Development of a Model Performance Metric and Its Application to Assess Summer Precipitation over the U.S. Great Plains in Downscaled Climate Simulations

BAIJUN TIAN,^a HUIKYO LEE,^a DUANE E. WALISER,^a ROBERT FERRARO,^a JINWON KIM,^b
JONATHAN CASE,^c TAKAMICHI IGUCHI,^{d,e} ERIC KEMP,^{e,f} DI WU,^{e,f}
WILLIAM PUTMAN,^e AND WEILE WANG^{g,h}

^a Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California

^b Joint Institute for Regional Earth System Science and Engineering, University of California, Los Angeles, Los Angeles, California

^c ENSCO, Inc., NASA Marshall Space Flight Center, Huntsville, Alabama

^d Earth System Science Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland

^e NASA Goddard Space Flight Center, Greenbelt, Maryland

^f Science Systems and Applications, Inc., Lanham, Maryland

^g California State University, Monterey Bay, Seaside, California

^h NASA Ames Research Center, Moffett Field, California

(Manuscript received 17 March 2017, in final form 8 July 2017)

ABSTRACT

Several dynamically downscaled climate simulations with various spatial resolutions (24, 12, and 4 km) and spectral nudging strengths (0, 600, and 2000 km) have been run over the contiguous United States from 2000 to 2009 using the high-resolution NASA Unified Weather and Research Forecasting (NU-WRF) regional model initialized and constrained by the Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). This paper summarizes the authors' efforts on the development of a model performance metric and its application to assess summer precipitation over the U.S. Great Plains (USGP) in these downscaled climate simulations. A new model performance metric T was first developed that uses both the linear correlation coefficient and mean square error and is consistent with other commonly used metrics, but gives a bigger separation between good and bad simulations. This metric T was then applied to the summer mean precipitation spatial pattern, diurnal Hovmöller diagram, and diurnal spatial pattern over the USGP from the simulations focusing on the summer precipitation diurnal cycle related to mesoscale convective systems (MCSs). The metric T skill scores increase significantly from the control simulation to the nudged simulations and from the nudged simulations with shorter wavelengths to the nudged simulations with longer wavelengths, but do not change much from MERRA-2 to the downscaled simulations or between the various downscaled simulations with different spatial resolutions. Thus, there is some credibility, but no significant value added compared to MERRA-2, of the downscaled climate simulations of the summer precipitation over the USGP.

1. Introduction

The summer precipitation over the U.S. Great Plains (USGP; 35°–45°N, 110°–90°W) exhibits a nocturnal diurnal peak, in stark contrast to the afternoon rainfall maximum over most other continental U.S. (CONUS) regions (e.g., Wallace 1975; Riley et al. 1987; Dai et al. 1999; Tian et al. 2005). The summer nocturnal

precipitation peak over the USGP is mainly associated with the diurnally eastward propagating mesoscale convective systems (MCSs) originating over the Rocky Mountains in the afternoon (e.g., Carbone et al. 2002; Tian et al. 2005; Jiang et al. 2006; Carbone and Tuttle 2008; Chang et al. 2016; Feng et al. 2016). On the other hand, the afternoon rainfall maximum over other CONUS regions is mainly the result of isolated convective events (e.g., Nesbitt and Zipser 2003; Tian et al. 2005; Jiang et al. 2006). Over the USGP, the MCSs account for 30%–70% of the total warm-season precipitation while isolated convective events are responsible for the remaining portion of the precipitation (Fritsch et al. 1986; Jiang et al. 2006; Nesbitt et al. 2006; Chen et al. 2009).

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-17-0045.s1>.

Corresponding author: Baijun Tian, baijun.tian@jpl.nasa.gov

Unfortunately, most global climate models (GCMs; e.g., Zhang 2003; Klein et al. 2006; Lee et al. 2007a) and global numerical weather prediction (NWP) models (e.g., Knievel et al. 2004) fail to capture this pronounced nocturnal rainfall peak over the USGP and the eastward propagating MCSs from the Rockies to the USGP on a daily time scale. Even some regional climate models (RCMs; e.g., Dai et al. 1999; Liang et al. 2004; Gao et al. 2017) and regional NWP models (e.g., Davis et al. 2003) cannot capture this nocturnal rainfall peak over the USGP. One possible reason for this failure is that the coarse-grid GCMs and RCMs lack the spatial and temporal resolutions to properly represent the dynamics of such MCSs (e.g., Clark et al. 2007; Lee et al. 2007b; Sun et al. 2016; Gao et al. 2017). Thus, we strive to investigate whether dynamically downscaled, high-resolution GCMs and RCMs can capture the summer nocturnal precipitation peak related to these eastward propagating MCSs over the USGP.

In the context of climate modeling, “dynamic downscaling” refers to the practice of driving an RCM with initial and boundary conditions derived from a previously executed low-resolution GCM simulation or assimilation data, affording potential advantages of higher spatial and temporal resolutions and/or a more comprehensive treatment of physical processes and topography given the savings in computational resources by only focusing on a specific region of Earth (Dickinson et al. 1989; Giorgi and Bates 1989; Giorgi 1990). The dynamic downscaling is designed to produce more detailed climate simulations at regional scales. This information is in high demand for regional stakeholders and decision-makers for assessing the local impacts of global climate change.

There has been a growth in using dynamic downscaling for regional climate assessment and impact studies during the past decade, such as the North American Regional Climate Change Assessment Program (NARCCAP; Mearns et al. 2012; 2013) and Coordinated Regional Climate Downscaling Experiment (CORDEX; Giorgi and Gutowski 2015). However, there are still some debates regarding the best methodologies and modeling frameworks used for downscaling as well as the credibility of the process in general (e.g., Pielke and Wilby 2012; Hall 2014).

Increasing model resolution has been found to have the largest positive impacts on climate simulations in regions of complex terrain where increased grid resolution notably improves simulations of climate processes that are strongly influenced by orography (e.g., Leung and Qian 2003; Leung et al. 2003; Leung et al. 2004; Hughes and Hall 2010; Bacmeister et al. 2014). However, increasing model resolution does not always improve model skill, such as simulating the

precipitation diurnal cycle (e.g., Dirmeyer et al. 2012; Jin et al. 2016).

Recently, the National Aeronautics and Space Administration (NASA) initiated a multicenter joint project to assess the credibility and value added of dynamically downscaled climate simulations (Ferraro et al. 2017). As a proxy for a prognostic climate forecast model, and so that ground truth in the form of satellite and in situ observations could be used for evaluation, the NASA Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2), reanalysis was used to drive and initialize the high-resolution NASA Unified Weather and Research Forecasting (NU-WRF) regional model. A number of experiments were conducted with various spatial resolutions and nudging strengths for simulations over the CONUS during the period 2000–09. The results of these experiments were compared to observational datasets to evaluate the model simulations. The evaluation analyses focused on the models’ ability to reproduce three high-impact phenomena affecting the CONUS using model performance metrics: atmospheric rivers on the west coast (Kim et al. 2017), MCSs in the central United States (this paper), and northeast winter storms in the mid-Atlantic and New England states (Loikith et al. 2017). The objective of this paper is to summarize our work on the development and application of performance metrics to assess the credibility and value added of the dynamically downscaled climate simulations of summer precipitation, especially that related to the MCSs, over the USGP and its dependencies on spatial resolution and spectral nudging strength.

The structure of this paper is organized as follows. We describe the models and various model run configurations of this downscaling project in section 2, the observational reference datasets that were used in section 3, a new generic performance metric for climate model evaluation we developed in section 4, and the application of this generic performance metric to the summer precipitation over the USGP for various model simulations in section 5. Summary and discussion are given in section 6.

2. Model descriptions and experimental designs

Dynamical downscaling is potentially attractive to evaluate regional impacts from future climate predicted by GCMs. However, since future climate predictions have no “truth” to use in an evaluation, for this NASA dynamic downscaling project, we constructed an experiment scenario in which an RCM is driven by a reanalysis over a historical period that includes observations as the ground truth. Here the reanalysis serves as the proxy for the coarse-grid GCM. Since a reanalysis is

expected to be the best possible representation of the atmospheric state at GCM resolutions, it avoids the problem of a free-running model drifting away from the historical weather patterns (which would make a comparison to observations not useful). These hindcast model simulations and boundary conditions were chosen so that in situ and satellite-based observation products could be used to evaluate the model simulations, to understand how much fidelity one should expect from the dynamic downscaling, and to test the effects of horizontal resolution and spectral nudging strength on regional model performance against available observational reference datasets.

