Geophysical Research Letters

RESEARCH LETTER
10.1029/2018GL081702

Key Points:
- The precipitation forecasting skill of FV3 with advanced physics schemes is evaluated at a convection-allowing resolution
- FV3 is shown to have comparable performance to the WRF model, despite the lack of storm-scale initialization, suggesting its promise for predicting precipitation at a convection-allowing resolution
- The Thompson microphysics scheme slightly outperforms the NSSL scheme in FV3, while no PBL scheme clearly outperforms others among the five schemes examined

Supporting Information:
- Supporting Information S1

Correspondence to:
M. Xue,
mxue@ou.edu

Citation:
Zhang, C., Xue, M., Supinie, T. A., Kong, F., Snook, N., Thomas, K. W., et al. (2019). How well does an FV3-based model predict precipitation at a convection-allowing resolution? Results from CAPS forecasts for the 2018 NOAA hazardous weather test bed with different physics combinations. *Geophysical Research Letters*, *46*, 3523–3531. https://doi.org/10.1029/2018GL081702

Received 20 DEC 2018
Accepted 13 FEB 2019
Accepted article online 18 MAR 2019
Published online 25 MAR 2019

How Well Does an FV3-Based Model Predict Precipitation at a Convection-Allowing Resolution? Results From CAPS Forecasts for the 2018 NOAA Hazardous Weather Test Bed With Different Physics Combinations

Chunxi Zhang[1], Ming Xue[1,2], Timothy A. Supinie[1], Fanyou Kong[1], Nathan Snook[1], Kevin W. Thomas[1], Keith Brewster[1], Youngsun Jung[1], Lucas M. Harris[3], and Shian-Jiann Lin[3]

[1]Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, OK, USA, [2]School of Meteorology, University of Oklahoma, Norman, OK, USA, [3]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA

Abstract The Geophysical Fluid Dynamics Laboratory (GFDL) Finite-Volume Cubed-Sphere (FV3) numerical forecast model was chosen in late 2016 by the National Weather Service (NWS) to serve as the dynamic core of the Next-Generation Global Prediction System (NGGPS). The operational Global Forecasting System (GFS) physics suite implemented in FV3, however, was not necessarily suitable for convective-scale prediction. We implemented several advanced physics schemes from the Weather Research and Forecasting (WRF) model within FV3 and ran 10 forecasts with combinations of five planetary boundary layer and two microphysics (MP) schemes, with an ~3.5-km convection-allowing grid two-way nested within am ~13-km grid spacing global grid during the 2018 Spring Forecasting Experiment at National Oceanic and Atmospheric Administration (NOAA)'s Hazardous Weather Testbed. Objective verification results show that the Thompson MP scheme slightly outperforms the National Severe Storms Laboratory MP scheme in precipitation forecast skill, while no planetary boundary layer scheme clearly stands out. The skill of FV3 is similar to that of the more-established WRF at a similar resolution. These results establish the viability of the FV3 dynamic core for convective-scale forecasting as part of the single-core unification of the NWS modeling suite.

Plain Language Summary In this paper we examine how well the Finite-Volume Cubed-Sphere (FV3) model predicts precipitation over the Contiguous United States (CONUS) when run at a resolution sufficient to explicitly predict convective storms. FV3 was chosen to replace the operational Global Forecast System (GFS) by the National Weather Services (NWS) in late 2016, but its performance for convective-scale regional forecasting was previously untested. The physical parameterization schemes available in FV3 were mostly from GFS and not necessarily suitable for convective-scale predictions. We implemented several advanced physics schemes taken from a more established and most widely used convective-scale model, the Weather Research and Forecasting (WRF) model, into FV3, including schemes for treating turbulence exchanges in atmospheric boundary layer and those for representing cloud and precipitation processes (microphysics). We ran 10 forecasts each day using different combinations of the physics schemes to evaluate their performance in FV3 as part of the 2018 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed Spring Forecasting Experiment. With these advanced schemes, FV3 is shown to be capable of predicting precipitation with skill comparable to WRF. The precipitation forecast is somewhat sensitive to the microphysics schemes used, but not particularly sensitive to the atmospheric boundary layer scheme. Our study shows that the newly selected FV3 model, when equipped with appropriate physics parameterization schemes, can serve as the foundation for the next-generation regional forecasting models of the NWS.

1. Introduction

After extensive intercomparison and evaluation of several candidates, the U.S. National Weather Service (NWS) selected the Geophysical Fluid Dynamics Laboratory (GFDL) Finite Volume Cubed-Sphere dynamical core (FV3, also called FV^3; Putman & Lin, 2007) in late 2016 to serve as the single dynamical core of the

Next-Generation operational Global Prediction System (NGGPS), which will become operational in 2019. The goal is to use the FV3 dynamic core for mesoscale and convective scale applications, replacing the current suite of operational regional models of the NWS, some of which are based on the Weather Research and Forecasting (WRF) model (Skamarock et al., 2005). Using a unified modeling framework for prediction at all scales has proven successful at other centers such as UK Met Office (Walters et al., 2017) and is expected to help the NWS focus its resources on building the best model for the nation.

The current NWS operational High-Resolution Rapid Refresh (HRRR; Benjamin et al., 2016) and the high-resolution window of the North American Mesoscale Forecast System (NAM) have horizontal grid spacings of about 3 km, typically referred to as *convection-allowing* resolution (Bauer et al., 2015; Clark et al., 2016). In the future, horizontal grid spacing of 1 km or finer will be required to support operational Warn-On-Forecast efforts (Stensrud et al., 2009). While the FV3 dynamic core is nonhydrostatic and is thus, in principle, capable of handling convective-scale flows at nonhydrostatic scales, its performance for convective-scale applications has not been systematically examined. Moreover, the GFS physics suite coupled to FV3 had been designed and tuned for global predictions on much coarser grids for hydrostatic flows. For convection-allowing forecasts, more advanced physics schemes/packages need to be implemented and evaluated.

