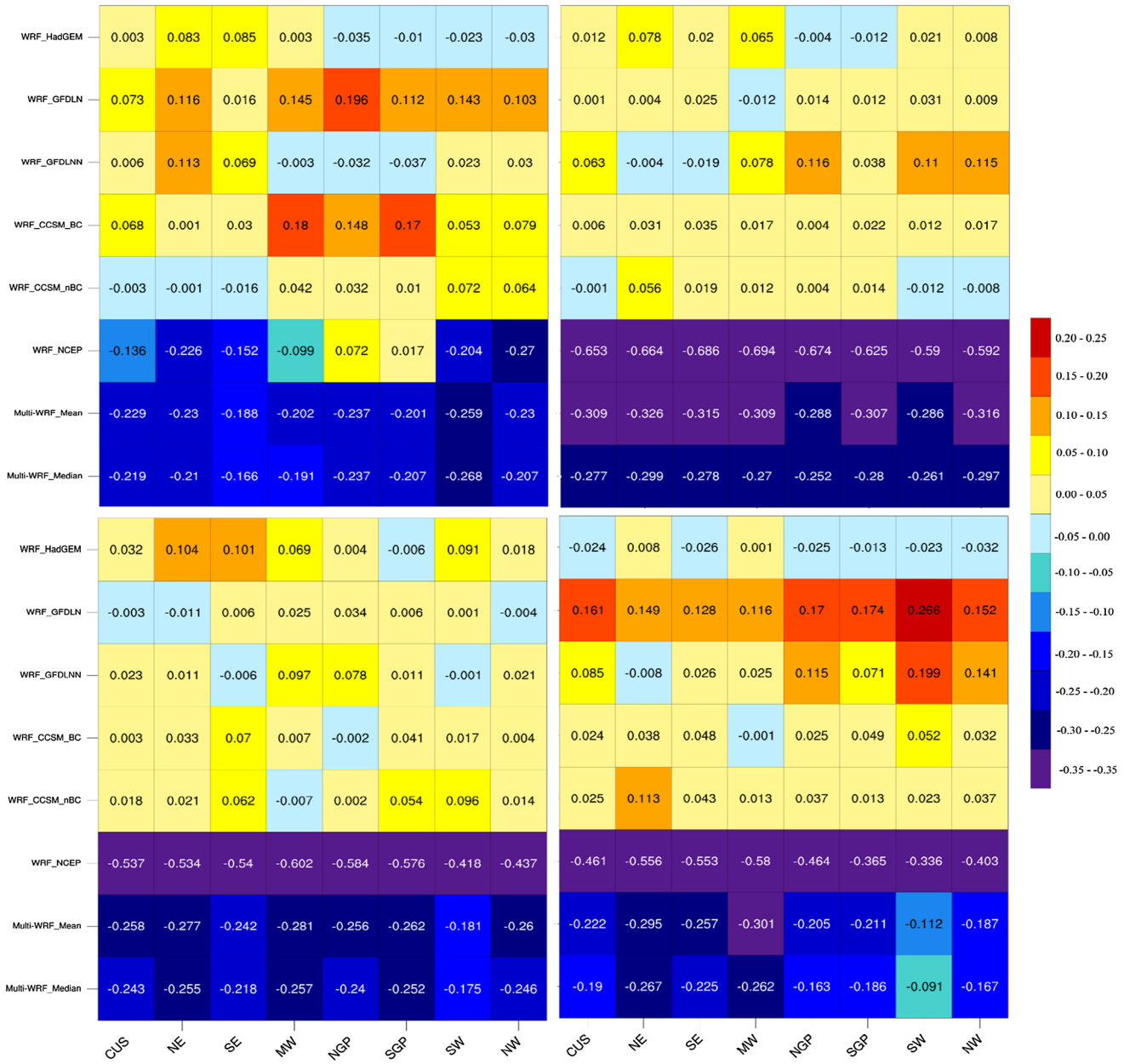

Fig. 2 Relative error of the RMSE for surface variables compared to observed gridded values: daily precipitation (*upper left*); Daily mean temperature (*upper right*); Daily maximum temperature (*lower left*); Daily minimum temperature (*lower right*)

The preferred GCM and the model setup for mean temperature (Fig. 2, upper right) and maximum temperature (Fig. 2, lower left) is regionally and simulation dependent. For minimum temperature (Fig. 2d), the WGNN shows smaller relative errors than does WGN for all the regions. WH shows the lowest relative error in comparison with other WRF simulations for all but two regions—MW and NE. There is not much difference between WCB and WCNB, but they both show far less error than the WGN run. WCB is significantly more accurate for all eight regions than the WGN for minimum temperatures. Since both simulations employ bias correction and nudging, much of the error in the WGN runs for minimum temperature is likely due to the biases in the boundary conditions of that GCM. Nudging does not necessarily improve the model performance in minimum temperature for the WGN.

Figure 3 shows relative errors of geopotential height at 500 hPa (Fig. 3, upper left), specific humidity at 850 hPa (Fig. 3, upper right), and zonal and meridional wind at 850 hPa (Fig. 3, lower left and Fig. 3, lower right) based on comparisons between WRF simulations and NARR. In general, the WN performs better than all GCM driven

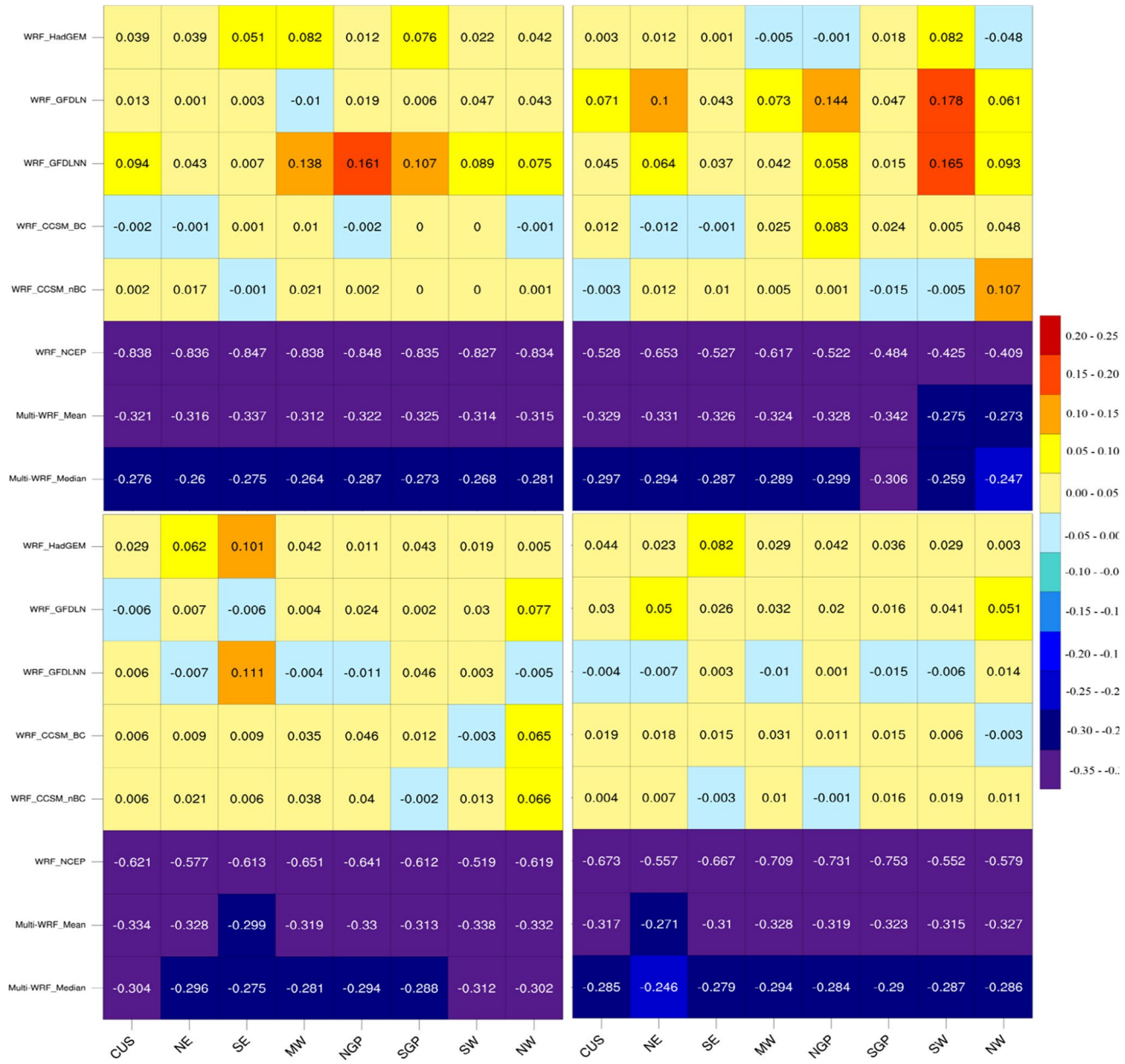

Fig. 3 Relative error of the RMSE for above surface variables compared to North American Regional Reanalysis (*NARR*): 500 hPa geopotential height (*upper left*); 850 hPa specific humidity (*upper right*); 850 hPa U component (*lower left*); 850 hPa V component (*lower right*)

simulations for above surface variables. For the rest of the ensemble, the ranks are different depending on region and variable. WGNN and WGN outperform the other GCM driven simulations for many of the regions for low-level winds at the 850 hPa (Fig. 3, lower left and lower right), but rank in the bottom two for most of the regions for 500 hPa geopotential height and 850 hPa specific humidity (Fig. 3, upper left and upper right)—where WH and WCB/WCNB runs are superior. Depending on the region and variable, there are several instances where nudging has a significant difference between WGN and WGNN in Fig. 3. For example, nudging helps reducing the relative error in all eight regions for 500 hPa geopotential height (Fig. 3, upper left), but yields higher RMSE values for specific humidity with the exception of the NW (Fig. 3, upper right). The use of bias correction does not cause large differences between the two CCSM4 driven WRF runs in relative errors, which have ranks that are mostly near the middle of pack.