The NU-WRF (Peters-Lidard et al. 2015) was used as the RCM for this experiment. NU-WRF has been developed at NASA's Goddard Space Flight Center (GSFC), in collaboration with NASA's Marshall Space Flight Center (MSFC) and university partners, as an observation-driven integrated modeling system that represents aerosol, cloud, precipitation, and land processes at satellite-resolved scales $O(1\text{--}25)$ km, thereby bridging the continuum between local (microscale), regional (mesoscale), and global (synoptic) processes (Peters-Lidard et al. 2015). NU-WRF was built upon the National Center for Atmospheric Research (NCAR)'s Advanced Research WRF (WRF-ARW) dynamical core model (Skamarock et al. 2005), with multiple additional modules developed at NASA GSFC. The present study employed a special version of NU-WRF, which was based on the WRF-ARW version 3.5.1 and included a bug fix that removed the accumulation of round-off errors in lateral boundary conditions for better long-term simulations (Dudhia 2015). The cumulus parameterization schemes employed in WRF-ARW version 3.5.1 are the Grell 3D ensemble deep convection scheme (G3D; Grell and Freitas 2014) and the University of Washington shallow cumulus scheme (UWSC; Bretherton et al. 2004; Park and Bretherton 2009). For the planetary boundary layer and subgrid-scale turbulence, the level 2.5 Mellor–Yamada–Janjić turbulence scheme (MYJ; Janjić 1990, 1994; Janjić 2002) was chosen, as this scheme has a long reliable history in the National Weather Service operational models (e.g., Eta and North American Mesoscale models). The corresponding Monin–Obukhov–Janjić Eta surface scheme was required when running the MYJ turbulence scheme. The GSFC modules include the GSFC Land Information System (LIS; Kumar et al. 2006; Peters-Lidard et al. 2007); the GSFC shortwave and longwave radiation schemes (Chou and Suarez 1999; Chou et al. 2001); and the GSFC single-moment, 3-ice bulk microphysics schemes including revised couplings to the aerosols (Tao et al. 2003; Lang et al. 2007, 2011). The GSFC LIS model

was used to run multiyear offline spinups of the land surface model prior to coupled NU-WRF initialization to improve coarsely resolved initial soil conditions obtained from reanalysis data alone. The Noah land surface model (Chen and Dudhia 2001; Ek et al. 2003) was then employed to calculate the land state from the initialization. The model physics setup of the NU-WRF is shown in Table 1, and readers can refer to Peters-Lidard et al. (2015) for more details.

The NU-WRF's running domain for the experiment is shown in Fig. 1 and covers the whole CONUS and nearby oceans and lands. The NU-WRF regional model simulations were done with three separate horizontal grid resolutions: 24, 12, and 4 km. We denote the model runs associated with these spatial grid resolutions as “B24,” “B12,” and “B4,” respectively. Both the B24 and B12 NU-WRF model runs are for 10 years from 2000 to 2009, while the B4 NU-WRF model run is only for 5 years from 2000 to 2004 due to its enormous demand on computational resources. The cumulus parameterization schemes were used in all model simulations, including the B4 model run.

The boundary and initial conditions of the NU-WRF model runs were provided by the NASA MERRA-2 reanalysis (Bosilovich et al. 2015), which has a longitudinal resolution of 0.625° and a latitudinal resolution of 0.5° . In addition, the NU-WRF was further constrained at every time step by MERRA-2 through spectral nudging (Miguez-Macho et al. 2004) for the calculation of the horizontal wind velocities, temperature, and geopotential heights above the boundary layer. Three nudging strengths were used in the experiments to test the effect of nudging on the downscaled results: a minimal spectral wavelength of 0 km (control, no nudging), 600 km, and 2000 km, in both east–west and north–south directions. The 2000-km nudging is a configuration that is similar to the WRF default and also similar to the 2500 km used by Miguez-Macho et al. (2004). According to Vukicevic and Errico (1990), waves of about 2500 km and larger scale have the most error growth in limited area models. The 600 km is consistent with the wavelength for constraint in the Goddard Earth Observing System Model, version 5 (GEOS-5) replay simulation (M2R12K, below). We denote the model runs associated with these spectral nudging configurations as “Con,” “N600,” and “N2000,” respectively. Both the B24 and B12 NU-WRF model runs have these three spectral nudging configurations, but the B4 NU-WRF model run used only the “N600” nudging. This nudging strength was chosen because computational resource limitations prevented us from conducting B4 experiments across all of the chosen nudging values and to be consistent with the spectral nudging used in the M2R12K simulation (below).

TABLE 1. NU-WRF model physics configuration.

Model framework	NU-WRF_v7_dscale_haswell_p1 (v7-lis7-3.5.1)
Initial and boundary conditions	MERRA-2
Land surface spinup	LIS 10-yr run from 1 Jan 1990 to 1 Nov 1999
Cloud microphysics	GCE 3ICE with graupel
Radiation	Goddard 2011
PBL	Mellor–Yamada–Janjić
Land surface model	Noah LSM
Cumulus parameterization	Grell-3D with native shallow cumulus component and UWSC
Nudging	Spectral nudging

The GEOS-5 replay capability was used to produce a 15-yr downscaled global reanalysis product at ~ 12.5 -km global resolution using a nonhydrostatic version of the GEOS-5 atmospheric model. This high-resolution version of GEOS-5 was nudged to the recent MERRA-2 reanalysis produced by the GEOS-5 atmospheric data assimilation system (ADAS). This downscaled global simulation is referred to as the MERRA-2 Replay at 12.5 km (M2R12K). The GEOS-5 physics includes relaxed Arakawa–Schubert convection scheme (Moorthi and Suarez 1992), the large-scale condensation scheme of Bacmeister et al. (2006), the Goddard shortwave/longwave radiation schemes (Chou and Suarez 1999; Chou et al. 2001), Lock turbulence scheme (Lock et al. 2000), and Koster et al. (2000) land surface scheme. While the MERRA-2 is produced at ~ 50 -km global resolution, a spectral analysis of the kinetic energy in MERRA-2 and M2R12K show that the MERRA-2 forcing stops adding value to the simulation below a spectral truncation of effectively T60 (~ 666 km). Thus, all increments from MERRA-2 to the M2R12K simulation

are filtered below T60, permitting the replay to closely follow the trajectory of the large scales (effectively T60) in the underlying MERRA-2 reanalysis, while allowing the M2R12K downscaling model to embed its own, internally developed, mesoscale organization. This allows us to directly compare the simulated mesoscale with observations for specific synoptic situations.

For this project, we used the GEOS-5 M2R12K replay run output covering the period from December 1999 through October 2010 over the CONUS. This study includes the analysis of the M2R12K results that overlapped in space and time with the NU-WRF experiments to see if downscaling using a global approach and model can produce significantly different results compared to using a regional model. These NU-WRF and GEOS-5 model run configurations are summarized in Table 2, and more details can be found in Ferraro et al. (2017).

3. Reference datasets

Two observational reference datasets for the surface precipitation were used in this study. The first is the 10-yr (2000–09) Tropical Rainfall Measurement Mission (TRMM) Multisatellite Precipitation Analysis (TMPA) 3B42 V7 precipitation product (Huffman et al. 2007). The TRMM TMPA provides a calibration-based sequential scheme for combining precipitation estimates from multiple satellites, as well as gauge analyses where feasible, at fine scales ($0.25^\circ \times 0.25^\circ$ and 3 hourly). The dataset covers the latitude band 50°N – 50°S for the period from 1998 to the delayed present.

The second is the 8-yr (2002–09) National Centers for Environmental Prediction (NCEP) Stage IV (ST4) precipitation analysis. The ST4 precipitation analysis is a near-real-time product based on the regional hourly/6-hourly multisensor (radar plus gauges) precipitation

FIG. 1. The domain map for the NU-WRF model runs.

TABLE 2. NU-WRF and GEOS-5 model run configurations.

Acronym	Description
M2R12K	GEOS-5 Replay, global, 12.5 km, hourly, 600-km spectral nudging
B24_Con	Control run at 24-km spatial resolution
B24_N600	Nudging run at 24-km spatial and 600-km spectral resolutions
B24_N2000	Nudging run at 24-km spatial and 2000-km spectral resolutions
B12_Con	Control run at 12-km spatial resolution
B12_N600	Nudging run at 12-km spatial and 600-km spectral resolutions
B12_N2000	Nudging run at 12-km spatial and 2000-km spectral resolutions
B4_N600	Nudging run at 4-km spatial and 600-km spectral resolutions

analyses and is generated at NCEP separately from the National Weather Service (NWS) Precipitation Processing System and the NWS River Forecast Center rainfall processing. These data are mosaicked into a national product of 4-km grid spacing (on a polar-stereographic grid) and are available for hourly, 6-hourly, and daily accumulation intervals since January 2002 (Lin and Mitchell 2005; Prat and Nelson 2015; Nelson et al. 2016). The ST4 precipitation product has an overall good agreement with other surface observations, although it tends to underestimate annual and seasonal means (Prat and Nelson 2015). Over the Rocky Mountains, it likely underestimates precipitation due to the lack of good gauge coverage and mountain blockage of radar beams.

To calculate the metric values related to the surface precipitation, we regridded all reference and model data into the TRMM spatial and temporal grids (3-hourly, $0.25^\circ \times 0.25^\circ$ longitude–latitude grids). Although this methodology does not test any spatial accuracy of the higher-resolution downscaled results, it does allow for testing the possibility that higher spatial resolution more accurately reproduces the aggregate precipitation and diurnal variability observed in the data.

4. A generic performance metric development and demonstration

For climate model performance evaluation, some diagnostics (e.g., spatial maps and distributions or time series) are typically employed to show how a model is similar or different from the reference and to lend understanding as to why a model exhibits a good/bad performance. However, to quantify the model performance, a scalar similar to some fundamental statistical measures [e.g., mean error or bias, mean square error (MSE) or

root-mean-square error, linear correlation coefficient, or variance] is typically used and referred to as a metric. Consistent with this nomenclature, we expect metrics to provide symptoms of problems, but to be less informative than diagnostics for illuminating their causes. However, metrics can be derived from diagnostics, generally resulting in a condensation of the original information of the diagnostics.