Toward the above goals, we have added several advanced microphysics (MP) and planetary boundary layer (PBL) schemes from WRF version 3.9 into the GFS physics suite coupled to the latest (as of May 2018) version of the FV3 core through the Interoperable Physics Driver (IPD) interface; henceforth, we will simply refer to this model as *FV3*. We ran 10 FV3 forecasts each day with different combinations of MP and PBL schemes during the 2018 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE). Forecasts were run with an ~3.5-km convection-allowing grid covering the Contiguous United States (CONUS) nested inside a global FV3 grid with a 13-km grid spacing. Since precipitation is both highly impactful and difficult to forecast, as a first effort, we evaluate the skill of the convection-allowing FV3 model with varying PBL and MP schemes in predicting precipitation relative to the skill of WRF forecasts using a similar grid spacing.

In this paper, we describe FV3 and its configurations for the 2018 HWT SFE forecasts in section 2. Section 3 evaluates the precipitation forecast skills of FV3 using the neighborhood-based equitable threat score (ETS; Clark et al., 2010) and the fractions skill score (FSS; Roberts & Lean, 2008). Conclusions and discussions are presented in section 4, along with a comparison of FV3 to other limited-area models.

2. Model Description and Configuration

FV3 solves the fully compressible Euler equations using the forward-in-time scheme and vertically Lagrangian discretization of Lin (2004), the horizontal discretization of Lin and Rood (1997), the scalar advection scheme of Lin and Rood (1996), and the finite-volume pressure gradient force scheme of Lin (1997). FV3 uses a hybrid D-grid staggering (Harris & Lin, 2013), and there are concerns in the community about its performance for convection-allowing-applications; evaluating the latter is the main purpose of this study. The horizontal discretization in FV3 is performed on a gnomonic equiangular cubed-sphere grid (Putman & Lin, 2007); six cubed-sphere tiles are used to cover the entire global domain (see Figure S1a). At the edges of those tiles, two one-sided third-order extrapolations are averaged to form a directionally symmetric scheme across the edges (Putman & Lin, 2007). The two-way grid nesting algorithm is described in Harris and Lin (2013). The nested regional and global grids are run concurrently on separate sets of processors. For the 2018 HWT SFE, the global grid has a uniform horizontal grid spacing of 13 km, and the nested regional grid has varying horizontal grid spacing, averaging around 3.5 km (Figure S1b). As in the operational GFS model, 63 vertical levels are used for both grids, although the level specifications are not the same. Scalar advection on the nest uses the piecewise-parabolic method with a monotonic limiter applied to condensate species, number concentrations, and ozone, and without a limiter applied to thermodynamic scalars (vorticity, mass, virtual potential temperature, etc.). All results shown in this paper are from the nested regional FV3 domain.

The Center for Analysis and Prediction of Storms (CAPS) implemented the partially two-moment Thompson MP scheme (Thompson et al., 2008) and the fully two-moment National Severe Storms Laboratory (NSSL) MP scheme (Mansell et al., 2010), as well as other schemes not tested during the 2018 HWT SFE, into FV3. CAPS also implemented the Yonsei University PBL scheme (YSU; Hong et al., 2006), the scale-aware

Table 1
The FV3 Model Configurations for 2018 HWT SFE

Forecast name	Microphysics	PBL
fv3-phys01	Thompson	SA-MYNN
fv3-phys02	Thompson	MYNN
fv3-phys03	Thompson	SA-YSU
fv3-phys04	Thompson	YSU
fv3-phys05	Thompson	EDMF
fv3-phys06	NSSL	SA-MYNN
fv3-phys07	NSSL	MYNN
fv3-phys08	NSSL	SA-YSU
fv3-phys09	NSSL	YSU
fv3-phys10	NSSL	EDMF

Note. All forecasts use Global Forecasting System (GFS) T1534 initial conditions, the Noah land surface model, Rapid Radiative Transfer Model for general circulation model (RRTMG) for long-wave and short-wave radiation, and the Tiedtke cumulus scheme for the global grid only. HWT, Hazardous Weather Testbed; NSSL, National Severe Storms Laboratory; PBL, planetary boundary layer; SFE, Spring Forecasting Experiment.

YSU PBL scheme (SA-YSU; Shin & Hong, 2015), and the Mellor-Yamada-Nakanishi-Niino PBL scheme (MYNN; Nakanishi & Niino, 2006), which has the option of using scale-aware capabilities (SA-MYNN). We also implemented the Tiedtke cumulus scheme (Tiedtke, 1989; Zhang et al., 2011; Zhang & Wang, 2017), into FV3, employing it on the global domain only. Notably, none of these schemes were tuned for FV3. Both the Thompson and NSSL MP schemes calculate the radiation effective radii and pass them to the Rapid Radiative Transfer Model for General circulation model (RRTMG) radiation scheme (Iacono et al., 2008), which is used to calculate short-wave and long-wave radiation. The Xu and Randall (1996) cloud fraction calculation, NOAH land surface model (Ek et al., 2003), and surface layer scheme are as in the operational GFS.

For the 2018 HWT SFE, the combinations of PBL and MP schemes used in the 10 FV3 forecasts are shown in Table 1. The first group of five forecasts uses the Thompson MP scheme with five different PBL schemes, and the second group of five uses the NSSL MP scheme with the same five PBL schemes. In addition to the four PBL schemes implemented by CAPS, the fifth PBL scheme used is the hybrid Eddy-Diffusivity Mass-Flux (EDMF; Han et al., 2016) scheme from the operational GFS. With everything else being the same, the relative performance of the two MP and five PBL schemes can be evaluated. The forecasts were run from 00 UTC out to 84 hr on each of 25 days during the 2018 HWT SFE (on weekdays from 30 April 2018 to 1 June 2018). The 00 UTC operational GFS (T1534, ~13-km resolution) analyses were used to initialize both FV3 grids; no specific storm-scale initialization was employed.

