Flash Flood Verification: Pondering Precipitation Proxies

GREGORY R. HERMAN AND RUSS S. SCHUMACHER

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

(Manuscript received 5 May 2018, in final form 10 September 2018)

ABSTRACT

Quantitative precipitation estimate (QPE) exceedances of numerous different heavy precipitation thresholds—including spatially varying average recurrence interval (ARI) and flash flood guidance (FFG) thresholds—are compared among each other and against reported and warned flash floods to quantify existing deficiencies with QPEs and to identify best practices for using QPE for flash flood forecasting and analysis. QPEs from three different sources—NCEP Stage IV Precipitation Analysis (ST4), Climatology Calibrated Precipitation Analysis (CCPA), and Multi-Radar Multi-Sensor (MRMS) QPE—are evaluated across the United States from January 2015 to June 2017. In addition to evaluating different QPE sources, threshold types, and magnitudes, QPE accumulation interval lengths from hourly to daily are considered. Systematic errors with QPE sources are identified, including a radar distance dependence on extreme rainfall frequency in MRMS, spurious occurrences of locally extreme precipitation in the complex terrain of the West in ST4, and insufficient QPEs for many legitimate heavy precipitation events in CCPA. Overall, flash flood warnings and reports corresponded to each other far more than any QPE exceedances. Correspondence between all sources was at a maximum in the East and worst in the West, with ST4, CCPA, and MRMS QPE exceedances locally yielding maximal correspondence in the East, Plains, and West, respectively. Surprisingly, using a fixed 2.5 in. $(24\,\mathrm{h})^{-1}$ proxy outperformed shorter accumulation exceedances and the use of ARIs and FFGs. On regional scales, different ARI exceedances achieved superior performance to the selection of any fixed threshold; FFG exceedances were consistently too rare to achieve optimal correspondence with observed flash flooding.

1. Introduction

Flash flooding is both a highly complex and immensely important forecast problem, being one of the leading causes of weather-related fatalities over the past several decades in addition to causing billions of dollars in economic damages in the annual mean (e.g., NWS 2017b). Part of the complexity compared with other weather hazards derives from the addition of hydrologic considerations alongside the purely meteorological ones. Antecedent soil conditions and the current levels of rivers and streams have a considerable influence on the proportion of rainfall that becomes surface runoff (e.g., Wood 1976; Castillo et al. 2003; Brocca et al. 2008). Land type and land use can also play a critical role (e.g., Ogden et al. 2000; Hapuarachchi et al. 2011), spanning the gamut from extremely absorbent sands to pavement, which can effectively saturate with very little rainfall. Urban effects such as pavement curvature and storm drain networks can also affect whether a flash flood is observed (e.g., Smith et al. 2005; Meierdiercks et al. 2010; Wolff 2013). Particularly in areas of complex terrain, the hydrologic response may also be highly sensitive to the precise spatiotemporal distribution of the precipitation; slight spatial displacements or differences in storm intensity may change whether a flash flood is observed. (e.g., Yatheendradas et al. 2008; Versini et al. 2010). Beyond the challenges from the hydrologic perspective, meteorologically, a complex combination of ingredients must come together to generate and sustain rainfall rates sufficient to produce flash flooding (e.g., Doswell et al. 1996; Davis 2001; Schumacher 2017). Flash flood–producing precipitation, which predominantly originates from warm-season moist convection over most of the contiguous United States (CONUS;

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-18-0092.s1.

Corresponding author: Gregory R. Herman, gherman@atmos.colostate.edu

DOI: 10.1175/JHM-D-18-0092.1

e.g., Schumacher and Johnson 2005, 2006; Stevenson and Schumacher 2014; Herman and Schumacher 2016), is consequently one of the most challenging and poorly forecasted sensible weather elements in contemporary numerical weather prediction (NWP; e.g., Fritsch and Carbone 2004; Novak et al. 2014).

