Evaluation of high-resolution rapid refresh (HRRR) forecasts for extreme precipitation

Haowen Yue and Mekonnen Gebremichael

University of California, Los Angeles, Civil and Environmental Engineering, CA, United States of America

E-mail: yuehaowen@g.ucla.edu and mekonnen@seas.ucla.edu

Keywords: precipitation, HRRR, precipitation forecast, QPF, extreme events

Abstract

This study evaluates the accuracy of short-range (1-h to 18-hr lead-time) forecasts from the High-Resolution Rapid Refresh (HRRR) model for five extreme storms in the United States: (1) the September 21–23, 2016, frontal storms in Iowa, (2) the April 28-May 1, 2017, frontal storms in the Southern Midwestern US, (3) the August 25–31, 2017, Hurricane Harvey storms in Texas, (4) the September 13–17, 2018, Hurricane Florence storms in the Carolinas, and (5) the September 4–6, 2019, Hurricane Dorian storms in the Carolinas. The evaluation was carried out by comparison with gauge-corrected Multi-Radar/Multi-Sensor (MRMS-GC) products. In terms of temporal variability, there was a good agreement between the forecasted and observed precipitation on an hourly basis. Thus, the HRRR products provide relatively reliable forecasts. However, the forecasts were mostly biased: they tend to overestimate rainfall for both hurricanes, underestimate rainfall for tropical storms in Iowa, and produce almost unbiased estimates for the frontal storms in Southern Midwestern US. In terms of spatial pattern, the forecasts are able to capture the spatial pattern of hurricanes, however, they produce too many, localized, high-rain intensities for the frontal storms than what the observations show. With regard to the effect of lead times, the 1-h lead forecasts have often lower accuracy than the other lead-time forecasts, while there was no much systematic difference in accuracy among the 2-h to 18-h lead-time forecasts. The bias estimates in the forecast are also examined at different spatial scales, ranging from 2 km × 2 km all the way to 128 km × 128 km. The results show that the bias estimated at smaller spatial scales vary within a large range, mostly within the range of −100% to +100%, indicating that the bias estimates obtained at large scale (hundreds of km grids) are not applicable to bias estimates at small scales, and vice versa. Local-bias correction approaches are therefore preferable over global bias-correction approaches.

1. Introduction

Quantitative precipitation forecasts (QPFs) are increasingly used for a number of hydrological applications, such as, flood warning (Bartholmes and Todini 2005, Cuo et al 2011), reservoir management (Collischonn et al 2007, Saavedra Valeriano et al 2010), drought and water supply forecasting (Morss 2003, Ceppi et al 2014), among others. However, QPFs, typically obtained from numerical weather prediction (NWP) models, are subject to errors resulting from uncertainties in model initial conditions, model formulations, and chaotic behavior of the atmosphere (e.g., Berner et al 2015). The QPF errors are generally considered to be large for extreme precipitation due to the difficulty of modeling extreme precipitation in NWP models (e.g., Sukovich et al 2014). Information on QPF uncertainty, particularly for extreme precipitation, is necessary for guiding the development and improvement of a forecasting process as well as for decision making.

The High-Resolution Rapid Refresh (HRRR) model of the National Weather Service (NWS) provides state-of-the-art, short-range (lead times ranging from 1 h to 18 h), high-resolution (3 km, hourly) QPFs across the continental United States (CONUS), which are used for a number of applications. For example, the HRRR
forecasts are used as forcing inputs into the NWS’ National Water Model (NWM) that generates short-range hydrological forecasts at approximately 4000 locations across the CONUS. Studies evaluating the accuracy of QPF, particularly during extreme events, are very limited as shown below.

Gowan et al (2018) assessed the performance of forecasts by HRRR and coarse-resolution NWP models (GFS, NAM-3 km CONUS nest and NCAR Ensemble) over mountainous western US by comparing the forecasts to precipitation observation from the and precipitation analyses from the Parameter–Elevation Regressions on Independent Slopes Model (PRISM), and reported that the HRRR forecasts outperformed the coarse-resolution NWP models partly due to better representation of topography in the HRRR model. Pinto et al (2015) reported that the HRRR predicted too many mesoscale convective systems (MCSs) over the Great Plains and too few MCSs over the southeastern United States, and that the skill of the HRRR at predicting specific MCS events increased between 2012 and 2013, coinciding with changes in model physics and data assimilation technique. Cai and Dumais (2015) evaluated three weeks of HRRR forecasts during the summer of 2010 over the eastern United States, and reported that the HRRR model was able to capture the overall convective storm characteristics but with significant bias which varies substantially by region.

Bytheway and Kummerow (2015) estimated the performance of the 2013 experimental version of the HRRR forecasts for warm-season convective storms over central US by comparing the forecasts to Stage IV Nexrad Radar rainfall products (Ref), and reported the following results: a southward displacement of the storm structure by about 30 km, large overestimation in intensity and areal extent of precipitation, overestimation of heavy rain, and relatively better forecast skill for the 3-h lead forecast. Two years later, Bytheway et al (2017) performed a similar study using the research versions of latter-year (2013–2015) HRRR forecasts, and concluded overall improvements in the bias for maximum and mean rainfall intensity and storm location as a result of model upgrades, compared to the 2013 forecasts.

Ikeda et al (2013) investigated the accuracy of winter-season HRRR forecasts with a particular focus on areal extent of precipitation and the timing and location of regions of different precipitation phases during the 2010/11 cold season by comparing the forecasts to observations from the Automated Surface Observing System (ASOS) station network across the Eastern United States, and revealed that larger synoptically forced weather systems were better predicted than smaller weather systems. Ikeda et al (2017) examined HRRR model’s ability to forecast the surface precipitation phase for cold-storm cases from the 2013/14 and 2014/15 winter seasons over the east of the Rocky Mountains, and concluded that the HRRR model was able to represent the overall vertical thermodynamic structure in the mixed-phase precipitation regions. See et al (2018) evaluated the performance of HRRR forecasts for the September 20–23, 2016, frontal storms over Eastern Iowa by comparing the forecasts to the Multi-Radar/Multi-Sensor Gauge-Corrected (MRMS-GC) products, and reported different behavior for different statistics: while the forecast skill decreased with lead time (maximum skill achieved at 1-h lead), the bias and stand error improved with lead time.