3. Results

Categorical verification scores, such as ETS, are frequently used for verification of precipitation forecasts (Ebert, 2009). Categorical scores are based on the distribution of forecasted and observed events. In precipitation verification, an event is typically defined as the occurrence of rainfall exceeding a given amount over a defined time period (e.g., rainfall exceeding 10 mm over 3 hr). In categorical verification scores, a 2×2 contingency table, including hits, misses, false alarms, and correct negatives, is commonly used (Jolliffe & Stephenson, 2003; Wilks, 2006).

Due to the double-penalty issue associated with the traditional point-wise skill scores in the presence of position error (Ebert, 2008), neighborhood-based verifications are more suitable for convection-allowing predictions (Clark et al., 2010; Ebert, 2009; Roberts & Lean, 2008). Here we follow Clark et al. (2010) to calculate neighborhood ETS (NETS); an ETS of 1.0 denotes a perfect forecast, and 0.0 denotes a forecast with no skill. The FSS (Roberts & Lean, 2008) is another metric that tolerates position error via the use of a neighborhood. For FSS, events in a certain area are counted to generate fractions or probabilities. We mainly focus on NETS and FSS in this paper.

3.1. Choice of Precipitation Thresholds

Stage IV multisensor precipitation estimate data (ST4; Lin & Mitchell, 2005) from National Centers for Environmental Prediction are used as observations for precipitation verification. The ~4-km ST4 data are interpolated to the FV3 regional grid using a nearest neighbor method. Given the closeness of the resolutions of the grids, interpolation error should be limited to the structures of two grid spacings. The hourly accumulated precipitation tends to be localized in ST4 and FV3 forecasts (Figure S2), while localized heavy precipitation is of particular interest for storm-scale prediction. In terms of precipitation probability density function, there are noticeable differences among the forecasts. The Thompson scheme tends to overpredict heavier precipitation compared to METAR observations, while the NSSL scheme predicts precipitation probability density functions that are very close to that of observations (Figure S3). There are no clear differences among PBL schemes (Figure S3). Because forecast bias is known to have significant impact on ETS scores, and positive bias tends to give higher ETS scores, we present scores mainly using percentile thresholds (Zhu et al., 2015). Scores based on absolute thresholds are given as supporting information.

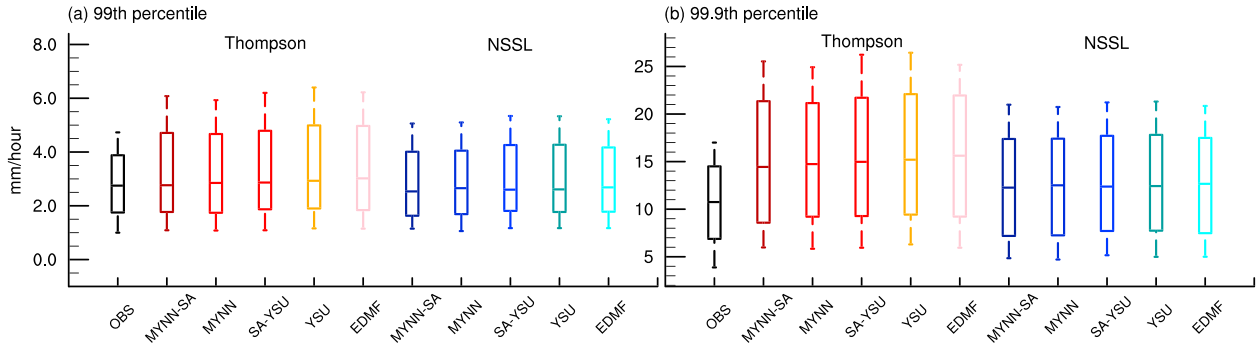

Figure 1. Box-and-whisker plots showing hourly accumulated precipitation values at (a) the 99th percentile and (b) the 99.9th percentile for Finite-Volume Cubed-Sphere (FV3) forecasts using different combinations of microphysics and planetary boundary layer schemes. The hourly accumulated precipitation value at the 99th (or 99.9th) percentile is calculated from all grid cells over Contiguous United States (CONUS) for 12- to 36 hr forecasts. Observed Stage-IV hourly precipitation data are interpolated onto the FV3 grid and plotted in black. Forecasts using the Thompson microphysics scheme are plotted in warm colors, and forecasts using National Severe Storms Laboratory (NSSL) scheme are plotted in cool colors. From bottom to top, each box-and-whisker plot shows the 10th, 25th, 50th, 75th, and 90th percentiles of hourly precipitation values at the 99th percentile (a) or the 99.9th percentile (b). The sample size, 600 (24 forecast hours times 25 cases) for each forecast, is believed to be sufficient to calculate the range of spread.

Box-and-whisker plots of hourly accumulated precipitation are shown in Figure 1 for the 99th percentile (a) and the 99.9th percentile (b) for all cases between 12- and 36-hr forecasts and for different combinations of MP and PBL schemes. Observed ST4 hourly precipitation is also shown. The observed median value for the 99th percentile is around 2.5 mm/hr, and both Thompson and NSSL schemes have similar median values. However, the spread of values is larger in the forecasts. The median value for the 99.9th percentile is around 11 mm/hr in the observations. The Thompson scheme has much higher median values (around 15 mm/hr), while the NSSL scheme has values of around 13 mm/hr. The spread of values remains wide in the forecasts. Again, differences among PBL schemes are relatively small.