Figure 4 shows relative errors for more near surface variables (10-m wind, Fig. 4 upper left and upper right) and mean sea level pressure (SLP) (Fig. 4 lower left). The GFDL driven runs perform slightly better than the other

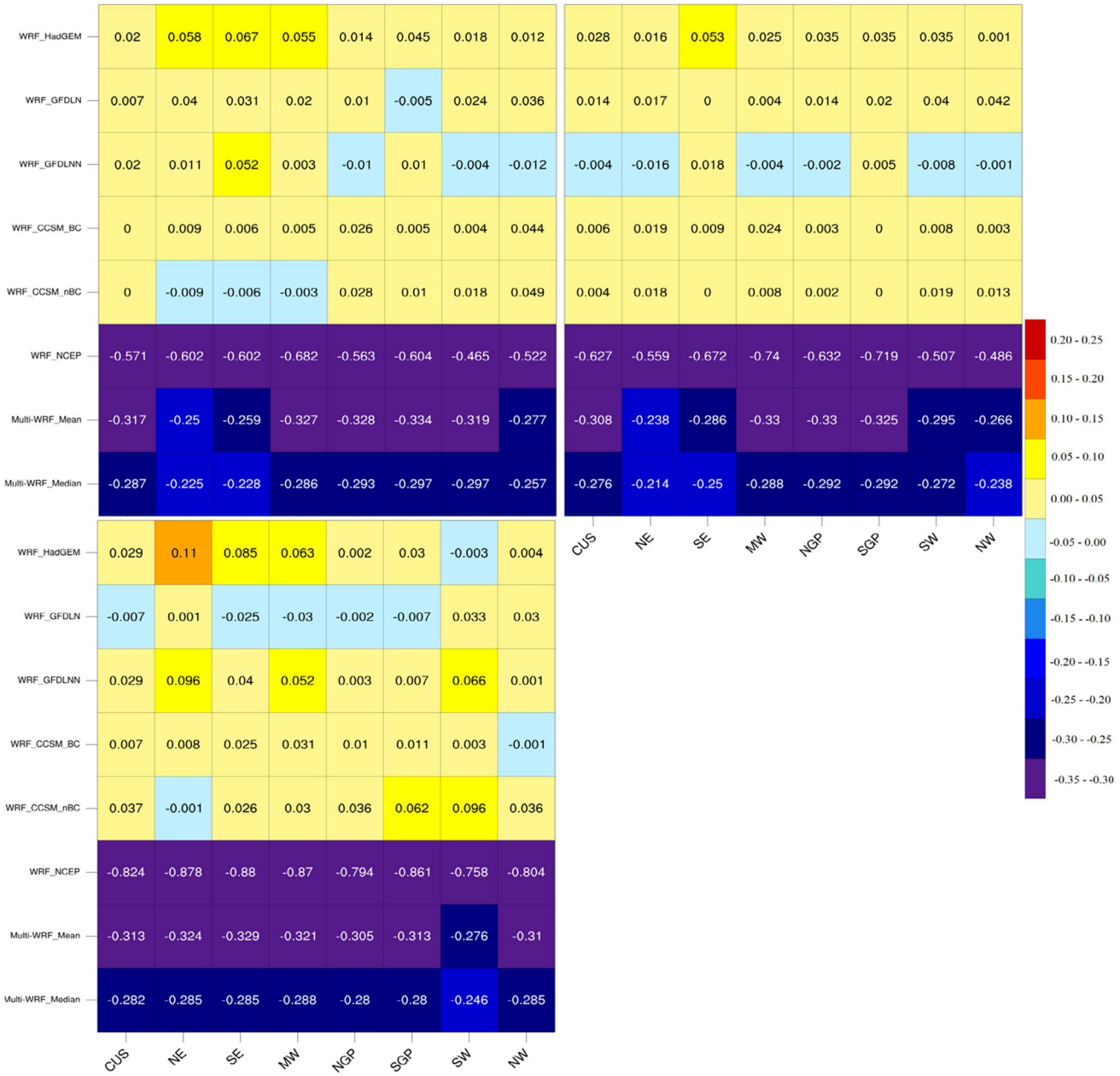

Fig. 4 Relative error of the RMSE values for: 10-m U wind (*upper left*), and 10-m V wind (*upper right*); and sea-level pressure (*lower left*)

WRF runs for 10 m-winds in several of the regions, especially the meridional wind (v component of the wind). The WGNN generally performs better than WRF_GFDLN for both the U and V component. For SLP, nudging significantly reduces the overall error and WGN shows less relative error than WGNN in all the regions except for the NW—where the difference is minor. Bias correction in the WRF simulations driven by CCSM4 does not result in lower RMSEs in all of the regions for SLP. Where WH ranks in terms of the other GCM simulations for Figs. 2, 3, and 4 is different depending on the variable and the region, but there is no discernable trend where/when it consistently outperforms its GCM counterparts for any region or variable. It would provide significant value if one could develop a single index to evaluate individual model performance considering all the variables of interest (Gleckler et al. 2008). This way one could consider more weight on the "better" model than the "worse" model when considering future climate projection. However, different model output is related to different aspect of model physics and/or model setup. Subjectively ranking the model performance based on an average score of all the variables of interest would

substantially hide model errors for some aspects (Gleckler et al. 2008). The following section will show that performing well in simulating overall climatology (i.e. relative error) does not necessarily mean the simulation is accurate in terms of model extremes.

3.3 Extreme event

Another primary motivation for dynamically downscaling climate models is to gain a more comprehensive idea of regional extreme events, and eventually, make more accurate predictions of frequency and intensity of future anomalous climate events. In this section, we discuss the model reliability at forecasting frequency and/or intensity of extreme warm/cold temperature, heat stress, both single and multi-day heavy precipitation events.

3.3.1 Extreme temperature

Temperature values that are located in the left/right tail of the PDF curve provides valuable information as to how the model simulation captures the extreme minimum/maximum temperature for a given location. In this study, we calculate the 95% threshold of summer (June, July, and August) maximum and the 5% threshold of winter (December, January, and February) minimum near surface temperature in the reference data and WRF simulations to judge the model's ability in capturing the extreme high and low temperatures. Figure 5 shows the differences in extreme high temperature between model simulations and the observations based on NCA subregion averages. To calculate these values, first, the 95% thresholds for each grid point in both observations and the simulations are calculated, second, the differences are found for each grid point, last, regional average of the differences are displayed. Comparing with the six individual model output, the mean and median show the largest bias (cold bias) in extreme high temperature from the reference data set in most of the regions because mean and median filters out the day-to-day variability at each location, which reduces the variance of the PDF curve and acts to smooth out the real extremes. The two CCSM4 driven WRF simulations (with and without bias correction) also underestimate the extreme maximum temperature for all seven climate regions, but show smaller cold bias than do mean and median. The use of bias correction (WCB) reduces the cold bias over the Northwest, Southwest as well as Southeast regions by 0.5–2 °C, but increase the bias over the Midwest and Northeast regions by 1 °C in comparison with the run without bias correction (WCNB). Overall, the GFDL driven simulations have warm bias over the Great Plains and Midwest regions and a smaller cold bias than the CCSM4 driven runs. In the two runs where WGN and WCB use both spectral nudging and bias correction, there are large differences in all seven regions, indicating that the GCM used to force the WRF makes a larger difference than the use of bias correction and nudging does for maximum temperature. This is especially true for the Midwest and the Southern Plains where the two runs not only disagree on sign, but the difference in magnitude in the 95% threshold is greater than a 3 °C between WGN and WCB. Spectral nudging does improve the model performance in extreme high temperature over most of the regions. For example, the regions NGP, SGP, MW, SE, and NE all have smaller bias in the WGN run by 0.36–1.53 °C than in the WGNN.

The HadGEM2 driven simulation underestimates the extreme maximum temperature for all of the regions other than the Midwest. WH's proximity to the observed value for the Northwest and Northern Plains is closer than both WCB and WCNB, but this is not the case when compared to the WGN and WGNN. In the Northeast and Southeast, WH performs much worse than the other GCM driven runs with the threshold being missed by at least 3.5 °C in these regions. For the Southeast, WH is actually closer to the error for the mean and median runs than the other GCM simulations. Overall, in the central part of the country (MW, NGP, and SGP), there are ambiguous signs between the five GCM runs. The GCM driven runs underestimate the intensity of the extreme maximum temperature events for both the eastern (NE and SE) and western regions (NW and SW). The NCEP-R2 driven WRF also underestimate the NW and SW regions thresholds by as much as 2.8 °C.

Figure 6 shows the differences in extreme low temperature between model simulations and the observations based on NCA subregion averages. Similar to the maximum temperature, the tail in the PDF curve for GCM driven simulations are too close to the mean and underestimate the intensity of extreme cold temperature (with warm bias) for many of the regions. The only region where all the models (except WN) are consistently too cold is the Northwest region. The WH model was the closest to the reference data set for this region when compared to the other GCM runs. Spectral nudging and bias correction affect the GFDL and CCSM4 driven runs differently. Spectral nudging shifts the threshold value in WGN to the right for all the regions, making the extreme cold temperature closer to the observation than WGNN over Northwest, but further from the observation than WGNN over other regions. Different from the effects of nudging, bias correction reduces the bias of extreme cold temperature by as much 2 °C in most of the NCA subregions, with the exception of NW and SW. The two simulations (WGN and WCB) that use both of bias correction and nudging have the same sign and similar magnitudes with the exception for the Northeast. This is different from Fig. 5, where the GCM boundary conditions play a much more significant role in the biases for extreme maximum temperatures between these two runs.