The purpose of this study is to evaluate the accuracy of the HRRR QPFs for extreme precipitation, using the most recent HRRR model versions. We specifically focus on five selected extreme storms that cover different storm natures (namely, frontal storms and hurricanes) and different regions: (1) the September 21–23, 2016, frontal storms in Iowa, (2) the April 28-May 1, 2017, frontal storms in the Southern Midwestern US, (3) the August 25–31, 2017, Hurricane Harvey storms in Texas, (4) the September 13–17, 2018, Hurricane Florence storms in the Carolinas, and (5) the September 4–6, 2019, Hurricane Dorian storms in the Carolinas. The forecasts are evaluated by comparison with Multi-Radar/Multi-Sensor Gauge-Corrected (MRMS-GC) products, which is one of the best-quality and high-resolution (1-km, 1-h) precipitation products over CONUS often used for evaluating the accuracy climate models and satellite products (Zhang et al 2016). The short-range forecasts are quantitatively evaluated over a range of lead times (1 h to 18 h lead times) and on different spatiotemporal scales (hourly and storm-total, 2 km × 2 km to domain-averaged grids). Various techniques, such as bias, quantiles, and object-based methods are used to analyze different aspects of forecast quality.

2. Data and methods

2.1. Storm cases

Given our focus on extreme storms, we selected five disastrous storms from recent years as reported by the USGS Flood reports. These are: (1) the September 21–23, 2016, frontal storms in Iowa, (2) the April 28-May 1, 2017, frontal storms in the Southern Midwestern US, (3) the August 25–31, 2017, Hurricane Harvey storms in Texas, (4) the September 13–17, 2018, and (5) the September 4–6, 2019, Hurricane Dorian storms in the Carolinas., Hurricane Florence storms in the Carolinas. These cases allowed us to examine the uncertainties in forecasts during frontal storms as well as hurricanes in different parts of the U.S. Table 1 presents some of the key characteristics of these storms. The frontal system in Iowa had an accumulation of 91 mm over 69 h for the Cedar watershed (watershed area of 49,140 sq. km), and the frontal system in southern Midwestern had 158 mm
of rainfall over 95 h for the Meramec watershed (18,630 sq. km). The hurricanes had higher rainfall and lasted longer: hurricane Harvey in Texas resulted in 501 mm over 157 h for the Lower Sabine Watershed (17,856 sq. km), Hurricane Florence had 496 mm over 128 h for Lumber watershed (9,790 sq. km), and Hurricane Dorian had 109 mm over 62 h for Cape Fear watershed (9,120 sq. km).

### 2.2. HRRR forecast system

With the growing need for frequently-updated short-range weather guidance, the fine-resolution HRRR was developed and introduced into the operational model suite at the NOAA/National Centers for Environmental Prediction (NCEP) in September 2014. The HRRR has been undergoing continuous model improvement, almost on a yearly basis. The HRRR model is a real-time, hourly-updated, convection-allowing, storm resolving model running at 3 km horizontal resolution with 50 vertical levels over the CONUS. Its domain is nested within the 13 km Rapid Refresh mesoscale model, which also provides boundary conditions (Benjamin et al. 2016). The HRRR is built upon the Advanced Research version of Weather Research Forecast (WRF-ARW) model with the following physics options: the Goddard shortwave radiation scheme (Chou and Suarez 1994), the Rapid Radiative Transfer Model longwave radiation scheme (Mlawer et al. 1997), the Rapid Update Cycle smirnova land surface model (Smirnova et al. 1997), the Mellor-Yamada-Nakanishi-Niino boundary layer parameterization (Nakanishi and Niino 2004), and the Thompson mixed-phase microphysics scheme (Thompson et al. 2008). Initial fields are created using 3D-VAR data assimilation. The latent heating profile are calculated as a function of radar reflectivity (which came from the NWS WSR-88D network that is used to create the Stage IV products) which is assimilated at 3 km resolution every 15 min (Benjamin et al. 2016). Detailed information on model history and physics are provided in Benjamin et al. (2016).

The HRRR model provides operational forecasts at 3-km and hourly resolution, with lead times ranging from 1 h to 18 h. It also provides experimental sub-hourly products in order to test updated model performance and measure performances resulting from updates. Most recent versions are HRRRv2 available since July 2016.

### Table 1. Some of the key characteristics of the selected storms over nested watersheds.

| Watershed          | Domain Size (km²) | Elevation (m) | Major Land Cover       | Total Rainfall (mm) | Peak hourly rainfall (mm) | Duration (Hour) |
|--------------------|-------------------|---------------|------------------------|--------------------|--------------------------|-----------------|
| Cedar              | 49 140            | 326           | Cultivated Crops       | 91                 | 5.97                     | 69              |
| The September 21–23, 2016, frontal storms in Iowa |
| Flint              | 819               | 350           | Deciduous Forest       | 159                | 25.83                    | 56              |
| Big                | 7650              | 303           | Deciduous Forest       | 159                | 9.90                     | 91              |
| Gasconade          | 9108              | 286           | Deciduous Forest       | 140                | 9.85                     | 77              |
| Eleven Point       | 9486              | 260           | Deciduous Forest       | 109                | 10.81                    | 68              |
| Illinois           | 11 025            | 355           | Deciduous Forest       | 126                | 14.26                    | 84              |
| Meramec            | 18 630            | 278           | Deciduous Forest       | 158                | 8.94                     | 95              |
| The April 28–May 1, 2017, frontal storms in the Southern Midwestern US |
| Cypress            | 4104              | 73            | Pasture/Hay            | 608                | 19.12                    | 121             |
| WF San Jacinto     | 9180              | 80            | Pasture/Hay            | 554                | 17.86                    | 139             |
| Navidad            | 10 260            | 86            | Pasture/Hay            | 346                | 11.30                    | 112             |
| Lower Sabine       | 17 856            | 86            | Shrub/Scrub           | 501                | 12.93                    | 175             |
| The August 25–31, 2017, Hurricane Harvey storms in Texas |
| Rockfish           | 1134              | 92            | Cultivated Crops      | 334                | 18.24                    | 109             |
| Black              | 1450              | 80            | Cultivated Crops      | 222                | 11.58                    | 76              |
| Neuse              | 5304              | 29            | Cultivated Crops      | 316                | 12.91                    | 107             |
| Cape Fear          | 9120              | 28            | Cultivated Crops      | 496                | 12.76                    | 102             |
| Lumber             | 9790              | 118           | Cultivated Crops      | 333                | 11.78                    | 128             |
| The September 13–17, 2018, Hurricane Florence storms in the Carolinas |
| Rockfish           | 1134              | 92            | Cultivated Crops      | 58                 | 7.14                     | 27              |
| Black              | 1450              | 80            | Cultivated Crops      | 49                 | 5.47                     | 29              |
| Neuse              | 5304              | 29            | Cultivated Crops      | 119                | 13.79                    | 34              |
| Cape Fear          | 9120              | 28            | Cultivated Crops      | 109                | 6.82                     | 62              |
| Lumber             | 9790              | 118           | Cultivated Crops      | 66                 | 5.44                     | 42              |
| The September 4–6, 2019, Hurricane Dorian storms in the Carolinas |