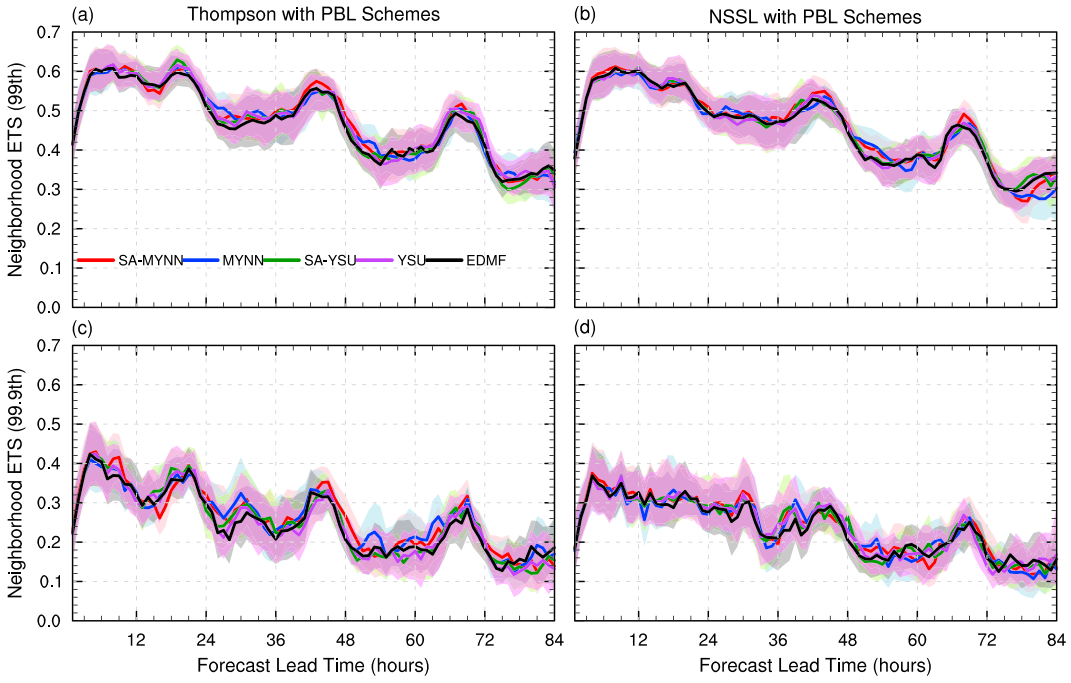

Figure 2. Time-series of neighborhood equitable threat score (NETS) (using a 45 km *r*) for forecasts using (a and c) the Thompson or (b and d) the National Severe Storms Laboratory (NSSL) scheme with different planetary boundary layer (PBL) schemes in Finite-Volume Cubed-Sphere (FV3) for hourly precipitation exceeding (a and b) the 99th percentile and (c and d) the 99.9th percentile. The lines are mean values from all 25 cases, while the shaded region indicates the 95% confidence interval of possible mean values based on 10,000 bootstrap resamplings from 25 cases for each forecast.

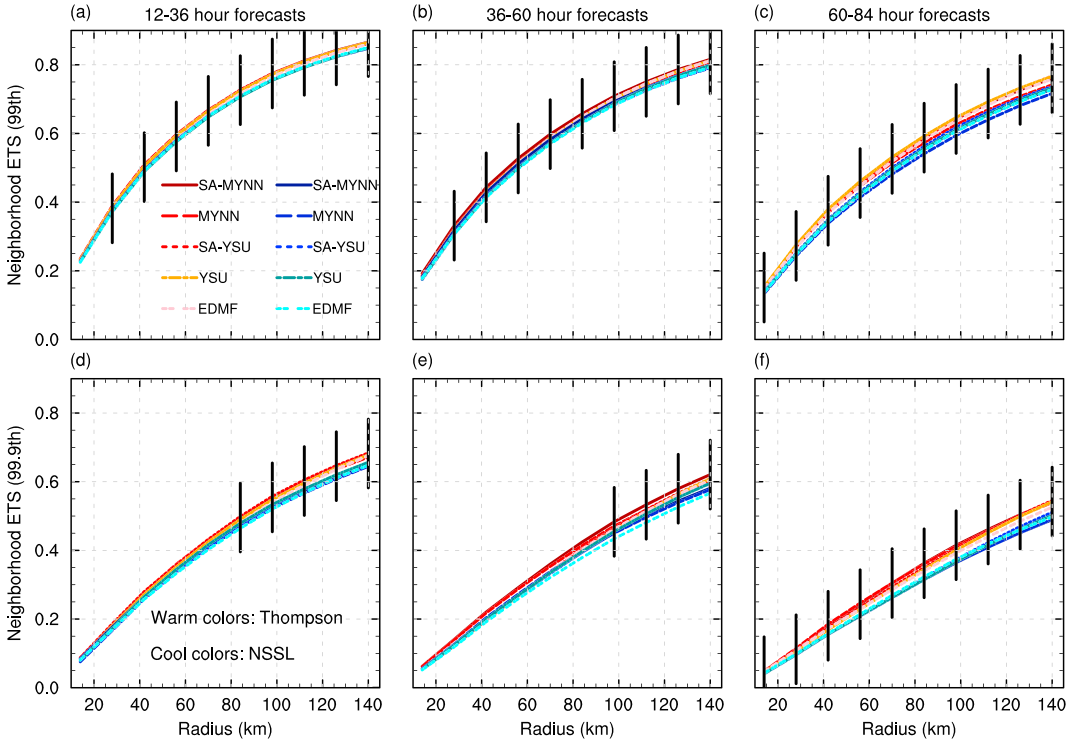

Figure 3. The aggregated neighborhood equitable threat score (NETS) as a function of ***r*** for hourly precipitation exceeding (a–c) the 99th percentile and (d–f) the 99.9th percentile. Plots are shown for (a and d) 12- to 36-, (b and e) 36- to 60-, and (c and f) 60- to 84-hr forecasts. Warm colors are used for forecasts running the Thompson microphysics scheme, and cool colors for forecasts running the National Severe Storms Laboratory (NSSL) microphysics scheme. A paired *t* test is employed to compare Thompson and NSSL forecasts. The vertical lines indicate neighborhood radii at which the differences of NETS between Thompson and NSSL forecasts are significant at the 99% confidence level.