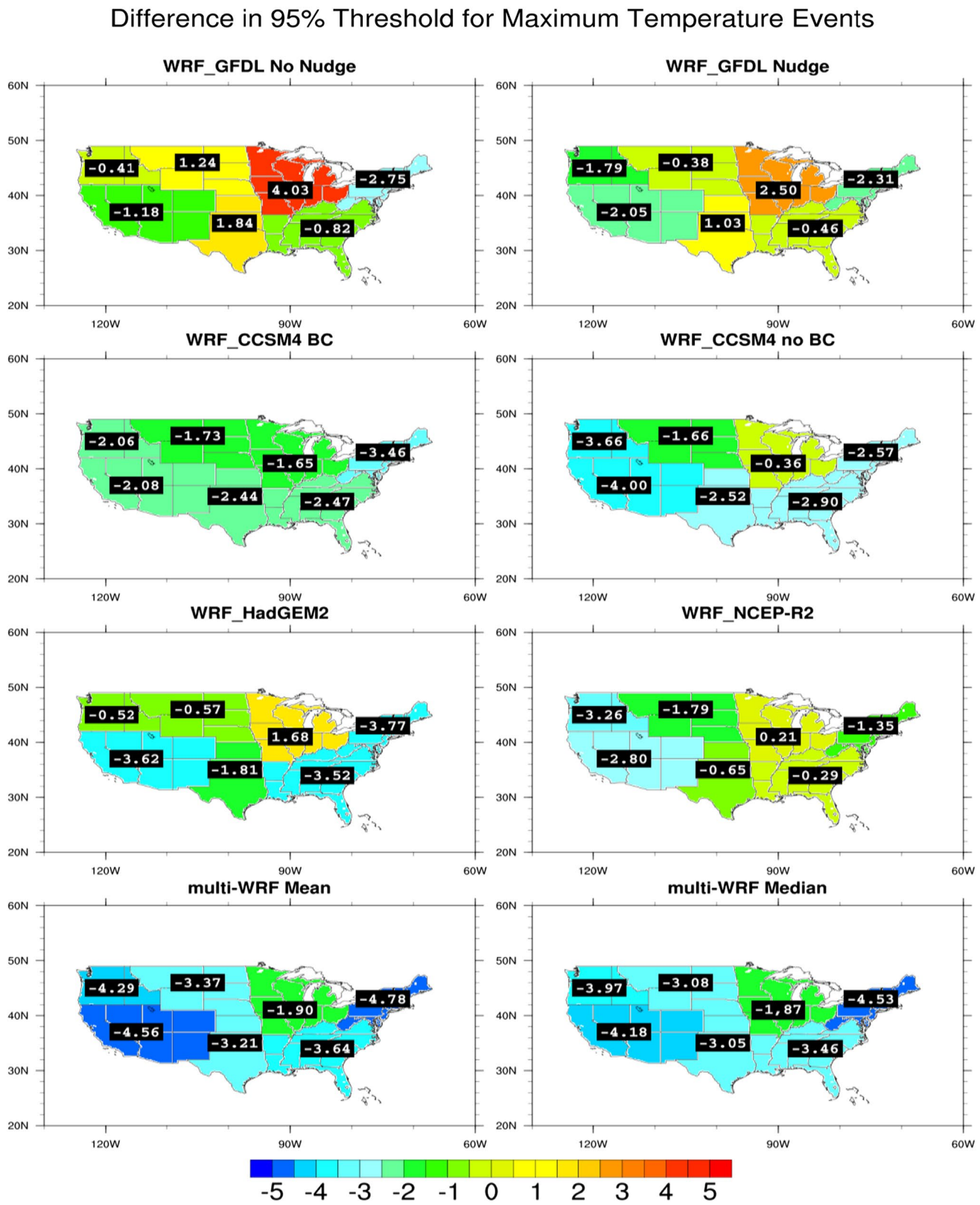

Fig. 5 Average subregional difference (model-observations) in 95% threshold of daily maximum summer (June, July, and August) temperature (°C)

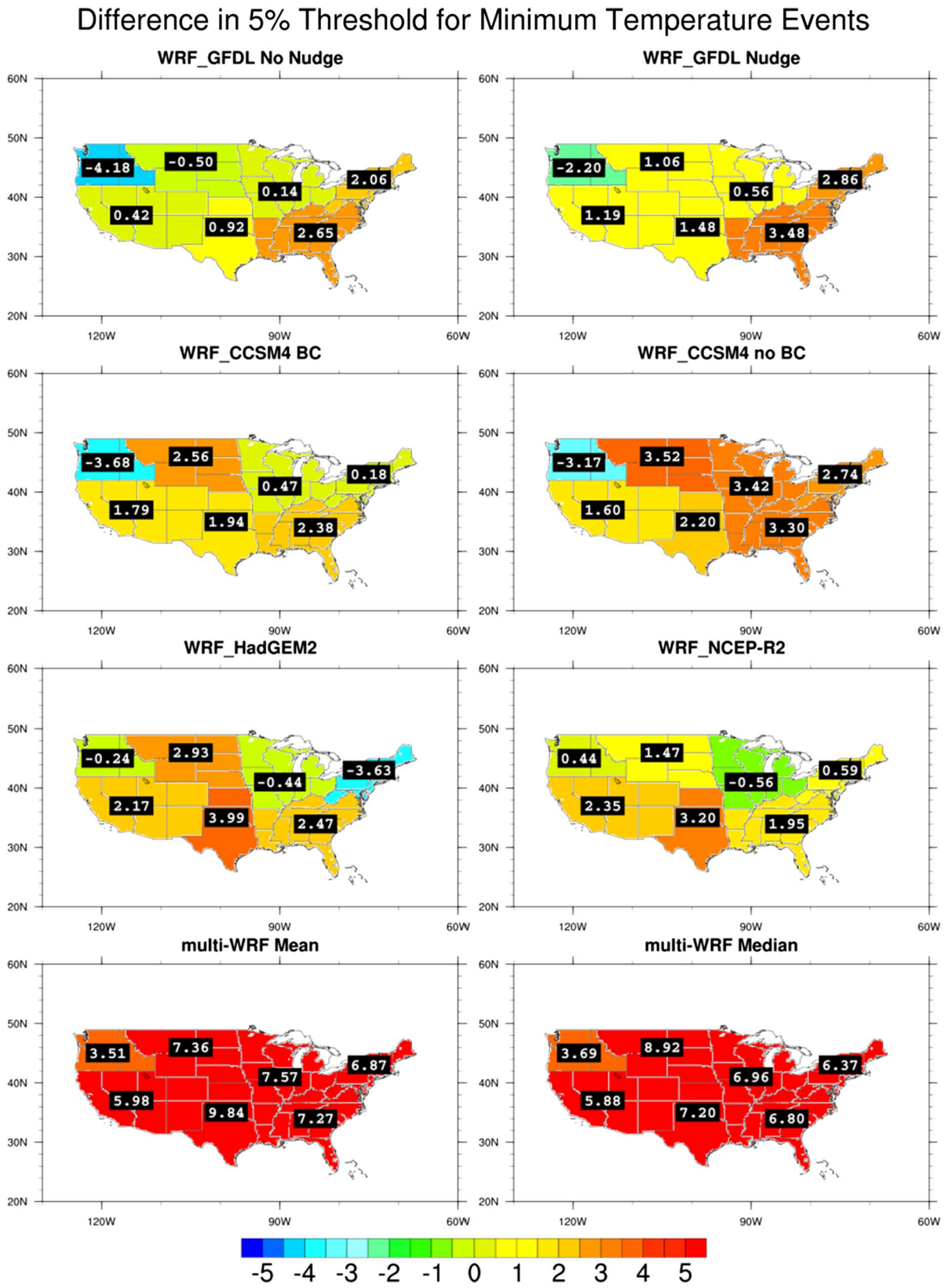

Fig. 6 Average subregional difference (model-observations) in 5% threshold of daily minimum winter (December, January, and February) temperature (°C)

3.3.2 Heat index

In addition to temperature, relative humidity (RH) plays an equally important role on the amount of stress the human body can endure in hot conditions. Thus, heat index (HI) was developed world-wide (Buzan et al. 2015). In this study, we apply one of the heat indices, which is developed by Rothfusz (1990) and using temperature and RH and is applied primarily by the National Weather Service in the United States. Rothfusz (1990) performed a multiple regression analysis on the original table of HI that was computed by Steadman (1979). However, the equation calculated by Rothfusz (1990) is not applicable for all ranges of RH and temperatures. An adjustment of the HI equation is needed (http://www.wpc.ncep.noaa.gov/html/heatindex_equation.shtml). Using the equation of Rothfusz (1990) and the adjustment mentioned above, we are able to compare how well the models capture the right tail of the PDF for HI. The HI value for each location is calculated for both the simulations and the observations, so that the difference can be plotted in gridded form. The maximum temperatures and near surface RH values are used to calculate the HI. Maximum temperature is used instead of mean temperature to help determine if the errors in HI are a result from the known extreme maximum temperature biases discussed for Fig. 5 or if RH biases could affect the results more significantly in some regions.

Figure 7 shows the difference in HI for each simulation's 95% threshold and the observations. Generally, the WRF_NCEP shows the smallest bias over the entire CONUS, followed by WCB. There is large positive bias for HI in the southern Plains and western CONUS for WGNN, WGN, and WCB that is not evident in the maximum temperature, indicating that the RH is overestimated in those regions. In comparison with WGNN, nudging reduce the bias for HI in the Northwest. In the Southeast, where HI values tend to be the highest during the summer, the models without bias correction underestimate the 95% threshold. In contrast, the models that use bias correction have a slight overestimation of HI and perform better over Southeast. Overall, there are significant differences in extreme HI values between the two GFDL runs when nudging is applied and the two CCSM4 runs when bias correction is used, especially over Great Plains and western part of the CONUS. When comparing the GFDL and CCSM4 runs that use both bias correction and spectral nudging, there are still a couple important differences. For example, in parts of the Midwest and central Plains, the differences in HI threshold are as high ~6 °C in some locations, indicating the biases in the boundary conditions are still significant in those areas.