The table above shows some of the key characteristics of the selected storms over nested watersheds. The storms are categorized by the month and location, such as the September 21–23, 2016, frontal storms in Iowa, the August 25–31, 2017, Hurricane Harvey storms in Texas, and the September 13–17, 2018, Hurricane Florence storms in the Carolinas. Each storm is represented by its domain size, elevation, major land cover, total rainfall, peak hourly rainfall, and duration. The table provides a comprehensive view of the rainfall characteristics and the geographical context of these storms.
and HRRRv3 forecasts available since 2018. These data are available to the public via the HRRR Archive managed by the University of Utah (Blaylock et al 2017). The NWS’ operational hydrologic modeling platform NWM first downscales the HRRR forecasts from 3 km to 1 km by bilinear interpolation methods, and uses the downscaled forecast as forcing input to generate short-range hydrological forecasts at lead times ranging from 1 h to 18 h (NOAA 2016).

For our analyses, we have used the recent HRRR forecasts, namely, the HRRRv2 forecasts for the frontal storms and hurricane Harvey and the HRRRv3 forecasts for the hurricane Florence. In order to provide results relevant for NWM, we have downscaled the hourly HRRR forecasts from 3 km to 1 km implementing identical downscaling methods used in NWM.

2.3. The reference observation data MRMS-GC

The gauge-corrected Multi-Radar/Multi-Sensor (MRMS-GC) quantitative precipitation estimates (QPE) are produced operationally at the National Centers for Environmental Prediction (NCEP) and distributed to NWS forecast offices and several external agencies. The MRMS-GC ingests radar reflectivity data from the WSR-88D network resulting in a spatial domain covering the CONUS. Radar-based estimates of precipitation are adjusted using a network of approximately 7000 gauges from the Hydrometeorological Automated Data System (HADS) network (Kim et al 2009). Data from the NWP model Rapid Update Cycle (RUP) are used in quality controlling the rain gauge data. The resulting MRMS-GC QPEs are hourly rainfall rates with a spatial resolution of 0.01° resolution. Detailed information on MRMS-GC can be found in REFs. The MRMS-GC product has a history of use as the reference product for validation of models and satellite products (e.g., Gourley et al 2017, Gebregiorgis et al 2017, Smalley et al 2017, Seo et al 2018). In this study, we used the MRMS-GC QPEs as reference to validate the HRRRv2 and HRRRv3 forecasts.

2.4. Evaluation methods

The evaluations of HRRR forecasts were carried out for various quality features. These include: (1) hourly comparisons of domain-averaged precipitation, fraction of domain covered by rain, and conditional coefficient of variation (i.e. standard deviation of rainfall within the domain normalized by domain-averaged precipitation—only taking into account pixels where rainfall occurred), (2) spatial maps of accumulated rainfall over the entire storm duration, and (3) The geometric features of accumulated rainfall using the Amplitude Structure Location (SAL) method.

The SAL method (Wernli et al 2008) evaluates storm features (size, variability, and location) by identifying precipitation objects in both the forecast and the observed storm at a given time, and decomposing differences (i.e. errors) into three components. The errors are normalized by the size of the domain and domain-wide
accumulation such that results from different domain sizes and rainfall accumulation can be compared. The amplitude (A) component (between \(-2\) and \(+2\)) corresponds to the normalized difference of the domain-averaged precipitation values. \(A = 0\) denotes perfect forecast, \(A > 0\) indicates overestimation, \(A < 0\) indicates underestimation of domain-averaged precipitation. The values \(A = 0.4, 0.67, 1\) indicate overestimation of domain-averaged precipitation by a factor of 1.5, 2, and 3, respectively. Along the same lines, \(A = -1\) indicates underestimation by a factor of 3. The Structure (S) component (between \(-2\) and \(+2\)) captures information about the size and shape of precipitation objects. The value \(S = 0\) indicates perfect field, \(S > 0\) indicates widespread precipitation forecast in a situation of localized events, while \(S < 0\) indicates too small precipitation objects or too peaked objects, or a combination of these factors. The location component \(L\) (between 0 and 2) consists of two parts. One part measures location differences in centers of mass for the domain-wide observed and forecast fields; the other part accounts for location differences of all objects weighted by their integrated precipitation. \(L = 0\) indicates a forecast field, where both the center of mass as well as the averaged distance between the

![Figure 2](image_url)

**Figure 2.** Event overview of frontal system driven storms in Iowa: (a) Time series of mean areal precipitation; (b) Time series of fractional area coverage; (c) Time series of conditional coefficient of variation.
precipitation objects and the center of mass agree with the observations. The ASL equations are provided in Wernli et al (2008).

3. Results and discussion

3.1. The september 21–23, 2016, frontal storms in Iowa

A tropical air mass interacting with a stationary front triggered several rounds of heavy storms in northern Iowa and southern Minnesota during September 21–23, 2016 (Seo et al 2018). Figure 1(a) shows spatial map of accumulated rainfall accumulation during the entire storm period over a selected domain. Most of the domain was covered by rain. There was a supercell, where rainfall ranged from 250 mm to 300 mm at the core of the cell and from 200 mm to 250 mm at the peripheries. Let us first examine how the spatial pattern of the storm was portrayed by the forecasts. The spatial maps of accumulated rainfall for HRRR forecasts are shown in figures 1(b)–(e). The forecasts accurately showed rainfall occurrence for most of the domain. However, there were differences in the spatial pattern of rainfall: the 1-h lead forecast captured the observed supercell but overestimated its spatial coverage, the 6-h and 18-h lead forecasts missed the supercell, and the 12-h lead forecast displaced the supercell to the northeast and also increased its size.