3.2. Neighborhood Equitable Threat Score (NETS) and Fractions Skill Score (FSS)

NETS scores are plotted in Figure 2 for forecasts using Thompson or NSSL scheme combined with different PBL schemes. The mean is denoted by a solid line, and the 95% confidence interval, calculated using 10,000 bootstrap (Efron, 1982) samples, is indicated by the shaded region. A neighborhood radius, *r*, of 45 km is used. For both the 99th and 99.9th percentiles of hourly precipitation, NETS remains relatively steady out to 48 hr of forecast time after the initial spin-up. Little systematic difference is noted among forecasts using differing PBL schemes, although some significant differences exist in individual cases (see Figure S4). Among MP schemes, the Thompson scheme tends to produce higher NETSs during the diurnal peak of precipitation, while NSSL scheme shows slightly lower average scores. For fixed thresholds of 2.5 and 12.7 mm/hr, corresponding roughly to the average median values of 99th and 99.9th percentiles, the relative performances among the forecasts are similar (Figure S5).

The aggregated NETSs as a function of neighborhood radius, *r*, for the 99th and 99.9th percentiles of hourly precipitation are plotted in Figure 3 for 12- to 36-, 36- to 60-, and 60- to 84-hr forecasts. As expected, NETS increases with the radius. Again, the differences among PBL schemes are small (and insignificant). However, NETSs from the Thompson scheme are always higher than those from the NSSL scheme. The paired *t* test (Rietveld & van Hout, 2017) is used to test whether the difference of the mean aggregated NETS between Thompson and NSSL schemes is significant at the 99% confidence level. The differences are statistically significant at almost all radii for the 99th percentile and at radii larger than 85 km for the 99.9th percentile for 12- to 84-hr forecasts. The differences are largest for 60- to 84-hr forecasts (Figures 3c and 3f).

The FSS can also be used to identify the minimum spatial scale at which forecasts are skillful. The minimum scale is defined in terms of the uniform value (see Figure 4 in Roberts & Lean, 2008), which is defined as $0.5 + f_0/2$, where f_0 is the percentage of grid cells where an event was observed. The