3.3.3 Extreme precipitation

Figure 8 is the difference between model and observed 95th percentile of daily precipitation. All the precipitation data is filtered to only include precipitation days that record at least 0.01 inches or 0.254 mm to shield from minimum unrealistic values. In comparison with GCM or reanalysis driven WRF runs, the mean and median of the six WRF simulations show significant dry bias in extreme

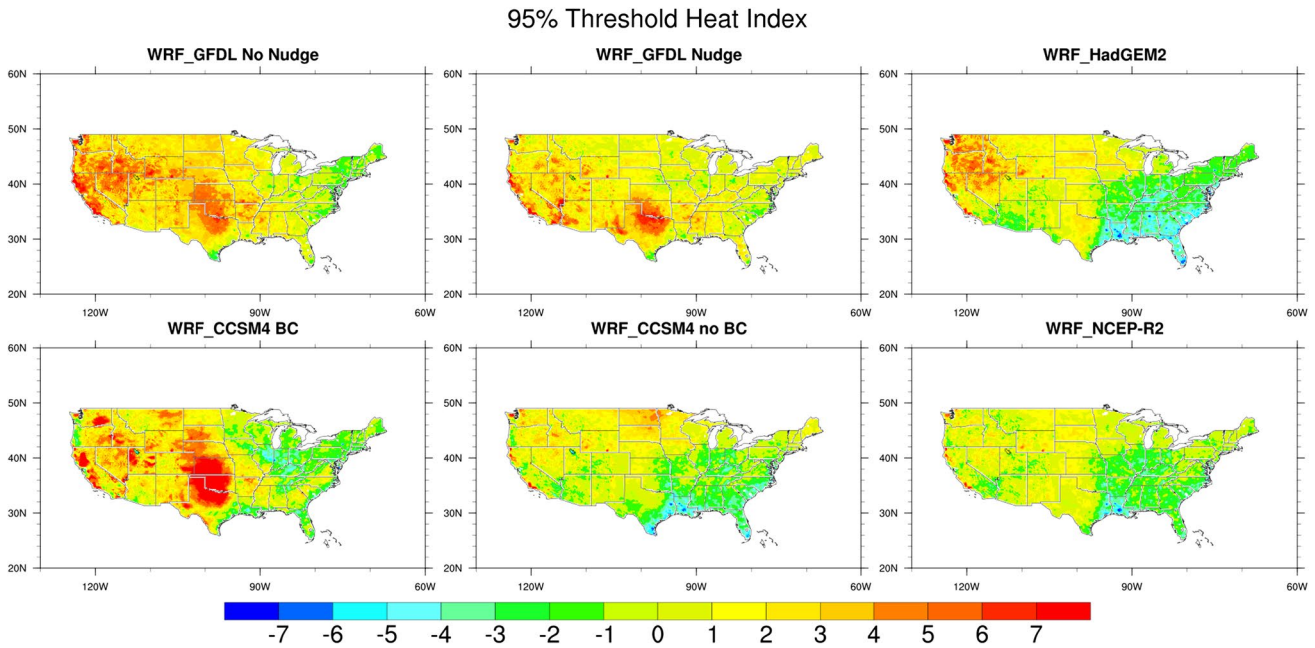

Fig. 7 The same as Fig. 5, but for heat index (unit: °C)

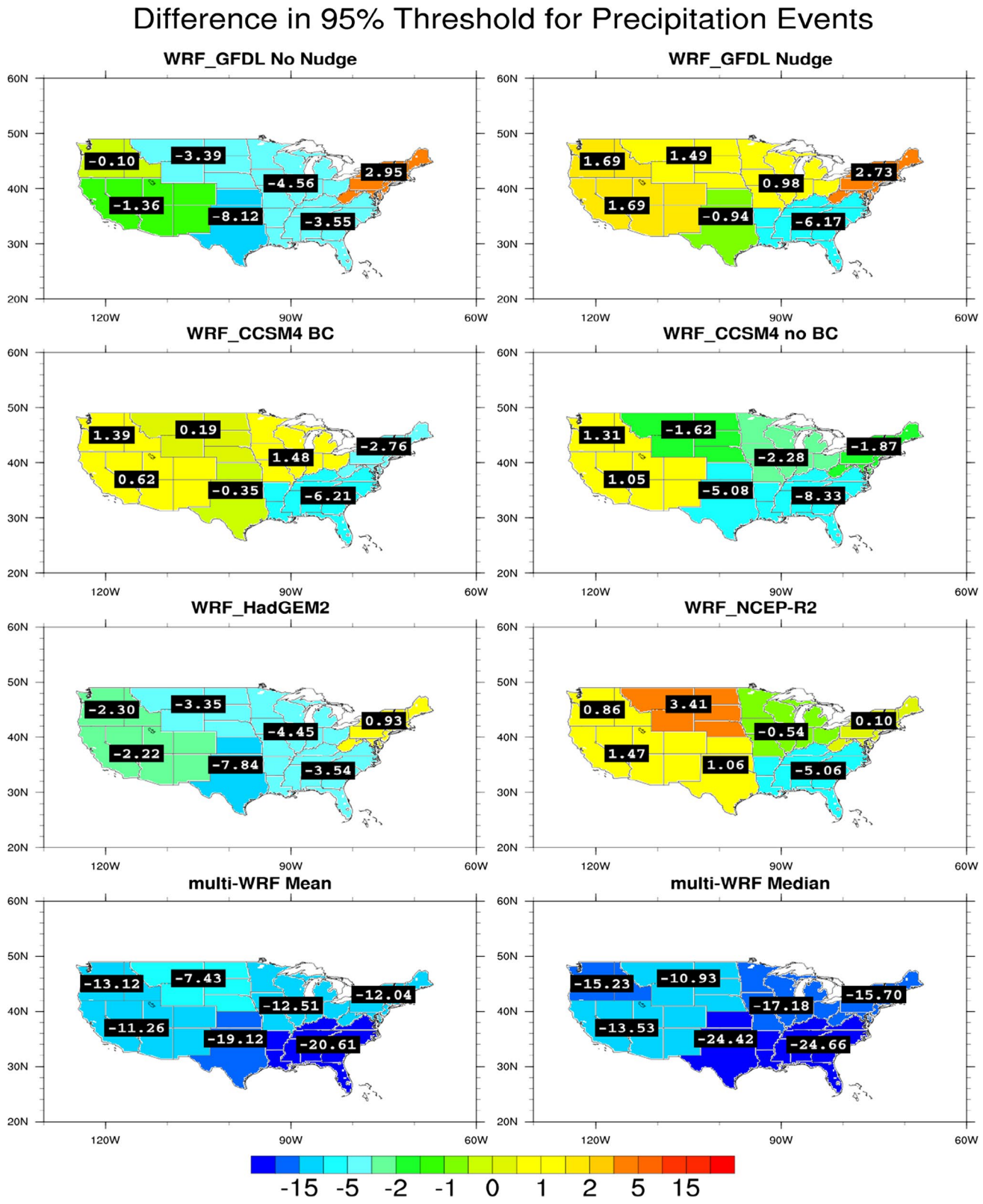

Fig. 8 The same as Fig. 5, but for precipitation (unit: mm)

precipitation. WN shows wet bias over Great Plains for not only the daily mean precipitation (as shown in Fig. 2 upper left), but also the extreme precipitation. The reason can be due to the short spin-up time and/or the strong nudging strength applied in this simulation (WK14), which has been modified in the GCM driven runs. The five GCM driven simulations underestimate the extreme precipitation by 3.5–8.3 mm over Southeast. This is likely because these simulations lack the ability to capture small scale convection that takes place regularly in this region that cannot be fully resolved with a 12-km horizontal resolution. In comparison with WGNN, WGN significantly reduces the model bias in extreme precipitation over Great Plains and Midwest. In comparison with WCNB, WCB significantly reduce the bias in extreme precipitation over all the regions except NW and NE. The WH run, which does not use bias correction or spectral nudging, shows much larger dry bias over most of the regions with the exception of the Northeast. It is worth mentioning that, knowing how the model performs in terms of relative error does not effectively forecast how the model predicts the extremes from the reference data. For example, the RMSE for the two models (WGN and WCBC) that use both bias correction and nudging is higher for the North Plains, South Plains, and Midwest regions, but they both show smaller bias in the same region for extreme precipitation than the other simulations.

Extreme precipitation events occur frequently when daily precipitation values are to the right of the 95% threshold in the PDF curve for multiple consecutive days (Janssen et al. 2014). In many cases, the heaviest precipitation events occur because a storm system is stagnant over similar areas for consecutive days (e.g. Francis and Vavrus 2012). While many other environmental factors determine the extent and magnitude of flash floods (Montz and Gruntfest 2002), the best these models can do is attempt to improve on forecasting frequency of long-term extreme precipitation events. For this reason, in addition to daily precipitation extremes, this study also analyzes the model's ability to simulate major precipitation events for 2 and 3-day storm totals. Figure 9 shows the differences in frequency of 99% threshold for 2-consecutive day precipitation. By finding the 99% average regional threshold for 2-day precipitation events from the observations, the difference in number of times the model predicts this occurring shows how well the simulation handles storm system movement across the US. This is calculated by ranking all the total 2-day precipitation events at each location that experienced at least a trace of precipitation and calculating the number of occurrences greater than the regional observed threshold for the whole decade in each of the six simulations. Figure 9 shows the number of times the model output was greater than the regionally averaged 99% threshold in the reference data and is standardized by subtracting the number of events in the reference data at each grid point that were greater than the 99% threshold. The reason difference is calculated is because, depending on the location or region, there may be a high frequency of precipitation days meaning it is expected that there would be more 99% events for these locations over the course of a decade. The 2-day and 3-day results for this metric are similar enough that only the difference in 2-day precipitation extremes is presented in this study.

The GCM driven simulations tend to underestimate the frequency of 99% events along the Gulf of Mexico in the Southeast and along the West Coast. Other regions, such as the Midwest, have differences in regional signs for each of the six simulations. Bias correction and nudging used together tend to slow storm system movement across the Plains and Midwest indicated by the positive 2-day precipitation positive anomalies. Without nudging or bias correcting the boundary conditions, the WRF simulations move storm systems across the central US. faster leading to fewer events that meet the observed 99% threshold criteria for that location. The addition of nudging in the GFDL runs enhances a strong positive bias in a large area of the Southwest as well as through most of the Northern Plain states that is not present in the no nudging run. The WGN run does reduce high negative bias in the WGNN in much of the Midwest. To a lesser extent, bias correction also reduces this same negative anomaly for the CCSM runs in most of the Midwest.