Further exacerbating the flash flood forecast problem is the considerable difficulty in verifying flash flood events (e.g., Welles et al. 2007; Gourley et al. 2012; Barthold et al. 2015), an essential component to forecasting any phenomenon. There is no observation source with sufficient accuracy and density to determine whether a flash flood has occurred at every location across the CONUS (e.g., Gourley et al. 2012, 2013; Barthold et al. 2015). Stream gauge measurements are useful, but they inherently cannot capture urban and other types of flash floods and are much too sparse even on streams and rivers to provide adequate spatial resolution (e.g., Gourley et al. 2013). Flash flood reports (FFRs) from human observations are subject to population bias, with report databases often missing transient floods in very rural areas or at night (e.g., Pielke et al. 2002), and also to varying reporting and report encoding practices in different regions of the United States (e.g., Ashley and Ashley 2008; Calianno et al. 2013). Flash flood warnings (FFWs) have similar inconsistencies associated with differing warning philosophies across weather forecast offices (WFOs; e.g., Barthold et al. 2015; Marjerison et al. 2016; Schroeder et al. 2016), different proclivities to warn rural areas (e.g., Marjerison et al. 2016), and the fact that they correspond to anticipated—rather than observed—impacts.

Nevertheless, because of the societal threat posed by excessive rainfall and flash flooding, there is immense utility in having accurate flood and flash flood analyses and forecasts. Given the sensitivities and complications associated with calculating the hydrologic response to precipitation and the importance and urgency of disseminating updated flash flood information, it is often attractive in operational flash flood analysis and very near-term forecasting to simplify the problem down to a matter of only QPE. In this simplified framework, the question becomes: is the precipitation a given location has received or is receiving over some duration, as estimated by the QPE, sufficient to induce a flash flood? This essentially amounts to a binary exceedance question of whether the QPE over time T is in excess of some unknown threshold Θ_T above which flash flooding will occur and below which it will not. Even in this simplified framework, there are many challenges, which can be classified into two broad areas: 1) the discrepancy between true precipitation and QPE and 2) the determination of T and Θ_T. On the former class of complications, current QPEs struggle with accurately quantifying extreme precipitation amounts (e.g., AghaKouchak et al. 2011; Hou et al. 2014; Zhang et al. 2016). Gauge observations have insufficient spatial resolution and density, while radar observation accuracy suffers from coarse and range-dependent vertical resolution, as well as having only indirect measurements of precipitation rate. Resultantly, QPE products are inherently too coarse to adequately capture local maxima corresponding to flash flooding. Even the highest-resolution products have substantial deficiencies (e.g., Nelson et al. 2016), which are also examined in this study.

Optimal threshold and interval determination is a complex, multilayered challenge as well. One approach that attempts to do just this is the Flash Flood Guidance (FFG) product issued routinely by NWS River Forecast Centers (RFCs; Sweeney 1992). Based on the antecedent conditions and characteristics of the basin, dynamic estimates of Θ_T are issued on a subdaily basis for $T =$ 1, 3, and 6 h. However, these are not a panacea; because the CONUS is so hydrometeorologically diverse and there is no agreed single best methodology to compute these thresholds, different RFCs apply different methodologies to calculate FFG thresholds (e.g., Sweeney 1992; Ntelekos et al. 2006; Schmidt et al. 2007; Villarini et al. 2010; Clark et al. 2014), which can often produce highly different estimates and large nonphysical discontinuities across RFC boundaries (e.g., Clark et al. 2014; Barthold et al. 2015). Other approaches simplify the Θ_T estimation question and avoid these nonphysical political discrepancies by considering QPE exceedances of static thresholds, themselves derived across the CONUS in a consistent manner. In particular, a fixed threshold (e.g., 2 in. h^{-1}) can be used as a proxy for flash flooding, as has been used in numerous previous studies (e.g., Brooks and Stensrud 2000; Hitchens et al. 2013; Novak et al. 2014). Exceedances of thresholds defined relative to the local precipitation climatology, such as average recurrence intervals (ARIs), can serve as Θ_T estimates as well. An ARI defines a fixed frequency relative to the hydrometeorological climatology of the region; in particular, it corresponds to the expected duration, given the local climatology, between exceedances of a given threshold. For example, the 1-yr ARI for 24-h precipitation accumulations describes the accumulation amount for which one would expect the mean duration between exceedances of said amount to be 1 year. Past research has shown that a fixed-frequency ARI-based framework can have better correspondence with heavy precipitation impacts than the use of any fixed threshold across the hydrometeorologically diverse regions of the CONUS (e.g., Reed et al. 2007).