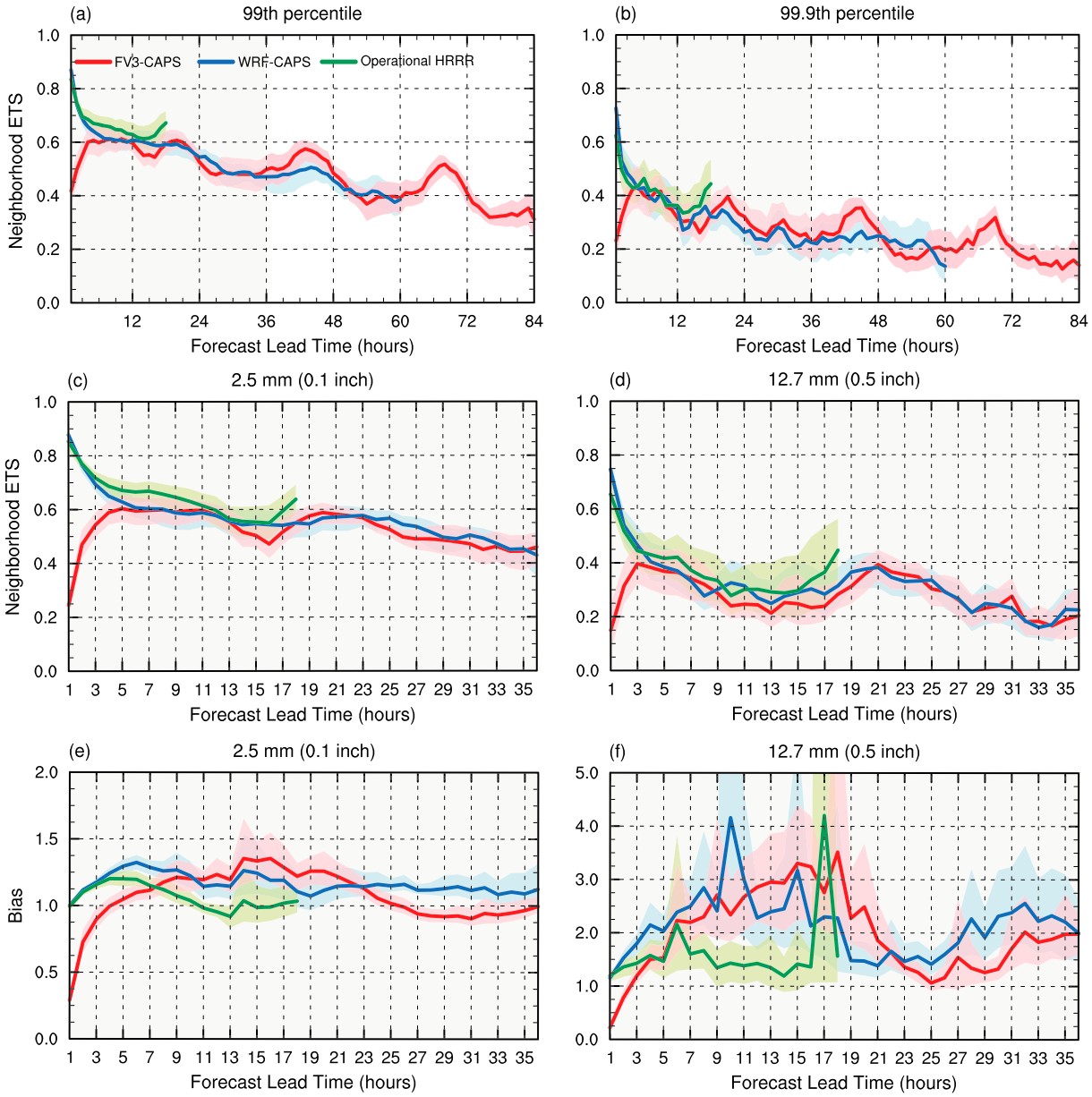

Figure 4. (a and b) Neighborhood equitable threat score (NETS; using a 45 km ***r*** and percentiles as thresholds) for Center for Analysis and Prediction of Storms (CAPS) Finite-Volume Cubed-Sphere (FV3; denoted FV3-CAPS), CAPS Weather Research and Forecasting (WRF; denoted WRF-CAPS), and operational High-Resolution Rapid Refresh (HRRR) version 2; (c and d) NETSs, and (e and f) frequency bias calculated with a 45 km ***r*** and absolute thresholds, for example, 2.5 and 12.7 mm/hr, respectively. FV3-CAPS is the FV3 forecast using the Thompson and SA-MYNN schemes, while WRF-CAPS is the WRF forecast using similar physics package as the operational HRRR. The lines are mean values for the 25 cases during the 2018 HWT SFE. The shading indicates the 95% confidence interval of possible mean values based on 10,000 bootstrap resamplings from the 25 cases for each forecast. Both CAPS WRF and operational HRRR include the assimilation of radar data in their initial conditions. CAPS WRF forecasts were run for 60 hr, while the operational HRRR was run for 18 hr. Only forecasts up to 36 hr (gray shading in (a) and (b)) are shown for bias (c–f). NETS and bias are calculated on the native grid for each forecast. Note that the vertical coordinate between (e) and (f) is different.

uniform value of FSS is thus equal to 0.505 for precipitation exceeding the 99th percentile or 0.5005 for the 99.9th percentile. The aggregated FSSs as a function of *neighborhood length* for the 99th and the 99.9th percentiles are plotted in Figure S6 for different forecast periods. The minimum scale is increased substantially with the forecast lead time. The differences among PBL schemes are once again quite small. Overall, the differences for 12- to 60-hr forecasts between Thompson and NSSL schemes are less substantial in terms of FSS than NETS.

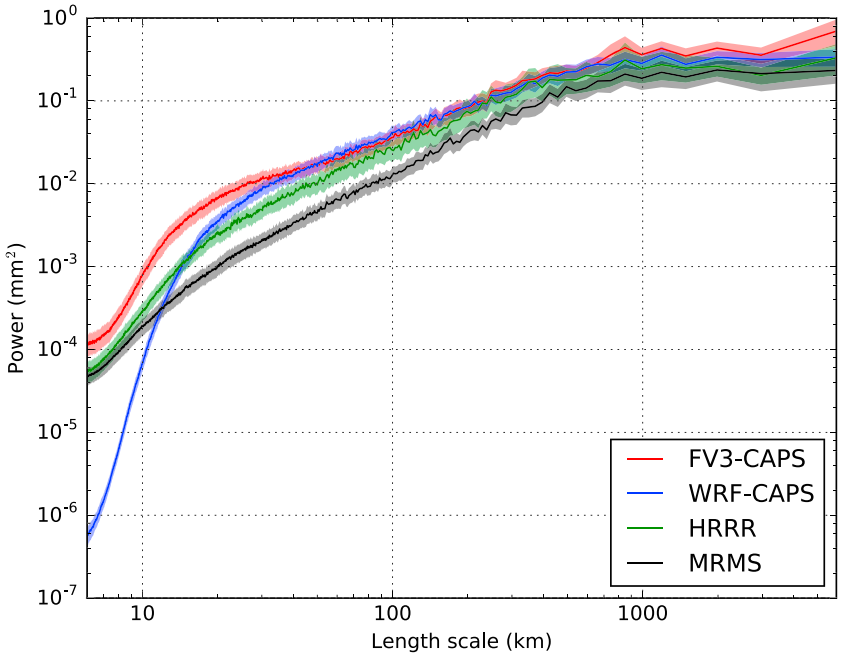

Figure 5. Two-dimensional spectral analyses of 6-hourly accumulated precipitation between 12 and 18 hr for Multi-Radar Multi-Sensor (MRMS), High-Resolution Rapid Refresh (HRRR), Center for Analysis and Prediction of Storms Finite-Volume Cubed-Sphere (CAPS FV3), and CAPS Weather Research and Forecasting (WRF). The lines are mean values from all 25 cases, while the shaded region indicates the 5th to 95th percentile range of possible mean values based on 10,000 bootstrap resamplings from 25 cases for each model.

3.3. Comparisons With WRF-Based Forecasts

The FV3 forecasts using the combination of Thompson (MP) and SA-MYNN (PBL) schemes (CAPS FV3) is compared to a 3-km WRF forecasts run as part of the CAPS Storm-Scale Ensemble Forecast (SSEF) system for the HWT SFE (details regarding configuration can be found at https://hwt.nssl.noaa.gov/sfe/2018/docs/HWT_SFE2018_operations_plan.pdf) and the 3-km operational HRRR forecasts. The HRRR has slightly higher NETSs for the 99th percentile and higher NETSs between 12 and 18 hr for the 99.9th percentile (Figures 4a and 4b), while the CAPS WRF and FV3 are comparable after hour 6. The HRRR also exhibits a slightly higher Critical Success Index for most forecasts (Schaefer, 1990), while Critical Success Index is comparable between FV3 and the CAPS WRF (Figure S7). The HRRR still has slightly higher NETS compared to that of CAPS WRF and FV3 at absolute thresholds of 2.5 and 12.7 mm/hr (which roughly match the 99th and 99.9th percentiles; Figures 4c and 4d). Differences in frequency bias (Wang, 2014) are relatively large (Figures 4e and 4f), with HRRR generally having lower biases. The low bias in FV3 before hour 3 is due to spin-up from coarse-resolution GFS analyses. Two-dimensional spectral analyses (Denis et al., 2002) of 6-hourly precipitation (Figure 5) show that the power below 30 km wavelength for HRRR is closest to that of Multi-Radar Multi-Sensor precipitation data (MRMS; Zhang et al., 2016), while CAPS FV3 has somewhat higher power than MRMS data at those wavelengths. The CAPS WRF shows a marked decrease in power for wavelengths shorter than 20 km, in large part due to the use of relatively strong sixth -order diffusion (Knievel et al., 2007; Xue, 2000) settings based on an experimental version of HRRR. The higher power at small scales in FV3 forecasts is believed to be linked to the use of a very low level of numerical diffusion; FV3 precipitation forecasts do tend to contain more near-grid-scale structures. The 6-hr accumulated precipitation fields from MRMS, HRRR, CAPS FV3, and WRF for 17 and 18 May are shown in Figure S8 as examples. The CAPS WRF forecast used similar WRF physics suite as the operational HRRR. Both HRRR and CAPS WRF included assimilation of radar data, resulting in high NETS for the first few hours. The hourly cycled hybrid data assimilation (Benjamin et al., 2016) compared to the single cycle data assimilation used in the CAPS WRF, along with long-term tuning, may explain HRRR's higher NETS during its 18-hr forecast.