There are many diagnostics that could be applied to the downscaled results, which would certainly point to issues with the details of the simulated results and would be highly valuable in improving the RCM performance. However, our goal here is to look at metrics that arguably capture the climatology of the regional phenomena and can be applied to the downscaled results for objective comparison. These metrics must also be based on observations available at high enough resolution to capture the local variability.

Various performance metrics for climate model evaluations that rely on the MSE and/or the linear correlation coefficient have been proposed previously (Table 3; e.g., Taylor 2001; Murphy et al. 2004; Gleckler et al. 2008; Reichler and Kim 2008; Hirota et al. 2011; Watterson et al. 2014). These previous metrics have mainly relied on either the linear correlation coefficient or the MSE individually instead of the combination of both. For example, the performance metrics from Taylor (2001) and Hirota et al. (2011) have mainly relied on the linear correlation coefficient, while the performance metrics from Gleckler et al. (2008), Murphy et al. (2004), Reichler and Kim (2008), and Watterson et al. (2014) have mainly relied on the MSE.

Suppose f is a climate model simulated field and r is a corresponding reference field at N discrete points (in time and/or space). Let R denote the linear correlation coefficient between f and r in both space and time, and it is defined as

$$R = \sigma_{fr} / \sigma_f \sigma_r. \quad (1)$$

Let MSE denote the mean square error between f and r in both space and time, and it is defined as

$$\text{MSE} = \sum_{i=1}^N (f_i - r_i)^2 / N = \text{Bias}^2 + \sigma_f^2 + \sigma_r^2 - 2\sigma_{fr}. \quad (2)$$

Here, $\bar{f} = \sum_{i=1}^N f_i / N$, $\bar{r} = \sum_{i=1}^N r_i / N$, $\sigma_f^2 = \sum_{i=1}^N (f_i - \bar{f})^2 / N$, $\sigma_r^2 = \sum_{i=1}^N (r_i - \bar{r})^2 / N$, and $\sigma_{fr} = \sum_{i=1}^N (f_i - \bar{f})(r_i - \bar{r}) / N$ are the mean, variance, and covariance of (between) f and r , respectively, and $\text{Bias} = \bar{f} - \bar{r}$ is the bias between f and r .

The larger (or closer to 1) R is, the more similar these two fields f and r are in terms of structure and phase. However, from a large R alone it is impossible

TABLE 3. Previously proposed climate model performance metrics.

Study	Climate model metric
Taylor (2001)	Skill score $S = 4(1 + R)/(a + 1/a)^2/(1 + R_0)$, f is a simulated field, and r is a corresponding reference dataset; R is the correlation coefficient between f and r in both space and time, a is the ratio of the standard deviations of f and r (σ_f/σ_r), and R_0 is the maximum correlation attainable.
Hirota et al. (2011)	Skill score $S = (1 + R)^4/(a + 1/a)^2/4$
Murphy et al. (2004)	Climate Prediction Index: $\text{CPI}^2 = \text{MSE}/V(f)$. The CPI^2 for each model is the MSE between f and r , divided by the interannual variance V of f , and averaged in desired regions and over 32 climate variables using equal weight.
Gleckler et al. (2008)	Root-mean-square difference or error (RMSE) of various variables between f and r in both space and time.
Reichler and Kim (2008)	The performance metric for each model is the gridpoint difference between f and r , normalized by the observed interannual variance, scaled by the average error found in a reference ensemble of models, and averaged globally and over 14 climate variables using equal weight.
Watterson et al. (2014)	Arcsin Mielke measure $M = (2/\pi)\arcsin(1 - \text{MSE}/\{V(f) + V(r) + [G(f) - G(r)]^2\}) \times 1000$; V is the variance and G is the mean; M is the MSE of f and r , nondimensionalized by their spatial variances of f and r .

to determine whether these two fields have the same amplitude or not (Watterson 1996). For example, in Fig. 2, f_1 and f_2 have the exact same spatial structure and their R is 1, but the magnitude of f_2 is twice as big as the magnitude of f_1 . In addition, from a small R (close to 0) alone it is impossible to determine how much of the difference is due to the difference in structure/phase or amplitude. For example, in Fig. 2, the R between f_1 and f_3 and the R between f_2 and f_3 are both 0.25, but the small R between f_1 and f_3 is due mainly to the difference in structure while the small R between f_2 and f_3 is due to the difference in both structure and magnitude.

The MSE approaches 0 as these two fields become more alike in magnitude. However, from a small MSE alone it is impossible to determine whether these two fields have the same structure or not. For example, in Fig. 2, the MSE between f_1 and f_3 is very small (0.04) due to their very similar magnitudes, but they have very different spatial structures. In addition, from a large MSE alone it is impossible to determine how much of the error is due to the difference in structure/phase or amplitude. For example, in Fig. 2, the MSE between f_1 and f_2 and the MSE between f_2 and f_3 are both large (0.15 and 0.31), but the large MSE between f_1 and f_2 is due mainly to the difference in magnitude while the large MSE between f_2 and f_3 is due to the difference in both structure and magnitude.

From the above discussion, we can see that both R and MSE provide important and complementary statistical information quantifying the correspondence between two fields, but neither is complete individually. Thus, Taylor (2001) has proposed a diagram consisting of both R and MSE of f and r , along with their variances, to provide a concise statistical summary of how well f and r match each other. Accordingly, we propose a new metric T that is a combination of the previous metrics

from Taylor (2001), Hirota et al. (2011), and Watterson et al. (2014) and uses both R and MSE:

$$T = [(1 + R)/2] \times [1 - \text{MSE}/(\text{Bias}^2 + \sigma_f^2 + \sigma_r^2)]. \quad (3)$$

For any given R except -1 , the metric T increases with decreasing MSE. For any given MSE, the metric

FIG. 2. Three artificial fields f_1 , f_2 , and f_3 to show the advantage and disadvantage of the linear correlation coefficient R and the MSE.

FIG. 3. The 2006 JJA (left) mean precipitation maps and (right) diurnal Hovmöller diagrams (meridionally averaged from 35° to 45°N) over the USGP from (top) TRMM and (bottom) MERRA.

T increases with increasing R . This metric T provides a skill score that has a maximum possible value of 1 (for $R = 1$ and $\text{MSE} = 0$) that indicates a perfect skill and a possible minimum value of 0 (for $R = -1$) that indicates no skill at all. When $\text{MSE} \neq 0$ and $R \neq -1$, then the metric T is between 0 and 1.

Next, we use a couple of examples to demonstrate the usefulness and advantage of this metric T to the previous ones. Figure 3 shows the 2006 summer (JJA) mean precipitation maps (left panels) and diurnal Hovmöller diagrams (right panels) over the USGP from TRMM (upper panels) and MERRA (lower panels). Here we treat the summer mean precipitation from TRMM as the reference field r and the summer mean precipitation from MERRA as the model simulated field f . Please note that we used the original MERRA instead of MERRA-2 here as an example, and MERRA-2 performs much better than MERRA in this test, as documented later in the paper. The metric T of the summer mean precipitation field from MERRA based on that from TRMM is calculated based on the two-dimensional (longitude and latitude) precipitation maps shown in the left two panels of Fig. 3. The final metric T is 0.87 with $R = 0.93$ and $\text{MSE} = 0.18$, which indicates a very good agreement between TRMM and MERRA, and this is supported by the visual inspection of the spatial maps. Both TRMM and MERRA show low precipitation over

the Rockies and the northwestern part of the USGP and high precipitation over the eastern part of the USGP. The metric T of the summer mean precipitation diurnal Hovmöller diagram (meridionally averaged from 35° to 45°N) from MERRA based on that from TRMM is calculated based on the two-dimensional (longitude and hour) precipitation diurnal Hovmöller diagrams shown in the right two panels of Fig. 3. The final metric T is 0.008 with $R = 0.014$ and $\text{MSE} = 3.95$, which indicates a very poor agreement between TRMM and MERRA, and this is supported by the visual inspection of the diurnal Hovmöller diagrams. TRMM shows an eastward propagating precipitation signal in the diurnal time scale while MERRA shows mainly a standing precipitation signal with little eastward propagation. These two examples demonstrate that this metric T is a very useful skill score to help quantify the model performance and discern good versus poor simulations.

Table 4 lists the metric T skill score and the metric skill scores from Taylor (2001), Hirota et al. (2011), and Watterson et al. (2014) for the 2006 summer mean precipitation maps (left panels) and diurnal Hovmöller diagrams (right panels) in Fig. 3. Here, the metric skill scores from Gleckler et al. (2008), Murphy et al. (2004), and Reichler and Kim (2008) are not listed because the metric skill score from Gleckler et al. (2008) is not normalized and thus is difficult to compare, whereas the

TABLE 4. The metric T and some metrics from Table 3 for Fig. 3. Note the metric from Watterson et al. (2014) is normalized by 1000, so all metrics are within 0 and 1 for an easy comparison.