3.3.4 Large-scale circulations

To further understand the impacts of bias correction and spectral nudging on the simulated precipitation, in this section, we investigate the large-scale circulation to figure out the dominant physical reasons for the differences between WCB and WCNB as well as between WGN and WGNN. The regions that have the largest differences in extreme precipitation occurs in the Plains, Midwest, and Southeast. Since the warm season is when most of the extreme events take place in these regions because of the Great Plains Low-Level Jet (GPLLJ) and large scale flow (Hitchens et al. 2013), we focus on summer average 500 hpa geopotential height and 850 hpa V wind, or the north–south component of the wind vector, as shown in Figs. 10 and 11. The north–south winds that are usually important for large daily precipitation events in these regions because it not only allows MCSs to grow upscale during the summer due to the southerly flow, but also provides an overall source of moisture to these regions. The figures show the average anomaly when WRF simulation is subtracted from the observations (here we use NARR) at each grid point. As shown in Fig. 10, the 500 hPa geopotential height anomalies between the WGNN and WGN runs are quite different.

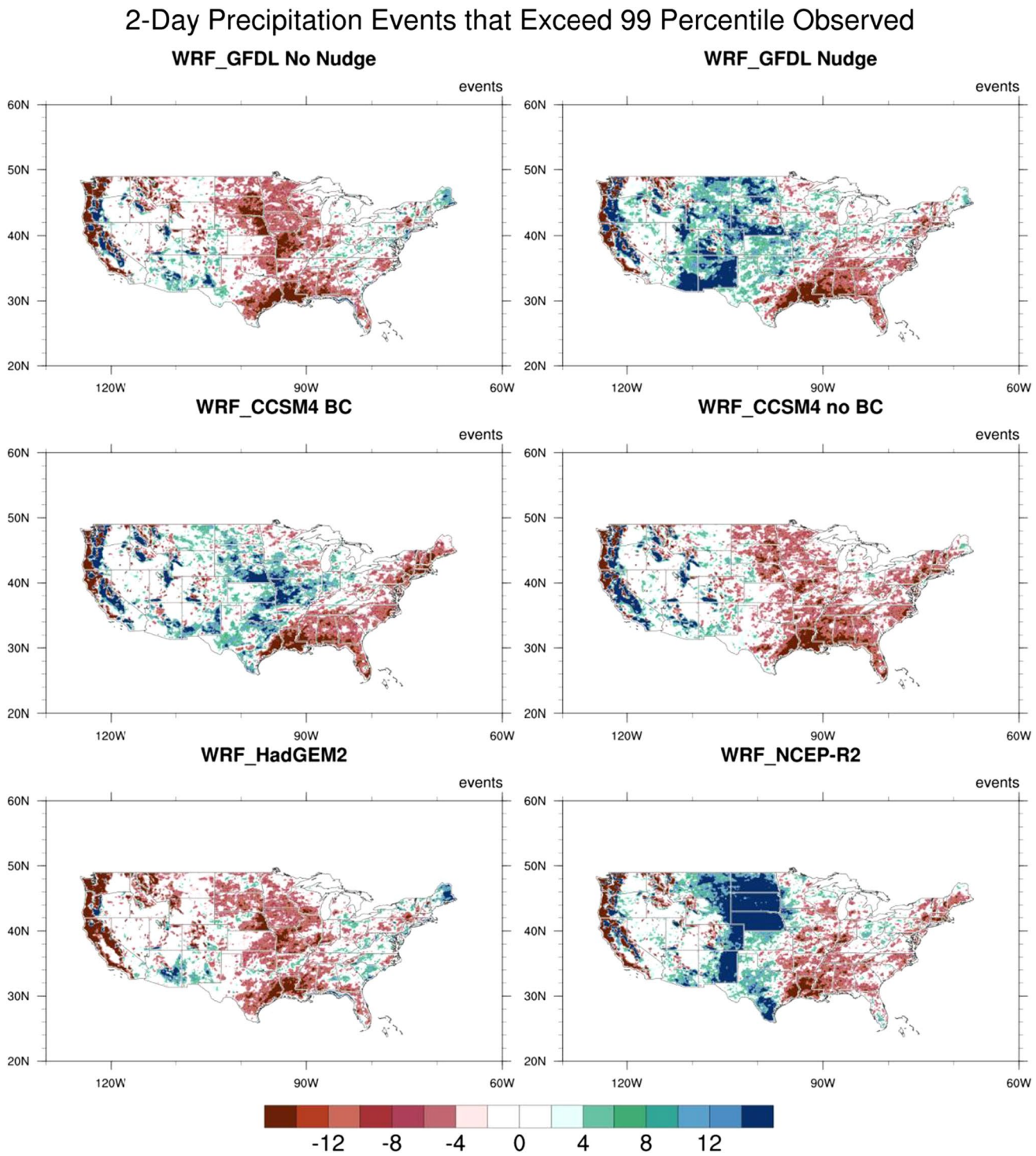

Fig. 9 Differences (model—observations) in the frequency of 99% threshold events for 2-day precipitation events. In order to be categorized as an "event" the grid point must experience at least a trace of precipitation for 2 consecutive days. To standardize these values, the difference between the number of 99% events in the observations is subtracted from the model values

WGNN produces significant ridging over much of the western third of the country and a trough centered most likely in southern Canada, but extending into the Northeast. The WGN run has an overall positive geopotential height bias over the country, but the strongest positive anomalies are located in the Southeast along the Georgia and Carolina coast. This strong positive anomaly in the WGN over the Southeast leads to the drier bias over Southeast as shown

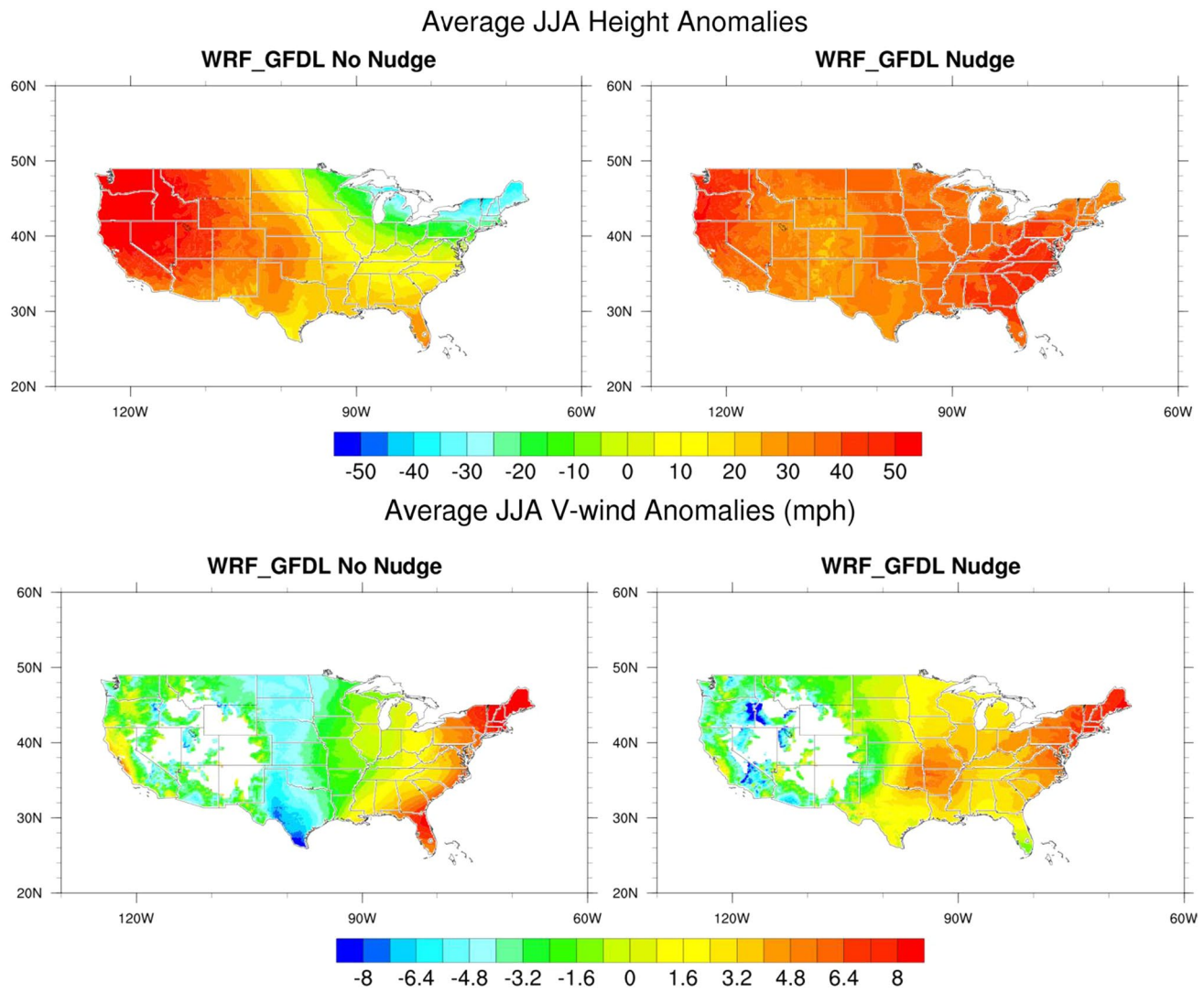

Fig. 10 Difference between model (WRF runs driven by GFDL with and without nudging) and observed 500 hPa geopotential height (in m; *upper panel*) and 850 hPa V-component of the wind (mph; *lower panel*) averaged across June, July, and August (JJA)

in Fig. 8. In the three central regions, the key difference is the large area of strong positive anomalies over the West Coast, Rocky Mountains, and extending into the western Plain states in the WGNN that is not as strong as that in WGN run. As shown in Fig. 10, in the WGNN simulation, the large negative anomaly, indicative of an overestimation of north wind strength, that persist in these regions is most certainly the primary driver of large negative anomalies for extreme precipitation events in the Plains and Midwest for this run. If the simulation has winds that move from north to south more than what is observed, like in the WGNN run, the air mass in place will tend to be drier Canadian air as opposed to the typical moist Gulf of Mexico air that would be the product of a southerly wind. In contrast, WGN captures the low level moisture transport across these regions more realistically than WGNN, with the exception of a positive anomaly across the Texas coast, and as a result is much close to the observed values.