In addition to spatial maps, hourly time series of precipitation patterns were evaluated. For this purpose, the domain-averaged precipitation values were evaluated at each hour during the storm period in figure 2(a). According to the observations, the storm lasted 72 h, and had a wave pattern with three peaks. The first and smaller peak was at hour 7 (counting from the beginning of the storm) with a peak rainfall of 1.12 mm h$^{-1}$, the second and largest peak at hour 34 with a peak rainfall of 5.78 mm h$^{-1}$, and the third and moderate peak occurred at hour 58 with a peak rainfall of 2.74 mm h$^{-1}$. Overall, the forecasts captured well the wave pattern with three peaks. However, the forecasts varied in how they forecasted the hydrographs for the peak events.

The second and largest peak (recall peak rainfall of 5.78 mm h$^{-1}$ for observed rainfall) was well-forecasted with only errors of $+9\%$ (1-h lead forecast), $-15\%$ (6-h), $+13\%$ (12-h), $+15\%$ (18-h). However, there were differences among the forecasts in the rising and falling limbs of the hyetograph for this event: the 1-h lead forecast displaced the storm by about 4 h (rising/falling limbs and peak were 4 h late than the observations); the 6-h lead forecast correctly reproduced the rising and falling limbs; the 12-h lead forecast showed unusual spike at the beginning of the storm, but captured well the remaining portion of the hyetograph.; the 18-h lead forecast missed almost the rising limb of the hyetograph, but captured the recession limb. For the third and moderate peak (recall peak rainfall of 2.74 mm h$^{-1}$ for observed rainfall), the 1-h lead forecast overestimated the peak (by about 39%), while the remaining forecasts substantially underestimated the peak (by 70% for 6-h, by 42% for 12-h, by 63% for 18-h lead forecasts). For this event, the 1-h lead forecast captured well the hydrograph with no delays, while the 6-h lead forecast missed the hydrograph. For the first and smallest-peak event (recall peak rainfall of 1.12 mm h$^{-1}$ for observed rainfall), the 1-h forecast best matched the observation albeit with some underestimation, while the remaining forecasts resulted in large underestimation of the peak.
Next, we evaluated the fraction of the domain covered by rain ($f_{\text{rain}}$) by dividing the number of pixels (1-km grids) with rainfall intensity larger than zero to the total number of pixels within the domain. Figure 2(b) shows the time series of $f_{\text{rain}}$ for the observation and forecasts. According to the observations, the temporal pattern of $f_{\text{rain}}$ follows the temporal pattern of the domain-averaged rainfall discussed above, indicating that the bigger the storms the larger the area they cover. The forecasts reproduced this temporal pattern of $f_{\text{rain}}$. However, there were some differences in the actual magnitudes of $f_{\text{rain}}$: for the second and largest-peak event, all the forecasts underestimated the areal coverage; for the third and moderate-peak event, the 1-h lead forecast agreed well with the observations, while the remaining lead-time forecasts underestimated the spatial coverage; and for the first and smallest-peak event, the 1-h lead forecast did well, while the remaining lead forecasts overestimates the areal coverage. Overall, the performance of the lead-time forecasts in forecasting the areal coverage is similar to their performance in forecasting the domain-averaged precipitation.

**Figure 4.** Event overview of frontal system driven storms in Southern Midwestern US: (a) Time series of mean areal precipitation; (b) Time series of fractional area coverage; (c) Time series of conditional coefficient of variation.
The 12-h and 18-h lead forecasts tended to underestimate precipitation during some hours. A convergence of two fronts produced thunderstorms that yielded abundant precipitation over the Southern Midwestern US during the period April 28th - May 1st, 2017 (Heimann et al. 2018). In figure 3, spatial maps of accumulated rainfall forecasts were compared to those of observations to get a first impression of the quality of the forecasts. According to the observations (figure 3(a)), there was a wide strip of moderate-intensity rain in the southwest-northeast direction, with large rainfall accumulation around the center of the domain. The 1-h lead forecast (figure 3(b)) placed the storm to the northwest, and produced quite a large number of scattered, localized, high-intensity rainfall events than what the observations show. The 6-h lead-time forecast (figure 3(c)) produced more-scattered, high-intensity rainfall fields in the northeast of the domain, while significantly underestimating the large rainfall field at the center of the field, and placing the storms in the southwest slightly to the north. The 12-hour lead forecast (figure 3(d)) tended to underestimate observed precipitation across the domain. The 18-h lead forecast (figure 3(e)) captured the large rainfall observed around the center of the domain but showed some scattered high-intensity rainfall fields in the southern part of the domain which did not appear in the observations.

The time series of observed domain-averaged rainfall (figure 4(a)) shows a major rain event (during the period of hours 24 to 70) that was characterized by a gradually varying rainfall accumulation. Figure 4 shows overall good agreement between the time series of domain-averaged precipitation observation and forecasts. The 12-h and 18-h lead forecasts tended to underestimate precipitation during some hours (e.g., hours 50–60), and the 1-h lead forecast tended to overestimate in some hours (e.g., hours 53–57).

The time series of \( f_{\text{rain}} \) (figure 4(a)) was bimodal and differed from the unimodal pattern of domain-averaged precipitation. Apparently, the light rain events during the first 24 h (which are not noticeable in the domain-averaged precipitation time series) covered large areas of the domain (up to 50%) causing one of the peaks in the \( f_{\text{rain}} \) time series. The major rainy period (from hour 34 to 65), which covered 60% to 100% of the domain, caused the second peak. Therefore, both light and heavy rain events covered large areas of the domain. How did the forecasts saw this bimodal pattern? All the forecasts captured well this bimodal pattern, but tended to underestimate the actual magnitudes of \( f_{\text{rain}} \) during both peaks.

Figure 4(c) shows that the conditional coefficient of variation of observed rainfall fields is mostly constant around 1, however, the forecasts produced a value of 2, indicating that the forecasts produced more spatially variable rainfall fields (doubling the spatial standard deviation) than what the observations show. The higher spatial variability of the forecasted fields compared to the observations is not surprising as the forecasts produce relatively large number of scattered high-intensity rainfall fields (see figures 3(b)–(e)).