Physical patterns	Metric skill scores				Metric T	Taylor (2001)	Hirota et al. (2011)	Watterson et al. (2014)
	R	MSE	$1 - \text{MSE}/(\text{Bias}^2 + \sigma_f^2 + \sigma_r^2)$					
2006 summer mean precipitation pattern	0.93	0.18	0.90		0.87	0.93	0.84	0.71
2006 summer mean precipitation diurnal Hovmöller diagram	0.014	3.95	0.015		0.008	0.46	0.06	0.009

metric skill scores from Murphy et al. (2004) and Reichler and Kim (2008) require the interannual variances that the current example lacks. Table 4 indicates that the metric T skill score is consistent with the metric skill scores from Taylor (2001), Hirota et al. (2011), and Watterson et al. (2014). However, the metric T skill score has a much bigger separation (from 0.87 to 0.008) than the metric skill score from Taylor (2001) (from 0.93 to 0.46) and a slightly bigger separation than the metric skill scores from Hirota et al. (2011) (from 0.84 to 0.06) and Watterson et al. (2014) (from 0.71 to 0.009) between the 2006 summer mean precipitation maps (left panels, a good simulation) and diurnal Hovmöller diagrams (right panels, a bad simulation). Thus, the metric T skill score is at least as good as the metric skill scores from Taylor (2001), Hirota et al. (2011), and Watterson et al. (2014) and probably all other metrics in Table 2 in distinguishing good and bad simulations. The major advantage of this metric T is that it uses both R and MSE, two fundamental and complementary statistics, while the previous metrics rely mainly on just one of them. Furthermore, the metric T is a generic one and can be applied to any model-simulated field f of any dimension (either spatial only or temporal only or spatial-temporal combined) and any resolution depending on the available reference dataset r .

If the reference dataset r is perfect, then the above-defined metric T will provide a useful skill score between 0 and 1 for the model-simulated field f . However, in reality, the reference dataset r usually has its own uncertainty. Thus, there is uncertainty of metric T for the model-simulated field f due to the uncertainty of the reference dataset r . Also, a metric T of 1 does not necessarily mean a perfect skill either. To solve this problem, similar to the solution used by Taylor (2001), we can use two reference datasets r_1 and r_2 to calculate the metric T for f (T_1 and T_2) and the metric T of r_2 with respect to r_1 (T_0), a possible highest T value. Then we can normalize T_1 and T_2 by T_0 , that is, $T_1 = T_1/T_0$, $T_2 = T_2/T_0$. Then a metric T_1 or T_2 of 1 will mean a perfect skill relative to r_1 or r_2 , and the mean of T_1 and T_2 [$\bar{T} = 0.5 \times (T_1 + T_2)$] and the difference between T_1 and T_2 [$\Delta T = \text{abs}(T_1 - T_2)$] can denote the best estimate and

its uncertainty of the metric T for the model-simulated field f based on the reference datasets r_1 and r_2 .

5. Performance metric application to the downscaled climate simulations of summer precipitation over the USGP

To evaluate the downscaled climate simulations of summer precipitation, especially that related to the MCSs, over the USGP, we applied this metric T developed in section 4 to three long-term JJA mean precipitation patterns over the USGP (the spatial pattern, the diurnal Hovmöller diagram, and the diurnal spatial pattern) for various NU-WRF runs as well as MERRA-2 and the GEOS-5 replay run (M2R12K). Two observational reference datasets (TRMM and ST4) are used to quantify the uncertainty of the metric T skill scores. The metric T skill score we developed is a generic metric and can be applied to any physical phenomenon for model evaluation. Whether or not the T skill score for each physical phenomenon is useful for model evaluation really depends on the physical phenomenon examined. Each physical phenomenon has its own advantages and disadvantages. Since both MCSs and isolated convection contribute to the summer precipitation over the USGP, and we have not made any attempt to distinguish the precipitation associated with MCSs and that associated with isolated convective events as previously done by Fritsch et al. (1986), Nesbitt et al. (2006), Feng et al. (2016), and Chang et al. (2016), the metric T scores for the mean spatial pattern, the diurnal Hovmöller diagram, and the diurnal spatial pattern describe the summer precipitation related to both MCSs and isolated convective events over the USGP. However, since the nocturnal peak of the precipitation and the diurnally eastward propagating rainfall feature over the USGP are mostly a result of the MCSs instead of the isolated convection in observations, the metric T scores for the diurnal Hovmöller diagram and the diurnal spatial pattern should provide some hint of evidence for MCS-like precipitation over the USGP (e.g., Gao et al. 2017). Nevertheless, to assess whether a model has true skills in simulating MCSs, it is best to

explicitly identify MCSs in the model in some way consistent with observations.

Please note that the long-term refers to 10 years (2000–09) for TRMM and most model datasets except 8 years (2002–09) for ST4 and 5 years (2000–04) for the B4 run. There is a time mismatch among the two observation datasets and various model simulation datasets. The TRMM data are from 2000 to 2009, while the ST4 data are from 2002 to 2009. The B4_N600 simulation is from 2000 to 2004, while the B12 and B24 simulations are from 2000 to 2009. There is an interannual variability in the summer season precipitation in the USGP, and our results may be sensitive to the period chosen. To test this sensitivity, we have examined our results based on the 3-yr period 2002–04 (all datasets are available), the 5-yr period 2000–04 (3 years of ST4 data), and the 10-yr period 2000–09 (8 years of ST4 data and 5 years of B4_N600 data). We found that the model simulations are very similar (with some differences for M2R12K), and the time period we chose did not affect our conclusions regarding the role of spatial resolution and nudging strength. However, the metric T skill scores are higher for the 10-yr model simulations because we are examining the climate features of the summer precipitation, such as the mean precipitation pattern and mean diurnal Hovmöller diagram (see supplementary material online). That was why we have focused our presentation on the 10-yr mean results.

Figure 4 shows the long-term summer mean precipitation spatial patterns over the USGP from the two observational reference datasets (TRMM and ST4), MERRA-2, M2R12K, and various NU-WRF model runs listed in Table 2. The metric T skill scores regarding the summer mean precipitation spatial patterns for MERRA-2, M2R12K, and various NU-WRF runs were calculated based on the two-dimensional (longitude and latitude) precipitation maps in Fig. 4 and are listed in red over the top right of each panel. The first (left) and second (middle) metric T skill score is based on TRMM and ST4 as the reference dataset, respectively, and the third (right) metric T skill score is their mean as our best estimate. The metric T skill score of ST4 based on TRMM ($T_0 = 0.86$) represents the possible highest metric T skill score based on these two reference datasets. All the metric T skill scores for various model runs have been normalized by T_0 (hereafter for Figs. 5 and 6).

Figure 5 shows the long-term summer mean precipitation diurnal Hovmöller diagrams over the USGP from the two observational reference datasets (TRMM and ST4), MERRA-2, M2R12K, and various NU-WRF model runs listed in Table 2. The metric T skill scores regarding the summer mean precipitation diurnal Hovmöller diagrams for MERRA-2, M2R12K, and

various NU-WRF runs were calculated based on the two-dimensional (longitude and hour) precipitation diurnal Hovmöller diagrams in Fig. 5 and are listed in red over the top right of each panel.

Figure 6 shows the long-term summer mean precipitation diurnal spatial patterns over the USGP from the two observational reference datasets (TRMM and ST4), MERRA-2, M2R12K, and various model runs (B24_Con) listed in Table 2. The metric T skill scores regarding the summer mean precipitation diurnal spatial patterns for MERRA-2, M2R12K, and various NU-WRF runs were calculated based on the three-dimensional (longitude, latitude, and hour) precipitation diurnal spatial patterns in Fig. 6 and are listed in red over the top right of each panel.

Table 5 summarizes the metric T skill scores regarding the three summer mean precipitation patterns over the USGP for MERRA-2, M2R12K, and various NU-WRF model runs. The first column indicates the analyzed patterns that each metric T skill score is based on, such as the summer mean precipitation spatial pattern (second row), the summer mean precipitation diurnal Hovmöller diagram (third row), and the summer mean precipitation diurnal spatial pattern (fourth row). The second column indicates the reference datasets (i.e., TRMM and ST4) based on which the metric T skill scores for each pattern (row) were calculated. The third column indicates the metric T skill scores for MERRA-2, M2R12K, and various NU-WRF model runs (column) for each pattern (row) based on the two reference datasets (each row within each pattern row). For each row of physical pattern, the first (upper) and second (middle) row indicates the metric T skill score based on TRMM (T_1) and ST4 (T_2) as the reference dataset, respectively, and the third (lower) row indicates their mean [$\bar{T} = 0.5 \times (T_1 + T_2)$] as our best estimate as well as their uncertainties [$\Delta T = \text{abs}(T_1 - T_2)$] for the metric T skill score. The metric T skill score of ST4 based on TRMM (T_0) represents the possible highest metric T skill score based on these two reference datasets. All the metric T skill scores for various model runs have been normalized by T_0 , so a value of 1 indicating a perfect skill. The metric T skill scores are generally between 0.5 and 0.7, and we consider this range as a fair simulation (neither too good nor too bad).