Fiugre 11 shows a similar dependency of atmospheric circulation on extreme precipitation for CCSM4 driven WRF runs. While both runs have positive height anomalies across the Northwest and NPlains, WCNB extends the ridge further south and east into the SPlains. These positive height anomalies are indicative of low-level anticyclonic circulation that results in a north wind on the eastern side of this circulation in both simulations across the Plain states on the lee side of the Rockies. However, since the positive anomaly is shifted to the southwest in the WCNB, the northerly winds extend further to east. Equally important to those anomalies, the WCB has a positive height anomaly centered in the Gulf of Mexico that brings southerly flow of moist air into Texas and eventually into parts

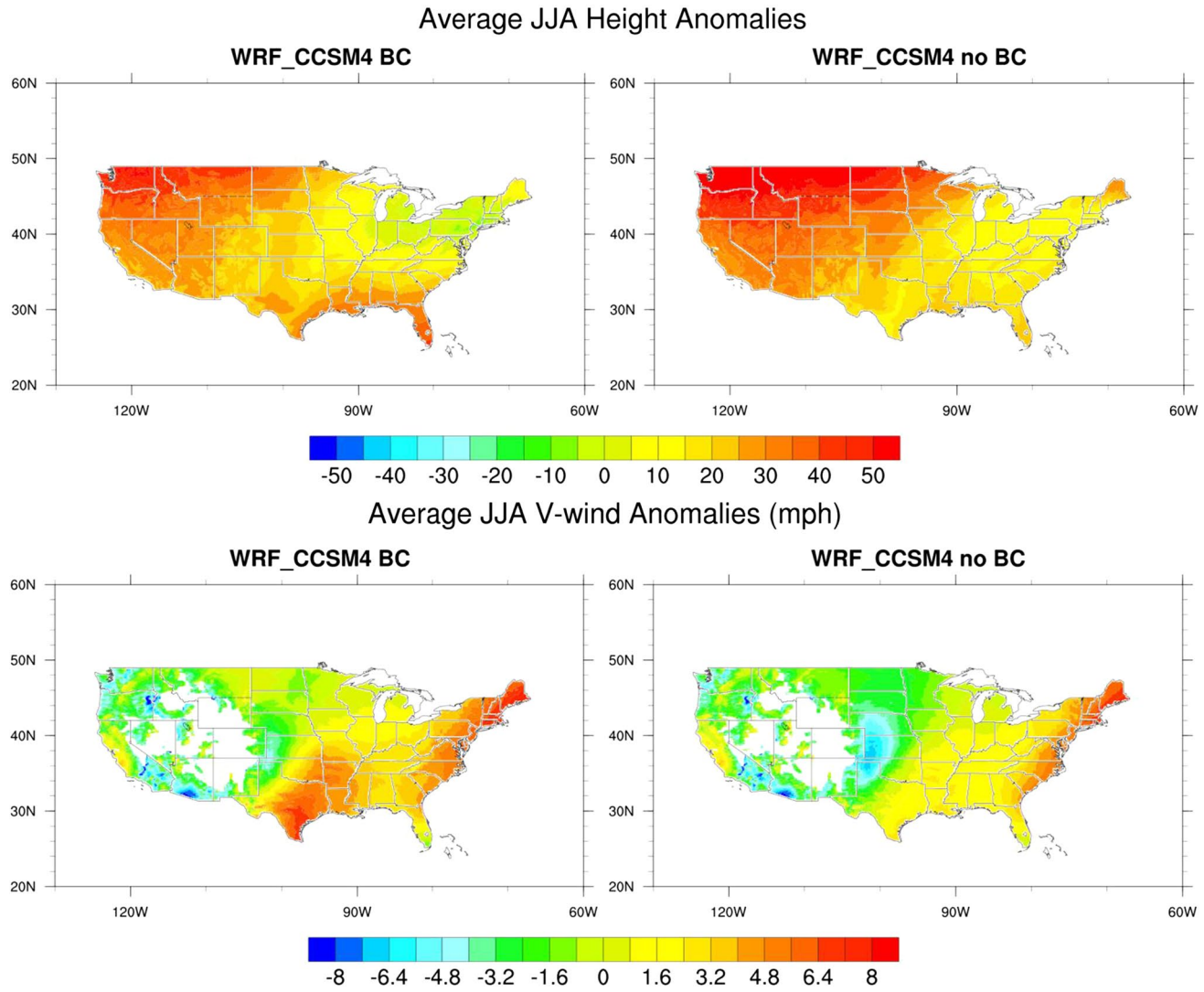

Fig. 11 The same as Fig. 9, expect for WRF runs driven by CCSM4 with and without bias correction

of the Midwest and Southeast. This is the reason that the SPlains and Southeast have larger dry bias for WCNB but smaller dry bias in WCB. In addition, because of the southerly flow which brings the moisture from Gulf of Mexico, WCB shows smaller bias than does WCNB which has significant dry bias over the Midwest and NPlains.

4 Discussion and summary

This study provides analysis for six WRF simulations by ranking their performance when evaluated based on measures of relative error and extreme climate events. Ranking the models based on relative error (Figs. 2, 3, 4) allows future researchers to make informed decisions on which type of boundary conditions and regional model settings are needed to achieve the most ideal results. Few downscaling projects have compared CMIP5 GCM-driven dynamical downscaled model performance for variables other than surface temperature and precipitation, especially as an ensemble (Fowler et al. 2007; Lee et al. 2014a, b). This study evaluates both lower and upper atmospheric variables and the dominant physics for extreme events, which can better inform researchers and users of the model results. The variables we study can aid in reconstructing dynamical profiles for the atmosphere to better understand their precipitation and temperature regional biases. Results show that when modeling climate extremes, the use of DD creates substantial "added value" or improved ability compared to low resolution GCM data. Model setup (i.e. bias correction and nudging) can be more important to predictability and biases than the GCM boundary conditions; A simulation that has low RMSE does not necessarily mean it is efficient at modeling extreme event frequency or

day-to-day fluctuations in these variables. In addition, our results show that many variables have the largest errors for surface variables in the wettest and driest regions of the continental United States. As mentioned above, high precipitation regions, such as the Southeast, yield higher errors because of the dominance of convective processes in these regions, which is challenging to predict at this resolution (Bryan et al. 2003). Similarly, drier regions have been shown to have greater errors or biases due to small scale processes that are hard to capture using downscaling techniques (Fowler et al. 2007).

One of the challenges in a study like this is to compare the model output to best reference data set available, but in reality, the "ground truth" for variables, such as precipitation, often have sources of biases and error themselves (Cosgrove et al. 2003). We compare the relative error when using PRSIM (Fig. 12, left panel) and NARR (Fig. 12, right panel) as the reference data sets for monthly precipitation. Overall, many of the regional ranks of the models are mostly similar between the two, but there are several cases where the difference in the magnitude of relative error is as high as 25% (e.g. Northeast region for GFDL using nudging and Southwest region for HadGEM2). Therefore, using multiple reference data sets that yield multiple results for errors for both relative error and extremes for a historical period, provides a more comprehensive understanding of the model performance. Understanding where the simulations fail, or do not closely match the observations, is the most important feature of this research and is vital to understanding future projections of climate extremes (Ekström et al. 2005). Similarly, the WRF model itself and the physical schemes used introduces an additional set of regional biases (e.g. Ruiz et al. 2010; Jankov et al. 2005; Ries and Schlünzen 2009; Cheng and Steenburgh 2005; Aligo et al. 2007). All of these studies discuss the importance of WRF configuration and that the ideal settings will have high temporal and geographical dependence based on the test variables. The large domain of this study evaluates regions with varying topography and climates making the choice in physical parameterizations for the WRF difficult despite the sensitivity experiments tested in WK14.

If there is a known overall bias in the dynamically downscaled method for a specific region in all members of the ensemble, that can now be accounted for when making projections of future climate change. As we mentioned above, each of these GCM's raw data have a different climate sensitivity. This is why the use of our ensemble could prove valuable at making analyses of uncertainties in projected extreme values. Since most of the uncertainty in future climate comes from choices such as the climate model used and the emission scenario (Déqué et al. 2007), our multi-climate model ensemble, while employing bias correction and spectral nudging, can prove valuable at analyzing the uncertainties in future climate extremes. In this study, we find the regions each metric most accurately is represented in the models. We show where both bias correction and spectral nudging can be beneficial using dynamical downscaling as well as which of the three GCMs tested have

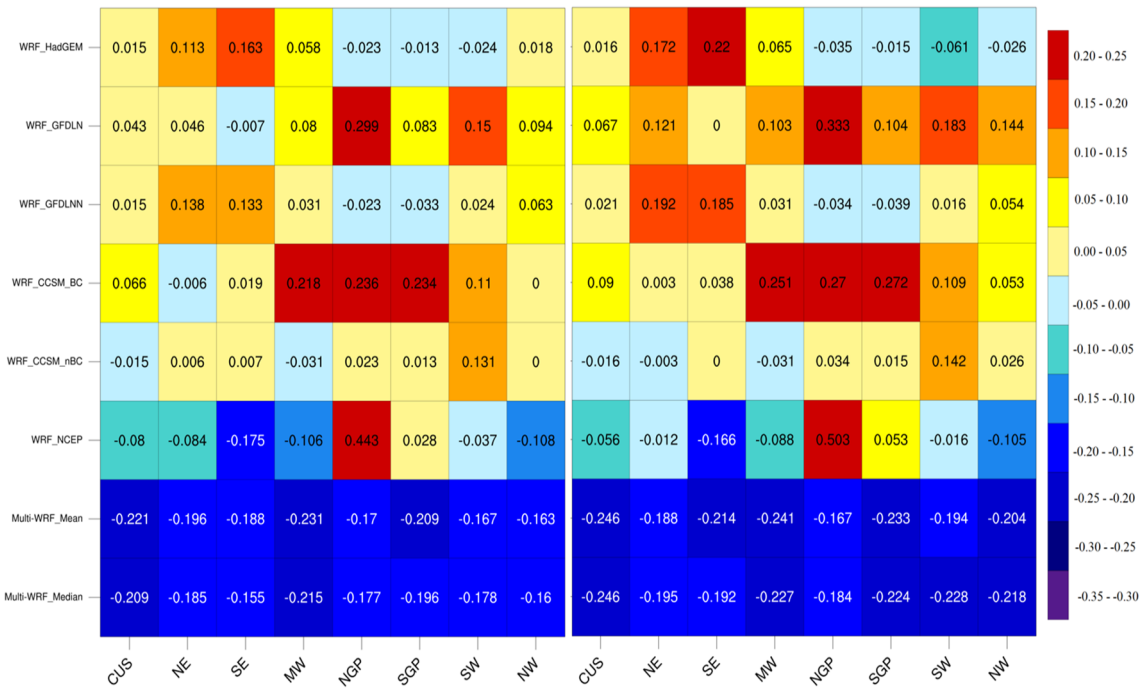

Fig. 12 RMSE figures for total monthly rainfall when compared to two reference data sets: **a** PRISM (*left*) **b** NARR (*right*)

model biases. The primary focus was on the shortcomings of the historical period in the model, but the information gathered from this analysis will be most useful in our future work on climate projections.