Finally, we evaluated the conditional coefficient of variation of rainfall (\( CV_{\text{con}} \)) defined as standard deviation of rainfall within the domain normalized by area-averaged rainfall, only considering pixels with rainfall exceeding zero. According to the observations, \( CV_{\text{con}} \) varied mostly between 0.9 and 2 and its temporal pattern was not associated with the magnitude of the storm. All the forecasts reproduced the temporal pattern of \( CV_{\text{con}} \) observed in the observed rainfall fields, however, they showed higher values at almost all hours, indicating that the forecasted fields had higher spatial variability than the observed rainfall fields.

### 3.2. The april 28-may 1, 2017, frontal storms in the Southern Midwestern US

A convergence of two fronts (a warm front that extended from southeast Missouri across west-central Arkansas into south-central Oklahoma moving north, and a tropical moisture from the Gulf of Mexico moving north) produced thunderstorms that yielded abundant precipitation over the Southern Midwestern US during the period April 28th - May 1st, 2017 (Heimann et al. 2018). According to the observations, \( CV_{\text{con}} \) varied mostly between 0.9 and 2 and its temporal pattern was not associated with the magnitude of the storm. All the forecasts reproduced the temporal pattern of \( CV_{\text{con}} \) observed in the observed rainfall fields, however, they showed higher values at almost all hours, indicating that the forecasted fields had higher spatial variability than the observed rainfall fields.

![Figure 5. Precipitation accumulation map (7-days) of Hurricane Harvey driven storms in Texas: (a) Navidad River basin; (b) Cypress Creek basin; (c) West Fork San Jacinto River basin; (d) Lower Sabine River basin; (e) Observation (b) Forecasts at Lead Time 1 H; (c) Forecasts at Lead Time 6 H; (d) Forecasts at Lead Time 12 H; (e) Forecasts at Lead Time 18 H.](image-url)

The time-series of observed domain-averaged rainfall was bimodal and differed from the unimodal pattern of domain-averaged precipitation. Apparently, the light rain events during the first 24 h (which are not noticeable in the domain-averaged precipitation time series) covered large areas of the domain (up to 50%) causing one of the peaks in the \( f_{\text{rain}} \) time series. The major rainy period (from hour 34 to 65), which covered 60% to 100% of the domain, caused the second peak. Therefore, both light and heavy rain events covered large areas of the domain. How did the forecasts saw this bimodal pattern? All the forecasts captured well this bimodal pattern, but tended to underestimate the actual magnitudes of \( f_{\text{rain}} \) during both peaks.

Figure 4(c) shows that the conditional coefficient of variation of observed rainfall fields is mostly constant around 1, however, the forecasts produced a value of 2, indicating that the forecasts produced more spatially variable rainfall fields (doubling the spatial standard deviation) than what the observations show. The higher spatial variability of the forecasted fields compared to the observations is not surprising as the forecasts produce relatively large number of scattered high-intensity rainfall fields (see figures 3(b)–(e)).
3.3. The August 25-31, 2017, Hurricane Harvey Storms in Texas

Hurricane Harvey was the most significant cyclone rainfall event in United States history (Blake and Zelinsky 2017). On August 25th, Hurricane Harvey made landfall near Rockport, Texas. The forward motion of Hurricane Harvey slowed down as it moved inland, producing tremendous rainfall amounts in southeastern Texas and southwestern Louisiana (Watson et al 2018). Figure 5 shows the spatial map of accumulated rainfall according to observations. Hurricane Harvey brought very large amounts of rainfall to the region, with rainfall accumulation exceeding 1,000 mm in some large areas around the center of the domain. The 6-h, 12-h, and 18-h lead forecasts captured well the spatial pattern of observed precipitation, while the 1-h lead forecast misplaced some part of the storm further north.

The time-series of domain-averaged accumulated precipitation is shown in figure 6(a). The hurricane brought strong storms through the entire storm duration, about five days. There was a high degree of agreement.
between the temporal pattern of forecasts and observations. In terms of actual magnitudes, all the forecasts produced mean-averaged values that are closer to the observations, except the 1-h lead forecast which resulted in overestimation throughout the entire storm period.

Figure 6(b) shows that the time series of $f_{\text{rain}}$ had a wide bell-shaped pattern, showing significant areas of the domain covered by rain at all times. There was a high degree of agreement between the forecasts (including the 1-h forecast) and observations in terms of hourly time series of fraction of domain covered by rain throughout the storm duration.

According to figure 6(c), the conditional coefficient of variation of observed rainfall fluctuates within 1 and 2, and does not show association with magnitudes domain-averaged rainfall. The $CV_{\text{con}}$ values for the forecasts were closer to those for the observed values: for almost half of the storm duration, the forecasted $CV_{\text{con}}$ values were slightly higher than the observed values (meaning the rainfall fields are spatially more variable), while for the other half of the storm hours the forecasted values were similar to the observed values.

3.4. The september 13–17, 2018, Hurricane Florence storms in the Carolinas

Hurricane Florence was the first major hurricane of the 2018 Atlantic hurricane season. On September 14th, 2018, Hurricane Florence moved inland, and the slowing down forward motion produced large amounts of rainfall across the Carolinas (Feaster et al. 2018). Figure 7 shows that Hurricane Florence brings large amount of rainfall to the eastern part of the domain with very large amounts (>700 mm) in some localized areas. All the forecasts captured this rainfall pattern, but produced larger areas with very high rainfall amounts than what the observations show. The overestimation by the rainfall forecasts holds for all the lead-time forecasts.

The temporal variability of the domain-averaged precipitation during the storm period was examined in figure 8. The time series rainfall intensity (figure 8(a)) shows that the hurricane brought large amounts of rainfall throughout the storm period, about 4 days, resulting in a wide bell-shaped pattern. There was a good agreement between the temporal variability of forecasts and observations. However, the forecast overestimated precipitation during the peak hours, consistent with the spatial maps discussed above.

As shown in figure 8(b), the fraction of the domain covered by rain also exhibited pronounced bell-shaped pattern. This pattern was well reproduced by all the forecasts. However, the forecasts tended to slightly underestimate $f_{\text{rain}}$ throughout the storm period. As displayed in figure 8(c), the conditional coefficient of variation fluctuated within 1.5 ± 1, and showed no association with rainfall amount. The forecasts produced this pattern, but tended to slightly overestimate CV.