We will next discuss our results shown in Table 5 and Figs. 4–6 in the following four topics.

a. The similarities and differences between the observations and model simulations

There are some similarities between the observations and MERRA-2, M2R12K, and various NU-WRF model runs in all three precipitation patterns. MERRA-2, M2R12K, and most NU-WRF model runs can roughly

FIG. 4. The long-term JJA mean precipitation maps over the USGP from observations (TRMM and ST4), MERRA-2, and various model runs listed in Table 2. The red numbers in the top-right corner of each panel are the metric T skill scores for each model run based on TRMM (first), ST4 (second), and the mean of the first two (third). The metric T skill score of ST4 based on TRMM ($T_0 = 0.86$) represents the possible highest metric T skill score based on these two reference datasets. All the metric T skill scores for various model runs have been normalized by T_0 .

capture the major features of the summer precipitation patterns over the USGP (a metric T skill score of ~ 0.6 except for B24_Con and B12_Con), such as the summer mean precipitation spatial pattern, the summer mean precipitation diurnal Hovmöller diagram, and the summer mean precipitation diurnal spatial pattern. For example, all MERRA-2, M2R12K, and various NU-WRF model runs show low precipitation over the Rockies and the northwestern part of the USGP and high precipitation over the eastern part of the USGP as well as its southwest-to-northeast orientation (Fig. 4). All MERRA-2, M2R12K, and various NU-WRF model runs show some hint of an eastward propagating precipitation signal in the diurnal time scale from the

Rockies to the eastern part of the USGP (Figs. 5, 6), although with some prominent differences from the observations (to be discussed later). Thus, there is some credibility of the downscaled climate simulations of the summer precipitation over the USGP.

There are also significant differences between the observations and MERRA-2, M2R12K, and various NU-WRF model runs in all three precipitation patterns. For example, Fig. 4 indicates that there is a spurious high precipitation area in the Rockies region of Colorado and northern New Mexico (35° – 40° N, 108° – 103° W) in all NU-WRF and M2R12K model simulations for the summer mean precipitation spatial pattern. This spurious high precipitation is a persistent feature in all years

FIG. 5. As in Fig. 4, but for the long-term summer mean precipitation diurnal Hovmöller diagrams.

with some interannual variation. The magnitude of this spurious high precipitation is reduced with the spectral nudging and the increase of the nudging wavelength from 600 to 2000 km. Figures 5 and 6 indicate that this high precipitation in model simulations is mainly a result of the too early (~ 3 h earlier) and too strong precipitation over the Rocky Mountains, a common diurnal cycle bias in GCM simulations (e.g., Dai and Trenberth 2004; Dirmeyer et al. 2012; Jin et al. 2016). This spurious high precipitation and too early (~ 3 h earlier) and too strong precipitation also exist in MERRA-2, but with a much smaller magnitude due probably to its coarser resolution (0.625° longitude $\times$ 0.5° latitude). Thus, this spurious high-precipitation area in Colorado and northern New Mexico in all NU-WRF and M2R12K

model simulations may come from MERRA-2, which provides the boundary conditions and nudging constraint for the model runs. This could also be due to physics issues (regardless of exact parameterization) shared by all models, including the model used by the reanalysis.

Figure 4 also indicates that the summer mean precipitation at the eastern part of the USGP between 95° and 90° W in the control runs without nudging (B24_Con and B12_Con) is underestimated compared to the observations. Figures 5 and 6 indicate that this precipitation underestimation bias between 95° and 90° W is due mainly to the weak eastward propagating precipitation feature from the Rockies to the eastern part of the USGP in the B24_Con and B12_Con model

FIG. 6. The long-term summer mean precipitation diurnal cycle spatial pattern over the USGP from observations (TRMM and ST4), MERRA-2, and various model runs listed in Table 2. Each row in the panels represents one UTC time every 3 h ranging from 0130 UTC at the top to 2230 UTC at the bottom. The red numbers in the top-right corner of each panel are the metric T skill scores for each model run based on TRMM (first), ST4 (second), and the mean of the first two (third). The metric T skill score of ST4 based on TRMM (T_0) represents the possible highest metric T skill score based on these two reference datasets. All the metric T skill scores for various model runs have been normalized by T_0 .

simulations. Figure 4 also indicates that the model precipitation underestimation bias between 95° and 90°W is progressively reduced when the model is nudged and the spectral nudging wavelength increases from 600 to 2000 km. This improvement is a combined effect of the increased eastward propagating precipitation feature in the evening from the Rockies to the eastern part of the USGP and the increased precipitation due probably to isolated convection over the eastern part of the USGP.

Figures 5 and 6 also indicate that the diurnally eastward propagating feature related to MCSs dominates the diurnal cycle over the USGP, while the diurnal cycle from isolated convection over the eastern part of the USGP between 95° and 90°W is rather weak in TRMM and ST4 observations. However, the diurnally eastward propagating feature related to MCSs is underestimated, and the diurnal cycle from isolated convection over the eastern part of the USGP between 95° and 90°W is overestimated in model simulations. As a result, the simulated precipitation fields (NU-WRF and M2R12K) typically miss the smooth eastward propagation from

afternoon to early morning seen in observations. For example, there exist local minima in the model-simulated precipitation around 100°W and 0600 UTC not found in observations. These local minima in model simulations suggest the simulated propagating precipitation signals tend to dissipate at 100°W more frequently than observed. The model-simulated precipitation also exhibits a diurnal peak around 2100–0900 UTC over the eastern part of the USGP from 95° to 90°W , which does not appear in observations. This afternoon precipitation peak is merely a standing signal and not associated with propagating signals. Again, these errors also exist in MERRA-2 but with a much smaller magnitude, due probably to its coarser resolution (0.625° longitude $\times$ 0.5° latitude). Thus, these diurnal cycle errors in all NU-WRF and M2R12K model simulations may come from MERRA-2, which provides the boundary conditions and nudging constraint for the model runs. This could also be due to physics issues (regardless of exact parameterization) shared by all models, including the model used by the reanalysis.

FIG. 6. (Continued)

b. Value added of running the high-resolution RCM or GCM

The differences of the metric T skill scores between various NU-WRF model runs and MERRA-2 are mixed. There is an increase of the metric T skill score from MERRA-2 to some model runs, such as B24_N2000 and B12_N2000 for the summer mean precipitation spatial pattern and M2R12K for the summer mean precipitation diurnal Hovmöller diagram. There is also a decrease of the metric T skill score from MERRA-2 to some model runs, such as B24_Con and B12_Con for all physical patterns and B4_N600 for the summer mean precipitation spatial pattern. However, there are no significant increases of the metric T skill score from MERRA-2 to most other model runs, and the metric T skill scores of most other model runs are around 0.6. The various diagnostics (e.g., spatial maps and Hovmöller diagrams) in Figs. 4–6 do corroborate

this conclusion based on the metric T skill scores. This implies that running the high-resolution RCM or GCM to downscale the coarse-grid reanalysis does not add much value for the summer mean precipitation over the USGP, at least for the comparison with the TRMM and ST4 products. There may be some improvement in the rainfall magnitudes from MERRA-2 to various model runs; however, this improvement might be obscured by the degradation in the detailed spatial-temporal structure. As a result, there is no significant improvement in the overall metric T skill scores.

c. Role of the horizontal grid resolution

Table 5 indicates that there are no significant differences among the metric T skill scores for various NU-WRF model runs at different spatial resolutions. For example, the metric T skill scores regarding the summer mean precipitation patterns for B24_N600, B12_N600, and B4_N600 are all around 0.6. The

TABLE 5. The metric T skill scores regarding the summer mean precipitation over the USGP for various model runs listed in Table 2 (uncertainty is denoted by “unc”).

Physical phenomena	Reference data r	Model simulated field f									
		MERRA-2	M2R12K	B24_Con	B24_N600	B24_N2000	B12_Con	B12_N600	B12_N2000	B4_N600	
Summer mean precipitation pattern	TRMM	0.76	0.67	0.49	0.67	0.92	0.51	0.64	0.85	0.57	
	ST4, 0.86	0.66	0.53	0.44	0.58	0.73	0.43	0.55	0.67	0.52	
	Mean $\pm$ unc	0.71 ± 0.10	0.60 ± 0.14	0.46 ± 0.05	0.63 ± 0.09	0.83 ± 0.19	0.47 ± 0.08	0.60 ± 0.09	0.76 ± 0.18	0.55 ± 0.03	
Summer mean precipitation diurnal Hovmöller diagram	TRMM	0.67	0.81	0.62	0.72	0.71	0.57	0.71	0.70	0.68	
	ST4, 0.89	0.70	0.82	0.69	0.77	0.75	0.62	0.77	0.75	0.75	
	Mean $\pm$ unc	0.69 ± 0.03	0.82 ± 0.01	0.66 ± 0.08	0.75 ± 0.05	0.73 ± 0.04	0.59 ± 0.05	0.74 ± 0.06	0.73 ± 0.05	0.72 ± 0.07	
Summer mean precipitation diurnal spatial pattern	TRMM	0.77	0.84	0.63	0.78	0.81	0.58	0.76	0.76	0.66	
	ST4, 0.67	0.60	0.61	0.56	0.64	0.71	0.49	0.64	0.68	0.57	
	Mean $\pm$ unc	0.69 ± 0.17	0.72 ± 0.27	0.59 ± 0.07	0.71 ± 0.14	0.76 ± 0.1	0.53 ± 0.09	0.70 ± 0.12	0.72 ± 0.08	0.62 ± 0.09	

metric T skill scores regarding the summer mean precipitation diurnal Hovmöller diagrams for B24_N600, B12_N600, and B4_N600 are all around 0.73. The metric T skill scores regarding the summer mean precipitation diurnal spatial patterns for B24_N600, B12_N600, and B4_N600 are all around 0.7. The diagnostics in Figs. 4–6 do corroborate this conclusion based on the metric T skill scores. Thus, the increase of the spatial resolution does not necessarily increase the performance of the dynamically downscaled RCM simulation for the summer mean precipitation over the USGP, at least for the comparison with the TRMM and ST4 products.