4. Conclusions and Discussions

Hourly precipitation forecasts from a nested regional FV3 grid run at convection-allowing resolution during the 2018 HWT SFE are objectively evaluated using two verification scores: NETS and FSS. Subjective evaluation results from HWT SFE will be reported elsewhere. Several advanced PBL and MP schemes recently implemented by CAPS into FV3 are used among forecasts verified against hourly ST4 precipitation data (and against 6-hourly ST4 data in supplementary materials; see Figures S9–S13). Percentiles are used as the thresholds to reduce the impact of any bias in rainfall amounts on the scores. Thresholds at the 99th and the 99.9th percentiles are examined. The median hourly precipitation intensity for the 99th percentile is around 2.5 mm/hr for both ST4 and FV3 forecasts, while the 99.9th percentile is around 11 mm/hr for ST4 and between 13 and 15 mm/hr for FV3 forecasts using different MP schemes.

No PBL scheme clearly outperforms the others in terms of hourly precipitation forecast skill. Between the two MP schemes tested, the Thompson scheme tends to produce higher precipitation rates and exhibits slightly higher skill than the NSSL scheme, particularly for 60- to 84-hr forecasts, in terms of neighborhood-based ETS and FSS. Excluding oscillations associated with the diurnal cycle, forecast skill gradually deceases with forecast lead time, and both NETS and FSS increase, as expected, with neighborhood radius/scale length. For precipitation exceeding the 99th percentile, the minimum skillful scale in terms of FSS increases rapidly from 85 km for 12- to 36-hr forecasts to 190 km for 60- to 84-hr forecasts. For the 99.9th percentile the minimum scales are much larger, indicating that forecasting heavy rainfall remains quite challenging. Comparisons across PBL and MP schemes suggest that MP schemes have larger impacts on precipitation forecasting. NETS scores based on absolute thresholds show similar relative performances among the forecasts.

The FV3 forecasts with Thompson and SA-MYNN exhibit similar NETSs and frequency bias as a 3-km WRF forecasts produced by CAPS for HWT SFE after the first 6 hr of FV3 spin-up, but both slightly underperform the operational HRRR. Nevertheless, with the nested global-regional configuration, FV3 produced precipitation forecasts at a convection-allowing resolution with skills comparable to WRF-based operational and experimental forecasts. The comparison of single-configuration FV3 forecasts with CAPS SSEF during the 2017 Hydrometeorology Flash Flood and Intense Rainfall (FFaIR) experiment draws similar conclusions. When combined with state-of-the-science data assimilation and better tuned advanced physics suites, precipitation forecast skills meeting or exceeding current operational forecasting skills can be expected. Moreover, a stand-alone regional version of FV3 has recently become available, which will enable more direct comparisons of FV3 with other limited-area models by using identical or similar physics packages, initial conditions, and boundary conditions. These are planned for future HWT SFEs.

Acknowledgments
This work was primarily supported by NOAA grants NA16OAR4320115, NA17OAR4590186, and NA16NWS4680002. Computing was performed primarily at the University of Texas Advanced Computing Center. All precipitation data used in this study are available at ftp://ftp.caps.ou.edu/zhang_2019_grl.

References

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., et al. (2016). A North American hourly assimilation and model forecast cycle: The rapid refresh. *Monthly Weather Review*, *144*(4), 1669–1694. https://doi.org/10.1175/MWR-D-15-0242.1

Clark, A. J., Gallus, W. A., & Weisman, M. L. (2010). Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Weather and Forecasting*, *25*(5), 1495–1509. https://doi.org/10.1175/2010WAF2222404.1

Clark, P., Roberts, N., Lean, H., Ballard, S. P., & Charlton-Perez, C. (2016). Convection-permitting models: A step-change in rainfall forecasting. *Meteorological Applications*, *23*(2), 165–181. https://doi.org/10.1002/met.1538

Denis, B., Cote, J., & Laprise, R. (2002). Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (DCT). *Monthly Weather Review*, *130*(7), 1812–1829. https://doi.org/10.1175/1520-0493(2002)130<1812:SDOTDA>2.0.CO;2

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteorological Applications*, *15*(1), 51–64. https://doi.org/10.1002/met.25

Ebert, E. E. (2009). Neighborhood verification: A strategy for rewarding close forecasts. *Weather and Forecasting*, *24*(6), 1498–1510. https://doi.org/10.1175/2009WAF2222251.1

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. Soc. Ind. *Appl. Math*., **92**.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., & Tarpley, J. D. (2003). Implementation of Noah land-surface model advances in the NCEP operational mesoscale Eta mode. *Journal of Geophysical Research*, *108*(D12), 8851. https://doi.org/10.1029/2002JD003296

Han, J., Witek, M. L., Teixeira, J., Sun, R., Pan, H. L., Fletcher, J. K., & Bretherton, C. S. (2016). Implementation in the NCEP GFS of a hybrid Eddy-diffusivity mass-flux (EDMF) boundary layer parameterization with dissipative heating and modified stable boundary layer mixing. *Weather and Forecasting*, *31*(1), 341–352. https://doi.org/10.1175/WAF-D-15-0053.1

Harris, L. M., & Lin, S.-J. (2013). A two-way nested global-regional dynamical core on the cubed-sphere grid. *Monthly Weather Review*, *141*(1), 283–306. https://doi.org/10.1175/MWR-D-11-00201.1

Hong, S.-Y., Noh, Y., & Dudhia, J. (2006). A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, *134*(9), 2318–2341. https://doi.org/10.1175/MWR3199.1

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., & Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of Geophysical Research*, *113*, D13103. https://doi.org/10.1029/2008JD009944

Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecast verification: A practitioner's guide in atmospheric science*. West Sussex, UK: John Wiley.