3.5. The september 4–6, 2019, Hurricane Dorian storms in the Carolinas

Hurricane Dorian was the first major hurricane of the 2019 Atlantic hurricane season and was one of the most powerful hurricanes on record in the open Atlantic region (Cerrai et al. 2019). Instead of landing on the Carolinas, Hurricane Dorian swiped the coastal area of the North Carolina and South Carolina, bringing moderate precipitation. The observed rainfall covers the central and eastern parts of the region, with higher rainfall values forming stripes (figure 9). All the forecasts detected the occurrence and severity of rainfall,
however, there were differences in the spatial location of high rainfall values, with the forecasts producing more scattered patterns.

According to the time-series of domain-averaged rainfall (figure 10(a)), all the forecasts captured well the temporal dynamics of observed rainfall, albeit with some overestimation. While the 1-h lead forecast tended to have relatively larger bias coverage (figure 10(b)), the observed rainfall shows a bimodal pattern, with one relatively smaller peak (around hour 20), and another much larger peak (during the period of hours 24 to 70) that was characterized by a gradually varying accumulation. All the forecasts reproduced the temporal dynamics of $f_{\text{rain}}$, with slight underestimation bias. Joint analysis of figures 10(a) and (b) reveals that the forecasted fields underestimated the areal coverage of conditional coefficient of variation (figure 10(c)), the observed rainfall has CV within the range between 1 and 2, whereas the forecasted fields have corresponding values much higher than this indicating that the forecasted fields produced ‘rougher’ spatial structure than observed.

Figure 8. Event overview of Hurricane Florence driven storms in North Carolina and South Carolina: (a) Time series of mean areal precipitation; (b) Time series of fractional area coverage; (c) Time series of conditional coefficient of variation.
3.6. Evaluation of precipitation geometric features

Here, we evaluate the ability of the forecasts to reproduce the geometric features of the accumulated precipitation, in terms of the Amplitude, Structure, and Location components. Figure 11 shows the ASL results for each storm, as a function of forecast lead time. The Amplitude values for both hurricanes are positive at all lead times, indicating overestimation of domain-averaged rainfall. The Amplitude for Hurricane Florence is constant around 0.20 at all lead times, indicating that the forecasts overestimate the domain-averaged rainfall by a factor of 0.6. Similarly, the Amplitude of Hurricane Dorian is constant around 0.3 from lead time 1 h to 12 h but drops to 0.1 to 0.2 thereafter. The Amplitude of Hurricane Harvey shows some variation with respect to lead time: it is generally higher at short lead times (1 h to 6 h), gets close to zero between 8 h and 13 h lead times, and slightly increases beyond 13 h. The Amplitude of Frontal system in Iowa is negative at all lead times indicating underestimation of domain-averaged rainfall, except for the 1-h lead time forecast which had a positive amplitude. For this storm, the negative amplitudes are generally higher at lead times from 4 h to 10 h (reaching amplitude of −0.5 at 6-h lead time, corresponding to underestimation of 1.5), and slightly improve beyond the 10-h lead forecast. The Amplitude of Frontal system in Southern Midwestern US was slightly positive during short lead times (1 h–10 h) but showed larger negative amplitude for lead times exceeding 10 h.

The Structure values for Hurricane Harvey and Hurricane Florence are close to zero, indicating that the forecasted rainfall fields have comparable spatial structure with the observed rainfall fields. The Structure values for Hurricane Dorian are slightly larger and negative up to lead time of 10 h, and get larger as the lead time increased further. However, the Structure values for both frontal storms are relatively large and negative, indicating that the forecasted rainfall fields contain too small precipitation objects or too peaked objects, or a combination of these factors, compared to the observed rainfall fields.

The Location values are very large for the frontal storm in Iowa, for lead times exceeding 5 h, indicating large misplacement of storm location. For lead times under 5 h, the frontal storm in Iowa had smaller storm location error. The Location values for Hurricane Harvey are the smallest at all lead times, indicating that the forecasts reproduced well the storm location. Hurricane Florence and Hurricane Dorian have constant but moderate locational value (around 0.18) at all lead times. The Frontal storm in Southern Midwestern US had moderate locational errors which depend on the lead time: the locational errors are relatively high at short lead times (1 h–7 h), get low for lead times of 8 h–13 h, and increase for lead times of 14 h and beyond.

3.7. Effect of spatial scale on bias estimation

In the above sections, we quantified the bias in rainfall forecasts averaged over a large domain. In hydrological simulations, where HRRR forecasts are used as input into a hydrologic model, estimates of bias are required over smaller spatial scales commensurate with the model computational grid. Here, we quantify the bias in the forecasts at smaller spatial scales, from 2 km × 2 km to increasingly coarser spatial scales all the way up to 128 km × 128 km. We perform sampling experiment, where we divide the domain into grids according to the spatial scale of interest and compare the grid-averaged forecast with the corresponding grid-averaged observed rainfall.
Figure 12 presents the relative bias (defined as: (forecasted rainfall—observed rainfall)/observed rainfall) distribution for each of the storms. The relative bias at the $2 \text{ km} \times 2 \text{ km}$ grids mostly varies in the range $-1$ to $+1$ (meaning up to 100% overestimation and up to 100% underestimation). Therefore, the bias estimates obtained at large (hundreds of km grids) are not transferable to bias estimates at smaller grids, indicating that local bias correction are more preferable over global bias correction methods. In addition, the bias estimates at $2 \text{ km} \times 2 \text{ km}$ grids are highly variable from one grid to another, indicating that the bias estimates at one grid is not transferable to another grid even within the same region. The relative bias does not show major reduction as the grid size increases.

To further examine the spread of the relative bias, we selected a few real watersheds from each domain, and calculated the relative bias for each watershed. The relative bias results for the 6-h lead forecast are shown in table 2. For the frontal storm in Southern Midwestern US, while the bias for the entire domain is about $+6\%$, the
bias for the watersheds varies from $-12\%$ for Illinois ($4,283 \text{ km}^2$) to $+47\%$ for Flint ($328 \text{ km}^2$). For the hurricane in Texas, the bias for the entire domain is $+19\%$, while the bias for the watershed varies from $-24\%$ for Navidad ($3,628 \text{ km}^2$) to $+56\%$ for WF San Jacinto ($2,801 \text{ km}^2$). For Hurricane Florence, the bias for the entire domain is

![Figure 11. SAL analysis results of five selected extreme events at all lead times (1–18 H): (a) Amplitude component; (b) Structure component; (c) Location component.](image-url)
+24%, while the bias ranges over a large range: -17% for Neuse (2,759 km²) to +110% for Lumber (4,540 km²). For Hurricane Dorian in the same region, the relative bias for the entire domain is about ±32%, while the bias of individual watersheds ranges from ±12% (Neuse) to ±81% (Black). These results underline that the bias estimates have large spatial variations, and therefore large-scale bias estimates are not transferable to smaller scales.