Furthermore, the B4_N600 simulations actually have the lowest metric T score compared to the B24 and B12 model simulations. The current result is contrary to some previous studies that have shown that the summer precipitation diurnal cycle over the USGP is improved in the 4-km model simulations (e.g., Clark et al. 2007; Sun et al. 2016; Gao et al. 2017). Possible reasons for this difference are speculated as follows. First, the cumulus parameterization schemes were used in the current B4_N600 model run but not in the previous 4-km model runs (e.g., Clark et al. 2007; Sun et al. 2016; Gao et al. 2017). The deficiencies in the cumulus parameterization schemes may overshadow the benefit of the increased spatial resolution in the model simulations. Our comparison of the convective and large-scale precipitation fields from the 4-km simulations (figures not shown) indicates that the majority of the precipitation is from the convective precipitation instead of the large-scale precipitation. This implies that the convective scheme is still very active at the 4-km resolution. Second, the spectral nudging was used in the current B4_N600 model run but not in the previous 4-km model runs (e.g., Clark et al. 2007; Sun et al. 2016; Gao et al. 2017). The deficiencies in MERRA-2 that provide the constraint for the calculation of the horizontal wind velocities, temperature, and geopotential heights above the boundary layer may overshadow the benefit of the increased spatial resolution in the simulations. Third, the precipitation related to MCSs might be a result of the propagating gravity waves or the mountain-valley winds, and their dynamical process might have been resolved already by the 24-km model resolution. Thus, the further increase of the model spatial resolution may not be very helpful. To test the validity of our speculation, further in-depth investigation is needed that is beyond the scope of this paper.

d. Role of the spectral nudging

As discussed earlier, Figs. 4–6 indicate that the spurious high precipitation and too early (~ 3 h earlier) and too strong precipitation in the Rockies region of

Colorado and northern New Mexico (35° – 40° N, 108° – 103° W) are reduced with the spectral nudging and the increase of the nudging wavelength from 600 to 2000 km. [Figures 4–6](#) also indicate that the model precipitation underestimation bias over the eastern part of the USGP between 95° and 90° W and the weaker eastward propagating precipitation feature in the evening from the Rockies to the eastern part of the USGP are progressively reduced when the model is nudged and the spectral nudging wavelength increases from 600 to 2000 km. As a result, [Table 5](#) indicates that the metric T skill scores are relatively higher for the spectrally nudged NU-WRF model runs (greater than 0.6) than the control NU-WRF model runs (mostly less than 0.6). For example, the metric T skill score regarding the summer mean precipitation spatial pattern is 0.46 for B24_Con, but it is 0.63 for B24_N600 and 0.83 for B24_N2000. The metric T skill score regarding the summer mean precipitation diurnal spatial pattern is 0.59 for B24_Con but over 0.7 for B24_N600 and B24_N2000. This indicates that the spectral nudging and the increase of the nudging wavelength from 600 to 2000 km clearly have a positive impact on the dynamically downscaled climate simulations of the summer precipitation over the USGP. The reason for this improvement needs further investigation in the future that is beyond the scope of this paper.

6. Summary and discussion

Recently NASA initiated a multicenter joint project to assess the credibility and value added of the dynamic downscaled climate simulations ([Ferraro et al. 2017](#)). The NASA MERRA-2 reanalysis was used to drive the NU-WRF regional model and the high-resolution GEOS-5 global model. A number of experiments were conducted with various spatial resolutions and nudging strengths for simulations over the CONUS during the period 2000–09. The results of these experiments were compared to observational datasets to evaluate the model simulations. The objective of this paper is to summarize our work of this downscaling project on the development of a model performance metric and its application to assess the credibility of downscaled climate simulations of summer precipitation features, especially the diurnal cycle related to MCSs, over the USGP and its dependency on spatial resolution and spectral nudging strength.

We first developed a new performance metric T for climate models that is a combination of the previous metrics from [Taylor \(2001\)](#), [Hirota et al. \(2011\)](#), and [Watterson et al. \(2014\)](#) and uses both linear correlation coefficient R and mean square error (MSE). For any given R , the metric T increases with the decreasing

MSE. For any given MSE, the metric T increases with the increasing R . This metric T provides a skill score that has a maximum possible value of 1 (for $R = 1$ and $\text{MSE} = 0$) that indicates a perfect skill and a possible minimum value of 0 (for $R = -1$) that indicates no skill at all. When $\text{MSE} \neq 0$ and $R \neq -1$, then the metric T is between 0 and 1. Through a couple of examples, we have shown that the metric T is consistent with, but with a bigger separation than, the previous metrics between good and bad simulations. This metric T is also a generic one and can be applied to any model-simulated field of any dimension (either spatial only or temporal only or spatial–temporal combined) and any resolution depending on the available reference data.

We then applied this metric T to three long-term summer mean precipitation patterns over the USGP—the mean spatial pattern, the diurnal Hovmöller diagram, and the diurnal spatial pattern—to evaluate the downscaled climate simulations of summer mean precipitation over the USGP for various NU-WRF runs as well as MERRA-2 and the GEOS-5 replay run (M2R12K). The diurnal Hovmöller diagram and the diurnal spatial pattern are particularly useful in describing the summer precipitation related to the MCSs over the USGP. Two observational reference datasets (TRMM and ST4) are used as the observational reference datasets and to quantify the uncertainty of the metric T skill scores. The major conclusions of our evaluation of the summer mean precipitation based on the metric T skill scores for MERRA-2, M2R12K, and various NU-WRF model runs are as follows.

First, there are some similarities between the observations and MERRA-2, M2R12K, and various NU-WRF model runs in all three precipitation patterns. MERRA-2, M2R12K, and most NU-WRF model runs can roughly capture the major features of the summer precipitation patterns over the USGP (a metric T skill score of ~ 0.6). Thus, there is some credibility of downscaled climate simulations of the summer precipitation over the USGP. However, there are also significant differences between the observations and various model runs in all three precipitation patterns. For example, the spurious high precipitation and too early (~ 3 h earlier) and too strong precipitation in the Rockies region of Colorado and northern New Mexico (35° – 40° N, 108° – 103° W) are found in all model simulations. The model precipitation over the eastern part of the USGP between 95° and 90° W and the eastward propagating precipitation feature in the evening from the Rockies to the eastern part of the USGP are also underestimated. Some of these model biases also exist in MERRA-2 and thus they may come from MERRA-2, which provides

the boundary conditions and nudging constraint for the model runs.

Second, there is no significant increase of the metric T skill scores from MERRA-2 to M2R12K and most NU-WRF model runs. This implies that running the high-resolution RCM or GCM to downscale the coarse-grid reanalysis (which were at approximately 0.5° resolution) does not add much value for the summer precipitation over the USGP, at least for the comparison with the TRMM and ST4 products and for the performance metric examined here.

Third, there are no significant differences among various NU-WRF model runs at different spatial resolutions. Thus, the increase of the spatial resolution does not necessarily increase the performance of the dynamically downscaled climate simulation of the summer precipitation over the USGP, at least for the spatial resolutions we tested here (i.e., 4, 12, and 24 km).

Fourth, the spectral nudging and the increase of the nudging wavelength from 600 to 2000 km clearly have a positive impact on the dynamically downscaled climate simulations of the summer precipitation over the USGP as evidenced by the relatively higher metric T skill scores for the spectrally nudged NU-WRF model runs than the control NU-WRF model runs and for the spectrally nudged NU-WRF model runs with longer wavelengths than the spectrally nudged NU-WRF model runs with shorter wavelengths. This indicates that the spectral nudging and the increase of the nudging wavelength from 600 to 2000 km do help the dynamically downscaled climate simulations of the summer precipitation over the USGP.

Overall, the mixed results of our experiments did not demonstrate a compelling positive value for the dynamic downscaling for the summer precipitation over the USGP. However, it is extremely important to recognize the limitations associated with the observations, metrics, and the evaluation process. For example, the current results were all based on the regridded coarse-resolution observational reference dataset (TRMM). It is likely that the benefit of the higher-resolution RCM and GCM model runs is underexplored. To do the evaluations, model results must be examined at the grid resolutions on the order of 1–10 km, where physical processes (e.g., convection, extreme weather) and relevant boundary conditions (e.g., land–sea contrast and topography) are expected to have considerable bearing on climate impacts, decision support, and associated assessments. Satellite products and even reanalysis products have limited useful information at these fine scales. Although gridded in situ data of precipitation exist (e.g., ST4), they are undersampled in some places, such as mountains or sparsely populated regions. Thus, to critically and better

assess the value added of the downscaled climate simulations of the summer precipitation, especially the precipitation related to the MCSs over the USGP, high spatial and temporal resolution observational datasets are definitely needed. Future efforts should be invested in developing these observational datasets.

Acknowledgments. This study was supported by the NASA Downscaling Project led by Dr. Tsengdar Lee at NASA headquarters. The authors want to thank Dr. Tsengdar Lee and all team members of this project for various discussions and contributions. The contribution from the Jet Propulsion Laboratory (JPL) personnel was performed at the JPL, California Institute of Technology, under a contract with NASA. We also want to thank the editor Dr. L. Ruby Leung and three anonymous reviewers for their constructive comments that helped to improve this paper.