The high-resolution modeling studies provide stakeholders and the public with a knowledge of the uncertainties on a range of climate indicators, including assessing effect on local hydrological processes; surface temperature changes and heat stress on humans in a warmer climate (Fowler et al. 2007; Buzan et al. 2015). Understanding these strengths and weakness of using dynamic downscaling methods is an important step in finding a way to access the risks of future climate. These types of ensemble downscaling studies can provide an evaluation of future uncertainties in societal impacts at spatial scales of interest to the impact assessment and adaptation community (Fowler et al. 2007). Wang and Kotamarthi (2015) used two members of this ensemble to discuss precipitation changes over the continental United States by first comparing the model's historical accuracy. We intend to build on those results in future research by incorporating these 5 GCM-driven simulations and by discussing more dynamical features rather than focusing strictly on the model surface output. The next step that needs to be made in establishing the utility of using these predictions and the uncertainties by user community.

Acknowledgements This work is supported under a military interdepartmental purchase request from the Strategic Environmental Research and Development Program, RC-2242, through US. Department of Energy (DOE) contract DE-AC02-06CH11357. The NARR data (in NetCDF format) are provided by the NOAA-ESRL Physical Sciences Division, Boulder, Colorado, at http://www.esrl.noaa.gov/psd/. The GCM data are downloaded from https://www.earthsystemgrid.org/home.htm. Computational resources are provided by the DOE-supported Argonne Leadership Computing Facility, the National Energy Research Scientific Computing Center, and National Center for Supercomputing Applications Blue Waters Supercomputer. All the model outputs generated in this study will be available online.

References

Aligo EA, Gallus Jr WA, Segal M (2007) Summer rainfall forecast spread in an ensemble initialized with different soil moisture analyses. Weather Forecast 22:299–214. doi:10.1175/WAF995.1

Bowden JH, Talgo KD, Spero TL, Nolte CG (2016) Assessing the added value of dynamical downscaling using the standardized precipitation index. Adv Meteorol. doi:10.1155/2016/8432064

Bryan GH, Wyngaard JC, Fritsch JM (2003) Resolution requirements for the simulation of deep moist convection. Mon Weather Rev 131(10):2394–2416. doi:10.1175/MWR-D-11-00046.1

Bukovsky MS, Karoly DJ (2007) A brief evaluation of precipitation from the North American Regional Reanalysis. J Hydrometeorol 8:837–846. doi:10.1175/JHM595.1

Buzan JR, Oleson K, Huber M (2015) Implementation and comparison of a suite of heat stress metrics within the Community Land Model version 4.5. Geosci Model Dev 8:151–170. doi:10.5194/gmd-8-151-2015

Carter TR, Parry ML, Harasawa H, Nishioka S (1994) IPCC technical guidelines for assessing climate change impacts and adaptions. Cambridge University Press, Cambridge p 59

Chen F, Dudhia J (2001) Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: model implementation and sensitivity. Mon Weather Rev 129:569–585. doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2

Cheng WYY, Steenburgh WJ (2005) Evaluation of surface sensible weather forecasts by the WRF and the Eta models over the western United States. Weather Forecast 20:812–821. doi:10.1175/WAF885.1

Christensen NS, Wood AW, Voisin N, Lettenmaier DP, Palmer RN (2004) The effects of climate change on the hydrology and water resources of the Colorado River basin. Clim Change 62:337–363

Christensen JH, Boberg F, Christensen OB, Lucas-Picher P (2008) On the need for bias correction of regional climate change projections of temperature and precipitation. Geophys Res Lett 35(20) doi:10.1029/2008GL035694

Christensen JH, Kjellström E, Giorgi F, Lenderink G, Rummukainen M (2010) Assigning relative weights to regional climate models: exploring the concept. Clim Res 44:179–194. doi:10.3354/cr00916

Cosgrove BA et al (2003) Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. J Geophys Res Atmos 108. doi:10.1029/2002JD003118

Daly C, Neilson RP, Phillips DL (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. J Appl Meteorol 33:140–158. doi:10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2

Daly C, Taylor GH, Gibson WP (1997) The PRISM approach to mapping precipitation and temperature. In Proc. 10th AMS Conf. on Applied Climatology (pp. 20–23)

Déqué M et al (2007) An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections. Clim Change 81:53–70

Di Luca A, de Elía R, Laprise R (2012) Potential for added value in precipitation simulated by high-resolution nested regional climate models and observations. Clim Dyn 38:1229–1247. doi:10.1007/s00382-011-1068-3

Doherty SJ et al (2009) Lessons learned from IPCC AR4: scientific developments needed to understand, predict, and respond to climate change. Bull Am Meteorol Soc 90:497

Donner LJ et al (2011) The dynamical core, physical parameterizations, and basic simulation characteristics of the atmospheric component AM3 of the GFDL global coupled model CM3. J Clim 24:3484–3519

Ekström M, Fowler HJ, Kilsby CG, Jones PD (2005) New estimates of future changes in extreme rainfall across the UK using regional climate model integrations. 2. Future estimates and use in impact studies. J Hydrol 300:234–251. doi:10.1016/j.jhydrol.2004.06.017

Fowler HJ, Blenkinsop S, Tebaldi C (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int J Climatol 27:1547–1578. doi:10.1002/joc.1556

Francis JA, Vavrus SJ (2012) Evidence linking Arctic amplification to extreme weather in mid—latitudes. Geophys Res Lett 39. doi:10.1029/2012GL051000

Gao Y, Fu JS, Drake JB, Liu Y, Lamarque JF (2012) Projected changes of extreme weather events in the Eastern United States based on a high-resolution climate modeling system. Environ Res Lett 7. doi:10.1088/17489326/7/4/044025

Gent PR et al (2011) The community climate system model version 4. J Clim 24:4973–4991. doi:10.1175/2011JCLI4083.1

Gimeno L, Dominguez F, Nieto R, Trigo R, Drumond A, Reason CJ, Taschetto AS, Ramos AM, Kumar R, Marengo J (2016) Major mechanisms of atmospheric moisture transport and their roles in extreme precipitation events. Annu Rev Environ Res 41. doi:10.1146/annurev-environ-110615-085558

Giorgi F (1990) Simulation of regional climate using a limited area model nested in a general circulation model. J Clim 3:941–963

Giorgi F, Bi X (2000) A study of internal variability of a regional climate model. J Geophys Res Atmos 105:29503–29521. doi:10.1029/2000JD900269

Giorgi F, Jones C, Asrar GR (2009) Addressing climate information needs at the regional level: the CORDEX framework. World Meteorol Organ (WMO) Bull 58(3):175

Giorgi F, Mearns LO (1991) Approaches to the simulation of regional climate change: a review. Rev Geophys 29:191–216. doi:10.1029/90RG02636

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. J Geophys Res Atmos 113. doi:10.1029/2007JD008972

Goodess CM et al (2007) An intercomparison of statistical downscaling methods for Europe and European regions–assessing their performance with respect to extreme temperature and precipitation events. Clim Change (**in press**)

Grell GA, Dévényi D (2002) A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. Geophys Res Lett 29. doi:10.1029/2002GL015311

Halmstad A, Najafi MR, Moradkhani H (2013) Analysis of precipitation extremes with the assessment of regional climate models over the Willamette River Basin, USA. Hydrol Process 27:2579–2590. doi:10.1002/hyp.9376

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25:1965–1978. doi:10.1002/joc.1276

Hitchens NM, Brooks HE, Schumacher RS (2013) Spatial and temporal characteristics of heavy hourly rainfall in the United States. Mon Weather Rev 141:4564–4575. doi:10.1175/MWR-D-12-00297.1

Hong SY, Kanamitsu M (2014) Dynamical downscaling: fundamental issues from an NWP point of view and recommendations. Asia Pac J Atmos Sci 50:83–104. doi:10.1007/s13143-014-0029-2

Hong SY, Lim JOJ (2006) The WRF single-moment 6-class microphysics scheme (WSM6). J Korean Meteorol Soc 42:129–151

Hong SY, Noh Y, Dudhia J (2006) A new vertical diffusion package with an explicit treatment of entrainment processes. Mon Weather Rev 134:2318–2341

Huth R (1999) Statistical downscaling in central. Eur Eval Methods Potential Predict Clim Res 13:91–101. doi:10.3354/cr013091

Iacono MJ, Delamere JS, Mlawer EJ, Shephard MW, Clough SA, Collins WD (2008) Radiative forcing by long—lived greenhouse gases: calculations with the AER radiative transfer models. J Geophys Res Atmos 113. doi:10.1029/2008JD009944

Jankov I, Gallus WA, Segal M, Shaw B, Koch SE (2005) The impact of different WRF model physical parameterizations and their interactions on warm season MCS rainfall. Weather Forecast 20:1048–1060. doi:10.1175/WAF888.1

Janssen E, Wuebbles DJ, Kunkel KE, Olsen SC, Goodman A (2014) Observational—and model—based trends and projections of extreme precipitation over the contiguous United States. Earth's Futur 2:99–113. doi:10.1002/2013EF000185

Jones CD et al (2011) The HadGEM2-ES implementation of CMIP5 centennial simulations. Geosci Model Dev 4:543–570. doi:10.5194/gmd-4-543-2011

Gutowski Jr et al (2010) Regional extreme monthly precipitation simulated by NARCCAP RCMs. J Hydrometeorol 11:1373–1379. doi:10.1175/2010JHM1297.1

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77:437–471

Kleinn J, Frei C, Gurtz J, Lüthi D, Vidale PL, Schär C (2005) Hydrologic simulations in the Rhine basin driven by a regional climate model. J Geophys Res Atmos 110. doi:10.1029/2004JD005143

Knutti R, Masson D, Gettelman A (2013) Climate model genealogy: generation CMIP5 and how we got there. Geophys Res Lett 40:1194–1199. doi:10.1002/grl.50256

Lee JW, Hong SY (2014) Potential for added value to downscaled climate extremes over Korea by increased resolution of a regional climate model. Theor Appl Climatol 117:667–677. doi:10.1007/s00704-013-1034-6.