4. Summary and conclusion

This study evaluated the accuracy of short-range (lead times ranging from 1 h to 18 h) forecasts, for five extreme events in the United States which covered two frontal storms and two hurricanes: the September 21–23, 2016, frontal storms in Iowa, (2) the April 28-May 1, 2017, frontal storms in the Southern Midwestern US, (3) the August 25–31, 2017, Hurricane Harvey storms in Texas, (4) the September 13–17, 2018, Hurricane Florence storms, and (5) the September 4–6, 2019, Hurricane Dorian storms in the Carolinas. The basis of the investigation was the HRRR operational forecasts, which are used as input into the National Weather Service’s...
National Water Model (NWM). Evaluation of the forecasts was carried out by comparison with high-quality and independent rainfall observational products known as the gauge-corrected Multi-Radar/Multi-Sensor (MRMS-GC). The main results are summarized as follows.

First, the temporal variability of precipitation during the storm period was examined. There was a good agreement between area-averaged forecasts and observations, on an hourly scale. However, the forecasts were mostly biased. The forecasts tend to overestimate rainfall for both hurricanes. However, the forecasts tend to underestimate the frontal storm in Iowa but produced almost unbiased estimates for the Southern Midwestern US.

Examination of the spatial precipitation pattern was additionally carried out. The forecasts were able to capture the spatial pattern of hurricanes, albeit with overestimation. However, the forecasts produced too many, localized, high-rain intensities for the frontal storms. In addition, the forecasts have difficulty locating the single supercell for the frontal storm in Iowa.

With regard to the effect of lead time, the 1 h lead forecast had lower accuracy compared to the other lead-time forecasts. For lead times ranging from 2 h to 18 h, there was not much systematic difference in accuracy among the various lead-time forecasts.

The bias in the forecast were also examined at different spatial scales, ranging from 2 km × 2 km all the way to 128 km × 128 km. The bias estimates for the small spatial scale varied quite a lot, mostly within the range of −100% to +100%, indicating that the bias estimates obtained at large scale (hundreds of km grids) are not

| Watershed Area (km²) | Relative Bias (Ratio) |
|----------------------|-----------------------|
| Full Domain 49 140    | −0.344                |
| Cedar 17 434         | −0.438                |

Table 2. Relative bias at Lead Time 6 H calculated over full domain and specific watersheds.

| Watershed Area (km²) | Relative Bias (Ratio) |
|----------------------|-----------------------|
| Full Domain 176 022  | 0.056                 |
| Meramec 5568         | 0.189                 |
| Illinois 4283        | −0.116                |
| Eleven Point 3115    | 0.380                 |
| Gasconade 2676       | 0.261                 |
| Big 2513             | 0.342                 |
| Flint 328            | 0.469                 |

| Watershed Area (km²) | Relative Bias (Ratio) |
|----------------------|-----------------------|
| Full Domain 127 575  | 0.187                 |
| Lower Sabine 6844    | 0.085                 |
| Navidad 3628         | −0.243                |
| WF San Jacinto 2801  | 0.560                 |
| Cypress 1960         | 0.486                 |

| Watershed Area (km²) | Relative Bias (Ratio) |
|----------------------|-----------------------|
| Full Domain 99 828   | 0.244                 |
| Lumber 4540          | 1.108                 |
| Cape Fear 2811       | 0.446                 |
| Neuse 2759           | −0.174                |
| Rockfish 801         | 0.944                 |
| Black 757            | 0.020                 |

| Watershed Area (km²) | Relative Bias (Ratio) |
|----------------------|-----------------------|
| Full Domain 99 828   | 0.316                 |
| Lumber 4540          | 0.554                 |
| Cape Fear 2811       | 0.480                 |
| Neuse 2759           | 0.117                 |
| Rockfish 801         | 0.170                 |
| Black 757            | 0.812                 |
applicable to bias estimates at smaller spatial scales, and vice versa. The bias did not also show significant reduction as the rainfall averaging grid increases from 2 km × 2 km all the way to 32 km × 32 km.

In conclusion, the results of our investigation show that the forecasts captured well the temporal variability of observed precipitation, indicating that the HRRR forecasts provide relatively reliable forecasts. In comparison, the forecasts have better accuracy for predicting hurricanes compared to frontal storms, particularly those frontal storms with single super cells. Our results also show that the 1-h lead forecasts showed generally lower accuracy than the other lead-time forecasts.

Finally, we point out that, although the selected storm cases are interesting from meteorological perspective, they are small in number. Thus, the findings of this study can only provide a first insight into the accuracy of HRRR forecasts for extreme precipitation. Additional analysis involving more storm cases and mechanistic approaches will have to be carried out in order to generalize the results.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

The HRRR archive is described and can be publicly accessed in Blaylock et al (2017).

The MRMS-GC data is described and can be publicly accessed in Zhang et al (2011).