REFERENCES

- Bacmeister, J. T., M. J. Suarez, and F. R. Robertson, 2006: Rain reevaporation, boundary layer–convection interactions, and Pacific rainfall patterns in an AGCM. *J. Atmos. Sci.*, **63**, 3383–3403, doi:[10.1175/JAS3791.1](https://doi.org/10.1175/JAS3791.1).
- , M. F. Wehner, R. B. Neale, A. Gettelman, C. Hannay, P. H. Lauritzen, J. M. Caron, and J. E. Truesdale, 2014: Exploratory high-resolution climate simulations using the Community Atmosphere Model (CAM). *J. Climate*, **27**, 3073–3099, doi:[10.1175/JCLI-D-13-00387.1](https://doi.org/10.1175/JCLI-D-13-00387.1).
- Bosilovich, M. G., and Coauthors, 2015: MERRA-2: Initial evaluation of the climate. NASA Tech. Memo. NASA/TM-2015-104606/Vol. 43, 145 pp., <https://gmao.gsfc.nasa.gov/pubs/docs/Bosilovich803.pdf>.
- Bretherton, C. S., J. R. McCaa, and H. Grenier, 2004: A new parameterization for shallow cumulus convection and its application to marine subtropical cloud-topped boundary layers. Part I: Description and 1D results. *Mon. Wea. Rev.*, **132**, 864–882, doi:[10.1175/1520-0493\(2004\)132<0864:ANPFSC>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<0864:ANPFSC>2.0.CO;2).
- Carbone, R. E., and J. D. Tuttle, 2008: Rainfall occurrence in the U.S. warm season: The diurnal cycle. *J. Climate*, **21**, 4132–4146, doi:[10.1175/2008JCLI2275.1](https://doi.org/10.1175/2008JCLI2275.1).
- , —, D. A. Ahijevych, and S. B. Trier, 2002: Inferences of predictability associated with warm season precipitation episodes. *J. Atmos. Sci.*, **59**, 2033–2056, doi:[10.1175/1520-0469\(2002\)059<2033:IOPAWW>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<2033:IOPAWW>2.0.CO;2).
- Chang, W., M. L. Stein, J. L. Wang, V. R. Kotamarthi, and E. J. Moyer, 2016: Changes in spatiotemporal precipitation patterns in changing climate conditions. *J. Climate*, **29**, 8355–8376, doi:[10.1175/JCLI-D-15-0844.1](https://doi.org/10.1175/JCLI-D-15-0844.1).
- Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Wea. Rev.*, **129**, 569–585, doi:[10.1175/1520-0493\(2001\)129<0569:CAALSH>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2).
- Chen, H. M., T. J. Zhou, R. C. Yu, and J. Li, 2009: Summer rain fall duration and its diurnal cycle over the US Great Plains. *Int. J. Climatol.*, **29**, 1515–1519, doi:[10.1002/joc.1806](https://doi.org/10.1002/joc.1806).

- Chou, M.-D., and M. J. Suarez, 1999: A solar radiation parameterization for atmospheric studies. Tech. Memo. NASA/TM-1999-104606, Vol. 15, 38 pp., <http://gmao.gsfc.nasa.gov/pubs/docs/Chou136.pdf>.
- , X.-Z. Liang, and M. M.-H. Yan, 2001: A thermal infrared radiation parameterization for atmospheric studies. NASA/TM-2001-104606, Vol. 19, 54 pp., <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20010072848.pdf>.
- Clark, A. J., W. A. Gallus, and T. C. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.*, **135**, 3456–3473, doi:10.1175/MWR3467.1.
- Dai, A., and K. E. Trenberth, 2004: The diurnal cycle and its depiction in the Community Climate System Model. *J. Climate*, **17**, 930–951, doi:10.1175/1520-0442(2004)017<0930:TDCAD>2.0.CO;2.
- , F. Giorgi, and K. E. Trenberth, 1999: Observed and model-simulated diurnal cycles of precipitation over the contiguous United States. *J. Geophys. Res.*, **104**, 6377–6402, doi:10.1029/98JD02720.
- Davis, C. A., K. W. Manning, R. E. Carbone, S. B. Trier, and J. D. Tuttle, 2003: Coherence of warm-season continental rainfall in numerical weather prediction models. *Mon. Wea. Rev.*, **131**, 2667–2679, doi:10.1175/1520-0493(2003)131<2667:COWCRI>2.0.CO;2.
- Dickinson, R. E., R. M. Errico, F. Giorgi, and G. T. Bates, 1989: A regional climate model for the western United States. *Climatic Change*, **15**, 383–422, doi:10.1007/BF00240465.
- Dirmeyer, P. A., and Coauthors, 2012: Simulating the diurnal cycle of rainfall in global climate models: Resolution versus parameterization. *Climate Dyn.*, **39**, 399–418, doi:10.1007/s00382-011-1127-9.
- Dudhia, J., 2015: The Weather Research and Forecasting model: 2015 annual update. *16th Annual WRF User's Workshop*, Boulder, CO, NCAR, 1.1, http://www2.mmm.ucar.edu/wrf/users/workshops/WS2015/extended_abstracts/1.1.pdf.
- Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.*, **108**, 8851, doi:10.1029/2002JD003296.
- Feng, Z., L. R. Leung, S. Hagos, R. A. Houze, C. D. Burleson, and K. Balaguru, 2016: More frequent intense and long-lived storms dominate the springtime trend in central US rainfall. *Nat. Commun.*, **7**, 13429, doi:10.1038/ncomms13429.
- Ferraro, R., D. E. Waliser, and C. Peters-Lidard, 2017: NASA Downscaling Project Final Report, NASA/TP-2017-219579, 52 pp., <https://trs.jpl.nasa.gov/handle/2014/45705>.
- Fritsch, J. M., R. J. Kane, and C. R. Chelius, 1986: The contribution of mesoscale convective weather systems to the warm-season precipitation in the United States. *J. Climate Appl. Meteor.*, **25**, 1333–1345, doi:10.1175/1520-0450(1986)025<1333:TCOMCW>2.0.CO;2.
- Gao, Y., L. R. Leung, C. Zhao, and S. Hagos, 2017: Sensitivity of U.S. summer precipitation to model resolution and convective parameterizations across gray zone resolutions. *J. Geophys. Res. Atmos.*, **122**, 2714–2733, doi:10.1002/2016JD025896.
- Giorgi, F., 1990: Simulation of regional climate using a limited area model nested in a general circulation model. *J. Climate*, **3**, 941–963, doi:10.1175/1520-0442(1990)003<0941:SORCUA>2.0.CO;2.
- , and G. T. Bates, 1989: The climatological skill of a regional model over complex terrain. *Mon. Wea. Rev.*, **117**, 2325–2347, doi:10.1175/1520-0493(1989)117<2325:TCSOAR>2.0.CO;2.
- , and W. J. Gutowski, 2015: Regional dynamical downscaling and the CORDEX initiative. *Annu. Rev. Environ. Resour.*, **40**, 467–490, doi:10.1146/annurev-environ-102014-021217.
- Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972.
- Grell, G. A., and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.*, **14**, 5233–5250, doi:10.5194/acp-14-5233-2014.
- Hall, A., 2014: Projecting regional change. *Science*, **346**, 1461–1462, doi:10.1126/science.aaa0629.
- Hirota, N., Y. N. Takayabu, M. Watanabe, and M. Kimoto, 2011: Precipitation reproducibility over tropical oceans and its relationship to the double ITCZ problem in CMIP3 and MIROC5 climate models. *J. Climate*, **24**, 4859–4873, doi:10.1175/2011JCLI4156.1.
- Huffman, G. J., and Coauthors, 2007: The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeorol.*, **8**, 38–55, doi:10.1175/JHM560.1.
- Hughes, M., and A. Hall, 2010: Local and synoptic mechanisms causing Southern California's Santa Ana winds. *Climate Dyn.*, **34**, 847–857, doi:10.1007/s00382-009-0650-4.
- Janjić, Z. I., 1990: The step-mountain coordinate: Physical package. *Mon. Wea. Rev.*, **118**, 1429–1443, doi:10.1175/1520-0493(1990)118<1429:TSMCPP>2.0.CO;2.
- , 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, doi:10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.
- , 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., <http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf>.
- Jiang, X. N., N. C. Lau, and S. A. Klein, 2006: Role of eastward propagating convection systems in the diurnal cycle and seasonal mean of summertime rainfall over the U.S. Great Plains. *Geophys. Res. Lett.*, **33**, L19809, doi:10.1029/2006GL027022.
- Jin, E. K., I. J. Choi, S. Y. Kim, and J. Y. Han, 2016: Impact of model resolution on the simulation of diurnal variations of precipitation over East Asia. *J. Geophys. Res. Atmos.*, **121**, 1652–1670, doi:10.1002/2015JD023948.
- Kim, J., and Coauthors, 2017: Winter precipitation characteristics in western US related to atmospheric river landfalls: Observations and model evaluations. *Climate Dyn.*, doi:10.1007/s00382-017-3601-5, in press.
- Klein, S. A., X. N. Jiang, J. Boyle, S. Malyshev, and S. C. Xie, 2006: Diagnosis of the summertime warm and dry bias over the U.S. Southern Great Plains in the GFDL climate model using a weather forecasting approach. *Geophys. Res. Lett.*, **33**, L18805, doi:10.1029/2006GL027567.
- Knierel, J. C., D. A. Ahijevych, and K. W. Manning, 2004: Using temporal modes of rainfall to evaluate the performance of a numerical weather prediction model. *Mon. Wea. Rev.*, **132**, 2995–3009, doi:10.1175/MWR2828.1.
- Koster, R. D., M. J. Suarez, A. Ducharme, M. Stieglitz, and P. Kumar, 2000: A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure. *J. Geophys. Res.*, **105**, 24 809–24 822, doi:10.1029/2000JD900327.
- Kumar, S. V., and Coauthors, 2006: Land information system: An interoperable framework for high resolution land surface