Lee JW, Hong SY, Chang EC, Suh MS, Kang HS (2014a) Assessment of future climate change over East Asia due to the RCP scenarios downscaled by GRIMs-RMP. Clim Dyn 42:733–747. doi:10.1007/s00382-013-1841-6

Lee JW, Ham S, Hong SY, Yoshimura K, Joh M (2014b) Future changes in surface runoff over Korea projected by a regional climate model under A1B scenario. Adv Meteorol 2014

Leung LR, Mearns LO, Giorgi F, Wilby RL (2003) Regional climate research: needs and opportunities. Bull Am Meteorol Soc 84:89–95. doi:10.1175/BAMS-84-1-89

Liu C, Ikeda K, Thompson G, Rasmussen R, Dudhia J (2011) High-resolution simulations of wintertime precipitation in the Colorado headwaters region: Sensitivity to physics parameterizations. Mon Weather Rev 139:3533–3553. doi:10.1175/MWR-D-11-00009.1

Liu P, Tsimpidi AP, Hu Y, Stone B, Russell AG, Nenes A (2012) Differences between downscaling with spectral and grid nudging using WRF. Atmos Chem Phys 12:3601–3610. doi:10.5194/acp-12-3601-2012

Loikith PC, Lintner BR, Kim J, Lee H, Neelin JD, Waliser DE (2013) Classifying reanalysis surface temperature probability density functions (PDFs) over North America with cluster analysis. Geophys Res Lett 40:3710–3714. doi:10.1002/grl.50688

Maurer EP, Hidalgo HG (2008) Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. Hydrol Earth Syst Sci 12:551–563. doi:10.5194/hess-12-551-2008

Maurer EP, Wood AW, Adam JC, Lettenmaier DP, Nijssen B (2002) A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States*. J Clim 15:3237–3251. doi:10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2

Mearns LO et al (2012) The North American regional climate change assessment program: overview of phase I results. Bull Am Meteorol Soc 93(9):1337–1362. doi:10.1175/BAMS-D-11-00223.1

Mearns LO, Giorgi F, McDaniel L, Shields C (1995) Analysis of daily variability of precipitation in a nested regional climate model: comparison with observations and doubled CO2 results. Global and Planet Change 10(1–4):55–78. doi:10.1016/0921-8181(94)00020-E

Melillo JM et al (2014) Climate change impacts in the United States: the third national climate assessment. US Glob Change Res Progr. doi:10.7930/J0Z31WJ2

Mesinger F et al (2006) North American regional reanalysis. Bull Am Meteorol Soc 87:343–360. doi:10.1175/BAMS-87-3-343

Montz BE, Gruntfest E (2002) Flash flood mitigation: recommendations for research and applications. Glob Environ Change Part B:Environ Hazards 4:15–22. doi:10.1016/S1464-2867(02)00011-6

Morrison H, Thompson G, Tatarskii V (2009) Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one-and two-moment schemes. Mon Weather Rev 137:991–1007. doi:10.1175/2008MWR2556.1

Noh Y, Cheon WG, Hong SY, Raasch S (2003) Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. Bound Layer Meteorol 107:401–427

Otte TL, Nolte CG, Otte MJ, Bowden JH (2012) Does nudging squelch the extremes in regional climate modeling? J Clim 25:7046–7066. doi:10.1175/JCLI-D-12-00048.1

Pryor SC, Nikulin G, Jones C (2012) Influence of spatial resolution on regional climate model derived wind climates. J Geophys Res 117:D03117. doi:10.1029/2011JD016822

Randall DA et al (2007) Climate models and their evaluation. In: Solomon S et al (eds) Climate change 2007: the physical science basis. Cambridge University Press, Cambridge, pp 589–662

Ries H, Schlünzen KH (2009) Evaluation of a mesoscale model with different surface parameterizations and vertical resolutions for the Bay of Valencia. Mon Wea Rev 137:2646–2661. doi:10.1175/2009MWR2836.1

Rothfusz L (1990) The heat index (or, more than you ever wanted to know about heat index), technical attachment SR 90–23. Scientific Services Division, National Weather Service, Fort Worth

Ruiz JJ, Saulo C, Nogués-Paegle J (2010) WRF model sensitivity to choice of parameterization over South America: validation against surface variables. Mon Weather Rev 138:3342–3355. doi:10.1175/2010MWR3358.1

Sheffield J et al (2013a) North American climate in CMIP5 experiments. Part II: evaluation of historical simulations of intraseasonal to decadal variability. J Clim 26:9247–9290. doi:10.1175/JCLI-D-12-00593.1

Sheffield J et al (2013b) North American climate in CMIP5 experiments. Part I: evaluation of historical simulations of continental and regional climatology*. J Clim 26:9209–9245. doi:10.1175/JCLI-D-12-00592.1

Shepard DS (1984) 'Computer mapping: the SYMAP interpolation algorithm'. In: Gaile GL, Willmott CJ (eds). Spatial statistics and models, Dordrecht pp 133–145

Sherwood SC, Bony S, Dufresne J-L (2014) Spread in model climate sensitivity traced to atmospheric convective mixing. Nature 505:37–42. doi:10.1038/nature12829

Skamarock WC et al (2008) A description of the advanced research WRF version 3. NCAR technical note NCAR/TN-475 + STR. doi:10.5065/D68S4MVH. http://www2.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf

Steadman RG (1979) The assessment of sultriness. Part I: a temperature-humidity index based on human physiology and clothing science. J Appl Meteorol 18:861–873. doi:10.1175/1520-0450(1979)018<0861:TAOSPI>2.0.CO;2

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res Atmos 106:7183–7192. doi:10.1029/2000JD900719

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485

Tripathi OP, Dominguez F (2013) Effects of spatial resolution in the simulation of daily and subdaily precipitation in the southwestern US. J Geophys Res Atmos 118:7591–7605. doi:10.1002/jgrd.50590

Van der Linden P, Mitchell JE (2009) ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, UK p 160

Vautard R et al (2013) The simulation of European heat waves from an ensemble of regional climate models within the EURO-CORDEX project. Clim Dyn 41:2555–2575. doi:10.1007/s00382-013-1714-z

von Storch H, Zorita E, Cubasch U (1993) Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. J Clim 6:1161–1171. doi:10.1175/1520-0442(1993)006<1161:DOGCCE>2.0.CO;2

Wang J, Kotamarthi VR (2013) Assessment of dynamical downscaling in near-surface fields with different spectral nudging approaches using the nested regional climate model (NRCM). J Appl Meteorol Climatol 52:1576–1591. doi:10.1007/s00382-016-3000-3

Wang J, Kotamarthi VR (2015) High—resolution dynamically downscaled projections of precipitation in the mid and late 21st century over North America. Earth's Futur 3:268–288. doi:10.1002/2015EF000304

Wang J, Swati FNU, Stein ML, Kotamarthi VR (2015) Model performance in spatiotemporal patterns of precipitation: new methods for identifying value added by a regional climate model. J Geophys Res Atmos 120:1239–1259. doi:10.1002/2014JD022434

Wang J, Kotamarthi VR (2014) Downscaling with a nested regional climate model in near-surface fields over the contiguous United States. Journal of Geophys Res: Atmosp 119(14):8778–8797. doi:10.1002/2014JD021696

Wang J, Han Y, Stein ML, Kotamarthi VR, Huang WK (2016) Evaluation of dynamically downscaled extreme temperature using a spatially-aggregated generalized extreme value (GEV) model. Clime Dyn 1–17

Wehner MF (2013) Very extreme seasonal precipitation in the NARCCAP ensemble: model performance and projections. Clim Dyn 40:59–80. doi:10.1007/s00382-012-1393-1

Widmann M, Bretherton CS (2000) Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States. J Clim 13:1936–1950. doi:10.1175/1520-0442(2000)013<1936:VOMPIT>2.0.CO;2

Wigley TML, Jones PD, Briffa KR, Smith G (1990) Obtaining sub—grid—scale information from coarse—resolution general circulation model output. J Geophys Res Atmos 95:1943–1953. doi:10.1029/JD095iD02p01943

Wilby RL, Wigley TML (1997) Downscaling general circulation model output: a review of methods and limitations. Prog Phys Geogr 21:530–548

Wilby R, Greenfield B, Glenny C (1994) A coupled synoptic-hydrological model for climate change impact assessment. J Hydrol 153:265–290. doi:10.1016/0022-1694(94)90195-3

Wood AW, Leung LR, Sridhar V, Lettenmaier DP (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Clim Change 62:189–216. doi:10.1023/B:CLIM.0000013685.99609.9e

Xue Y, Janjic Z, Dudhia J, Vasic R, De Sales F (2014) A review on regional dynamical downscaling in intraseasonal to seasonal simulation/prediction and major factors that affect downscaling ability. Atmos Res 147–148:68–85. doi:10.1016/j.atmosres.2014.05.001

Yongkang X et al (2014) A review on regional dynamical downscaling in intraseasonal to seasonal simulation/prediction and major factors that affect downscaling ability. Atmos Res 147:68–85. doi:10.1016/j.atmosres.2014.05.001