Knievel, J. C., Bryan, G. H., & Hacker, J. P. (2007). Explicit numerical diffusion in the WRF model. *Monthly Weather Review*, *135*(11), 3808–3824. https://doi.org/10.1175/2007MWR2100.1

Lin, S. J. (1997). A finite-volume integration method for computing pressure gradient force in general vertical coordinates. *Quarterly Journal of the Royal Meteorological Society*, *123*(542), 1749–1762.

Lin, S.-J. (2004). A "vertically Lagrangian" finite-volume dynamical Core for global models. *Monthly Weather Review*, *132*(10), 2293–2307. https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2

Lin, S. J., & Rood, R. B. (1997). An explicit flux-form semi-Lagrangian shallow-water model on the sphere. *Quarterly Journal of the Royal Meteorological Society*, *123*(544), 2477–2498. https://doi.org/10.1002/qj.49712354416

Lin, S.-J., & Rood, R. B. (1996). Multidimensional flux-form semi-Lagrangian transport schemes. *Monthly Weather Review*, *124*(9), 2046–2070. https://doi.org/10.1175/1520-0493(1996)124<2046:MFFSLT>2.0.CO;2

Lin, Y., & K. E. Mitchell (2005). The NCEP Stage II/IV hourly precipitation analyses: Development and applications. Paper presented at 19th Conf. Hydrology. San Diego, CA: American Meteorological Society.

Mansell, E. R., Ziegler, C. L., & Bruning, E. C. (2010). Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *Journal of the Atmospheric Sciences*, *67*(1), 171–194. https://doi.org/10.1175/2009JAS2965.1

Nakanishi, M., & Niino, H. (2006). An improved Mellor–Yamada Level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Boundary-Layer Meteorology*, *119*(2), 397–407. https://doi.org/10.1007/s10546-005-9030-8

Putman, W. M., & Lin, S. H. (2007). Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, *227*(1), 55–78. https://doi.org/10.1016/j.jcp.2007.07.022

Rietveld, T., & van Hout, R. (2017). The paired t test and beyond: Recommendations for testing the central tendencies of two paired samples in research on speech, language and hearing pathology. *Journal of Communication Disorders*, *69*, 44–57. https://doi.org/10.1016/j.jcomdis.2017.07.002

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Schaefer, J. T. (1990). The critical success index as an indicator of warning skill. *Weather and Forecasting*, *5*(4), 570–575. https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2

Shin, H. H., & Hong, S.-Y. (2015). Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions. *Monthly Weather Review*, *143*(1), 250–271. https://doi.org/10.1175/MWR-D-14-00116.1

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. D. (2005). A description of the Advanced Research WRF version 2Rep., 88 pp, NCAR.

Stensrud, D. J., Xue, M., Wicker, L. J., Kelleher, K. E., Foster, M. P., Schaefer, J. T., et al. (2009). Convective-scale warn on forecast system: A vision for 2020. *Bulletin of the American Meteorological Society*, *90*(10), 1487–1500. https://doi.org/10.1175/2009BAMS2795.1

Thompson, G., Field, P. R., Rasmussen, R. M., & Hall, W. D. (2008). Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Monthly Weather Review*, *136*(12), 5095–5115. https://doi.org/10.1175/2008MWR2387.1

Tiedtke, M. (1989). A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, *117*(8), 1779–1800. https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2

Walters, D., Boutle, I., Brooks, M., Melvin, T., Stratton, R., Vosper, S., et al. (2017). The Met Office Unified Model Global Atmosphere 6.0/6.1 and JULES Global Land 6.0/6.1 configurations. *Geoscientific Model Development*, *10*(4), 1487–1520. https://doi.org/10.5194/gmd-10-1487-2017

Wang, C. C. (2014). On the calculation and correction of equitable threat score for model quantitative precipitation forecasts for small verification areas: The example of Taiwan. *Weather and Forecasting*, *29*(4), 788–798. https://doi.org/10.1175/WAF-D-13-00087.1

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd ed.). Amsterdam: Academic Press, Elsevier.

Xu, K.-M., & Randall, D. A. (1996). A semiempirical cloudiness parameterization for use in climate models. *Journal of Atmospheric Sciences*, *53*, 3084–3102.

Xue, M. (2000). High-order monotonic numerical diffusion and smoothing. *Monthly Weather Review*, *128*(8), 2853–2864. https://doi.org/10.1175/1520-0493(2000)128<2853:HOMNDA>2.0.CO;2

Zhang, C., & Wang, Y. (2017). Projected future changes of tropical cyclone activity over the western North and South Pacific in a 20-km-mesh regional climate model. *Journal of Climate*, *30*(15), 5923–5941. https://doi.org/10.1175/JCLI-D-16-0597.1

Zhang, C., Wang, Y., & Hamilton, K. (2011). Improved representation of boundary layer clouds over the Southeast Pacific in ARW-WRF using a modified Tiedtke cumulus parameterization scheme*. *Monthly Weather Review*, *139*(11), 3489–3513. https://doi.org/10.1175/MWR-D-10-05091.1

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., et al. (2016). Multi-radar multi-sensor (Mrms) quantitative precipitation estimation initial operating capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621–638. https://doi.org/10.1175/BAMS-D-14-00174.1

Zhu, K. F., Yang, Y., & Xue, M. (2015). Percentile-based neighborhood precipitation verification and its application to a landfalling tropical storm case with radar data assimilation. *Advances in Atmospheric Sciences*, *32*(11), 1449–1459. https://doi.org/10.1007/s00376-015-5023-9