- modeling. *Environ. Modell. Software*, **21**, 1402–1415, doi:10.1016/j.envsoft.2005.07.004.
- Lang, S. E., W.-K. Tao, R. Cifelli, W. Olson, J. Halverson, S. Rutledge, and J. Simpson, 2007: Improving simulations of convective systems from TRMM LBA: Easterly and westerly regimes. *J. Atmos. Sci.*, **64**, 1141–1164, doi:10.1175/JAS3879.1.
- , —, X. Zeng, and Y. Li, 2011: Reducing the biases in simulated radar reflectivities from a bulk microphysics scheme: Tropical convective systems. *J. Atmos. Sci.*, **68**, 2306–2320, doi:10.1175/JAS-D-10-05000.1.
- Lee, M. I., and Coauthors, 2007a: An analysis of the warm-season diurnal cycle over the continental United States and northern Mexico in general circulation models. *J. Hydrometeor.*, **8**, 344–366, doi:10.1175/JHM581.1.
- , and Coauthors, 2007b: Sensitivity to horizontal resolution in the AGCM simulations of warm season diurnal cycle of precipitation over the United States and northern Mexico. *J. Climate*, **20**, 1862–1881, doi:10.1175/JCLI4090.1.
- Leung, L. R., and Y. Qian, 2003: The sensitivity of precipitation and snowpack simulations to model resolution via nesting in regions of complex terrain. *J. Hydrometeor.*, **4**, 1025–1043, doi:10.1175/1525-7541(2003)004<1025:TSOPAS>2.0.CO;2.
- , —, and X. D. Bian, 2003: Hydroclimate of the western United States based on observations and regional climate simulation of 1981–2000. Part I: Seasonal statistics. *J. Climate*, **16**, 1892–1911, doi:10.1175/1520-0442(2003)016<1892:HOTWUS>2.0.CO;2.
- , —, —, W. M. Washington, J. G. Han, and J. O. Roads, 2004: Mid-century ensemble regional climate change scenarios for the western United States. *Climatic Change*, **62**, 75–113, doi:10.1023/B:CLIM.0000013692.50640.55.
- Liang, X. Z., L. Li, A. Dai, and K. E. Kunkel, 2004: Regional climate model simulation of summer precipitation diurnal cycle over the United States. *Geophys. Res. Lett.*, **31**, L24208, doi:10.1029/2004GL021054.
- Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.
- Lock, A. P., A. R. Brown, M. R. Bush, G. M. Martin, and R. N. B. Smith, 2000: A new boundary layer mixing scheme. Part I: Scheme description and single-column model tests. *Mon. Wea. Rev.*, **128**, 3187–3199, doi:10.1175/1520-0493(2000)128<3187:ANBLMS>2.0.CO;2.
- Loikith, P. C., D. E. Waliser, J. Kim, and R. Ferraro, 2017: Evaluation of cool season precipitation event characteristics over the Northeast US in a suite of downscaled climate model hindcasts. *Climate Dyn.*, doi:10.1007/s00382-017-3837-0.
- Mearns, L. O., and Coauthors, 2012: The North American Regional Climate Change Assessment Program: Overview of phase I results. *Bull. Amer. Meteor. Soc.*, **93**, 1337–1362, doi:10.1175/BAMS-D-11-00223.1.
- , and Coauthors, 2013: Climate change projections of the North American Regional Climate Change Assessment Program (NARCCAP). *Climatic Change*, **120**, 965–975, doi:10.1007/s10584-013-0831-3.
- Miguez-Macho, G., G. L. Stenchikov, and A. Robock, 2004: Spectral nudging to eliminate the effects of domain position and geometry in regional climate model simulations. *J. Geophys. Res.*, **109**, D13104, doi:10.1029/2003JD004495.
- Moorthi, S., and M. J. Suarez, 1992: Relaxed Arakawa–Schubert: A parameterization of moist convection for general-circulation models. *Mon. Wea. Rev.*, **120**, 978–1002, doi:10.1175/1520-0493(1992)120<0978:RASAP0>2.0.CO;2.
- Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772, doi:10.1038/nature02771.
- Nelson, B. R., O. P. Prat, D. J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, doi:10.1175/WAF-D-14-00112.1.
- Nesbitt, S. W., and E. J. Zipser, 2003: The diurnal cycle of rainfall and convective intensity according to three years of TRMM measurements. *J. Climate*, **16**, 1456–1475, doi:10.1175/1520-0442-16.10.1456.
- , R. Cifelli, and S. A. Rutledge, 2006: Storm morphology and rainfall characteristics of TRMM precipitation features. *Mon. Wea. Rev.*, **134**, 2702–2721, doi:10.1175/MWR3200.1.
- Park, S., and C. S. Bretherton, 2009: The University of Washington shallow convection and moist turbulence schemes and their impact on climate simulations with the Community Atmosphere Model. *J. Climate*, **22**, 3449–3469, doi:10.1175/2008JCLI2557.1.
- Peters-Lidard, C. D., and Coauthors, 2007: High-performance Earth system modeling with NASA/GSFC's Land Information System. *Innov. Syst. Softw. Eng.*, **3**, 157–165, doi:10.1007/s11334-007-0028-x.
- , and Coauthors, 2015: Integrated modeling of aerosol, cloud, precipitation and land processes at satellite-resolved scales. *Environ. Modell. Software*, **67**, 149–159, doi:10.1016/j.envsoft.2015.01.007.
- Pielke, R. A., and R. L. Wilby, 2012: Regional climate downscaling: What's the point? *Eos, Trans. Amer. Geophys. Union*, **93**, 52–53, doi:10.1029/2012EO050008.
- Prat, O. P., and B. R. Nelson, 2015: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). *Hydrol. Earth Syst. Sci.*, **19**, 2037–2056, doi:10.5194/hess-19-2037-2015.
- Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.*, **89**, 303–311, doi:10.1175/BAMS-89-3-303.
- Riley, G. T., M. G. Landin, and L. F. Bosart, 1987: The diurnal variability of precipitation across the central Rockies and adjacent Great Plains. *Mon. Wea. Rev.*, **115**, 1161–1172, doi:10.1175/1520-0493(1987)115<1161:TDVOPA>2.0.CO;2.
- Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note NCAR/TN-468+STR, 88 pp., doi:10.5065/D6DZ069T.
- Sun, X., M. Xue, J. Brotzge, R. A. McPherson, X.-M. Hu, and X.-Q. Yang, 2016: An evaluation of dynamical downscaling of Central Plains summer precipitation using a WRF-based regional climate model at a convection-permitting 4 km resolution. *J. Geophys. Res. Atmos.*, **121**, 13 801–13 825, doi:10.1002/2016JD024796.
- Tao, W. K., and Coauthors, 2003: Microphysics, radiation and surface processes in the Goddard Cumulus Ensemble (GCE) model. *Meteor. Atmos. Phys.*, **82**, 97–137.
- Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, doi:10.1029/2000JD900719.
- Tian, B., I. M. Held, N. C. Lau, and B. J. Soden, 2005: Diurnal cycle of summertime deep convection over North America: A

- satellite perspective. *J. Geophys. Res.*, **110**, D08108, doi:[10.1029/2004JD005275](https://doi.org/10.1029/2004JD005275).
- Vukicevic, T., and R. M. Errico, 1990: The influence of artificial and physical factors upon predictability estimates using a complex limited-area model. *Mon. Wea. Rev.*, **118**, 1460–1482, doi:[10.1175/1520-0493\(1990\)118<1460:TIOAAP>2.0.CO;2](https://doi.org/10.1175/1520-0493(1990)118<1460:TIOAAP>2.0.CO;2).
- Wallace, J. M., 1975: Diurnal variations in precipitation and thunderstorm frequency over conterminous United States. *Mon. Wea. Rev.*, **103**, 406–419, doi:[10.1175/1520-0493\(1975\)103<0406:DVIPAT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1975)103<0406:DVIPAT>2.0.CO;2).
- Watterson, I. G., 1996: Non-dimensional measures of climate model performance. *Int. J. Climatol.*, **16**, 379–391, doi:[10.1002/\(SICI\)1097-0088\(199604\)16:4<379::AID-JOC18>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1097-0088(199604)16:4<379::AID-JOC18>3.0.CO;2-U).
- , J. Bathols, and C. Heady, 2014: What influences the skill of climate models over the continents? *Bull. Amer. Meteor. Soc.*, **95**, 689–700, doi:[10.1175/BAMS-D-12-00136.1](https://doi.org/10.1175/BAMS-D-12-00136.1).
- Zhang, G. J., 2003: Roles of tropospheric and boundary layer forcing in the diurnal cycle of convection in the U.S. southern Great Plains. *Geophys. Res. Lett.*, **30**, 2281, doi:[10.1029/2003GL018554](https://doi.org/10.1029/2003GL018554).