ORCID iDs

Haowen Yue https://orcid.org/0000-0002-4068-8810

References

Bartholmes J and Todini E 2005 Coupling meteorological and hydrological models for flood forecasting Hydrol. Earth Syst. Sci. Discuss. 9 333–46
Benjamin S G, Weygandt S S, Brown J M, Hu M, Alexander C R, Smirnova T G and Lin H 2016 A North American hourly assimilation and model forecast cycle: the rapid refresh Mon. Weather Rev. 144 1669–94
Berne R K, Fossell K R, Ha S Y, Hacker J P and Snyder C 2013 Increasing the skill of probabilistic forecasts: understanding performance improvements from model-error representations Mon. Weather Rev. 143 1295–320
Blake E S and Zelinsky D A 2017 Cloud archiving and data mining of high resolution rapid refresh model output Computers and Geosciences 109 43–50
Bytheway J L and Kummerow C D 2015 Toward an object-based assessment of high-resolution forecasts of long-lived convective precipitation in the central US Journal of Advances in Modeling Earth Systems 7 1248–64
Bytheway J L, Kummerow C D and Alexander C A 2017 A features-based assessment of the evolution of warm season precipitation forecasts from the HRRR model over three years of development Weather and Forecasting 32 1841–56
Cai H and Dumais R E Jr 2015 Increasing the skill of probabilistic forecasts: understanding performance improvements from model-error representations Mon. Weather Rev. 143 1295–320
Ceppi A, Ravazzani G, Corbari C, Salerno R, Meucci S and Mancini M 2014 Real-time drought forecasting system for irrigation management Hydrolog. Earth Syst. Sci. 18 3353–66
Chou M D and Suarez M J 1994 An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 85
Collischonn W, Tucci C E M, Clarke R T, Chou S C, Guilhon L G, Cataldi M and Allasia D 2007 Medium-range reservoir inflow predictions based on quantitative precipitation forecasts J. Hydrol. 344 112–22
Cuo L, Pagano T C and Wang Q J 2011 A review of quantitative precipitation forecasts and their use in short-to-medium range streamflow forecasting Journal of Hydrometeorology 12 713–28
Feaster T D, Weaver J C, Gotvald A J and Kolb K R 2018 Preliminary Peak Stage and Streamflow Data for Selected US Geological Survey Streamgaging Stations in North and South Carolina for Flooding Following Hurricane Florence, September 2018 (No. 2018-1172) (USGS National Center, Sunrise Valley Drive, Reston, Virginia: US Geological Survey)
Gebregiorgis A S, Kistetter P E, Hong Y E, Carr N J, Gourley J J, Petersen W and Zheng Y 2017 Understanding overland multisensor satellite precipitation error in TMPA-RT products Journal of Hydro meteorology 18 285–306
Gourley J J et al 2017 The FLASH Project: improving the tools for flash flood monitoring and prediction across the United States Bulletin of the American Meteorological Society 98 361–72
Gowan T M, Steenburgh W J and Schwartz C S 2018 Validation of mountain precipitation forecasts from the convection-permitting NCM ensemble and operational forecast systems over the Western United States Weather and Forecasting 33 739–65
Heimann D C, Holmes R R Jr and Harris T E 2018 Flooding in the Southern Midwestern United States, April–May 2017 (No. 2018-1004). HRRR data archive (https://doi.org/10.7278/S3(QZS5)
Ikeda K, Steiner M, Pinto J and Alexander C 2013 Evaluation of cold-season precipitation forecasts generated by the hourly updating high-resolution rapid refresh model Weather and Forecasting 28 921–39
Ikeda K, Steiner M and Thompson G 2017 Examination of mixed-phase precipitation forecasts from the high-resolution rapid refresh model using surface observations and sounding data Weather and Forecasting 32 949–67
Kim Dongsoo, Nelson Brian and Seo Dong-Jun 2009 Characteristics of Reprocessed Hydrometeorological Automated Data System (HADS) Hourly Precipitation Data Weather and Forecasting 24 1287–96
Mlawer E J, Taubman S J, Brown P D, Iacono M J and Clough S A 1997 Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave Journal of Geophysical Research: Atmospheres 102 16663–82
Mors R E 2003 Assessing the needs of users of warm season quantitative precipitation forecasts in Colorado Symp. on Impacts of Water Variability: Benefits and Challenges. Preprints, 83rd AMS Annual Meeting pp 9–13
Nakanishi Mikio and Niino Hiroshi 2004 An Improved Mellor–Yamada Level-3 Model with Condensation Physics: Its Design and Verification Boundary-Layer Meteorology 112 1–31
NOAA (National Oceanic and Atmospheric Administration) 2016 National Water Model: Improving NOAA’s Water Prediction Services https://water.noaa.gov/documents/wrn-national-water-model.pdf
Pinto James O, Grim Joseph A and Steiner Matthias 2015 Assessment of the High-Resolution Rapid Refresh Model’s Ability to Predict Mesoscale Convective Systems Using Object-Based Evaluation Weather and Forecasting 30 892–913
Saaavedra Valeriano O C, Koike T, Yang K, Graf T, Li X, Wang L and Han X 2010 Decision support for dam release during floods using a distributed biosphere hydrological model driven by quantitative precipitation forecasts Water Resour. Res. 46 W10544
Seo B C, Quintero F and Krajewski W F 2018 High-resolution QPF uncertainty and its implications for flood prediction: a case study for the eastern Iowa flood of 2016 Journal of Hydrometeorology 19 1289–304
Smalley M, Kirtteter P E and L’Ecuyer T 2017 How frequent is precipitation over the contiguous United States? Perspectives from ground-based and spaceborne radars Journal of Hydrometeorology 18 1657–72
Smirnova T G, Brown J M and Benjamin S G 1997 Evolution of soil moisture and temperature in the MAPS/RUC assimilation cycle. 13th Conf. on Hydrology (Long Beach, CA) 172–75 Perprints
Sukovich E M, Ralph F M, Barthold F E, Reynolds D W and Novak D R 2014 Extreme quantitative precipitation forecast performance at the weather prediction center from 2001 to 2011 Weather and Forecasting 29 894–911
Thompson G, Field P R, Rasmussen R M and 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 5095–115
Watson K M, Harwell G R, Wallace D S, Welborn T L, Stengel V G and McDowell J S 2018 Characterization of Peak Streamflows and Flood Inundation of Selected Areas in Southeastern Texas and Southwestern Louisiana from the August and September 2017 Flood Resulting from Hurricane Harvey (No. 2018-5070) (USGS National Center12201 Sunrise Valley DriveReston, Virginia 20192: US Geological Survey)
Wernli H, Paulat M, Hagen M and Frei C 2008 SAL—a novel quality measure for the verification of quantitative precipitation forecasts Mon. Weather Rev. 136 4470–87
Zhang J, Howard K, Langston C, Kaney B, Q Y, Tang I and Arthur A 2016 Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: initial operating capabilities Bull. Am. Meteorol. Soc. 97 621–38
Zhang J, Howard K, Langston C, Vasiloff S, Kaney B, Arthur A and Seo D J 2011 National mosaic and multi-sensor QPE (NMQ) system: description, results, and future plans Bull. Am. Meteorol. Soc. 92 1321–38