There are two primary objectives for this study. First, it seeks to evaluate the characteristics, deficiencies, and differences for existing QPE products and other tools and frameworks used in flash flood forecasting and analysis. Second, comparative evaluation of correspondence between QPE threshold exceedances and flash flood observations is performed to ascertain the merits of different QPE sources and the most effective ways to use QPE information for flash flood analysis and forecasting on both regional and national scales. Improved understanding of these properties can lead to more effective use of existing information in the short term and identify revisions that may be adapted to existing products and algorithms to remove these undesirable properties in the longer term, resulting in more useful products for flash flood forecasting and analysis across a range of time and spatial scales. This study investigates issues surrounding these two important classes of challenges by first examining the climatological characteristics of heavy precipitation in several popular QPE sources. The issue of threshold quantification and application in flash flood analysis and forecasting is investigated with extensive comparison between different QPE threshold exceedances and flash flood observations. Section 2 describes the numerous datasets used in the study, and section 3 describes the analysis methods employed therefrom. Section 4 presents characteristics of the various QPE threshold exceedances and other sources employed in this study, and section 5 assesses the correspondence between QPE exceedances and flash flood observations on regional and national scales. Section 6 summarizes the findings, describes the most important implications, and provides suggestions for future work. A number of acronyms are used in this study; to relieve memory burden, the reader is invited to consult appendix A for a complete listing of acronyms used.

2. Datasets

FFRs in this study come from NWS local storm reports (LSRs) so encoded as flash floods. Archived LSRs are obtained from Iowa State University's Iowa Environmental Mesonet (IEM) geographical information system (GIS) archive (available online at https://mesonet.agron.iastate.edu/request/gis/). NWS FFWs were also obtained from the IEM GIS archive. FFWs have been storm-based rather than county-based since 2008 (e.g., Waters et al. 2005; Ferree et al. 2006). Both warning type and report encoding are conducted for a given county warning area (CWA) by a governing WFO. Alternative report encoding options include "flood" and "heavy rain," while alternate weather warning and advisory options include flood warnings, flood advisories, and areal and small stream flood advisories. Practices on warning type, report encoding, and proclivity to issue warnings at all vary based on local WFO philosophy and practices (e.g., Barthold et al. 2015; Nielsen et al. 2015). Both FFRs and FFWs are available with temporal resolution to the minute.

There are several different gridded QPE sources currently in use in operational analysis and forecasting. Three leading sources are the National Centers for Environmental Prediction (NCEP) Stage IV Precipitation Analysis product (ST4; Lin and Mitchell 2005), the Climatology-Calibrated Precipitation Analysis (CCPA; Hou et al. 2014), and the Multi-Radar Multi-Sensor QPE product (MRMS; Zhang et al. 2016). ST4 provides QPEs across the CONUS on a ~4 km grid for 1-, 6-, and 24-h accumulations centered about 1200–1200 UTC meteorological days. It uses both rain gauge observations and radar-derived rainfall estimates to generate an analysis and is further quality controlled via RFCs, particularly for 6- and 24-h QPEs, to remove stray radar artifacts and other spurious anomalies (Lin and Mitchell 2005). ST4 products are generated by each RFC, and each center applies somewhat different treatments in generating the products. Most importantly, 1-h QPE is not in general provided by the Northwest RFC and has not been routinely generated by the California–Nevada RFC since early 2016. When provided, 1-h QPEs in this region are generally a simple disaggregation of 6-h QPE into 1-h intervals. CCPA is derived from two QPE sources, the ST4 QPE and Climate Prediction Center's unified global daily gauge analysis. In particular, because of more rigorous and uniform quality control, the CPC-based QPE product is thought to have more accurate estimates than ST4 but has lower spatial and temporal resolution, at 1/8° and 24-h, respectively. A linear regression technique is applied to upscaled and aggregated 6-h ST4 QPE to correct its distribution toward the more robust CPC-based estimates, and then downscaled back to the native resolution to derive more accurate estimates while maintaining the spatiotemporal resolution of ST4. However, due to the limitations of linear regression, extremes not in the original ST4 cannot be introduced in the calibration process employed in CCPA, and extremes in ST4 are inherently regressed to some extent toward more typical values in the local precipitation climatology (Hou et al. 2014). Both ST4 and CCPA have up to a full day of latency in generation and publication of the QPE products. Last, MRMS, which became operational in September 2014, employs approximately 180 operational radars to create CONUS-wide radar mosaics every 2 min on a 1-km grid. In conjunction with gauge, satellite, and other environmental data, these radar mosaics are used to create CONUS-wide QPE at very high spatial and temporal resolution (Zhang et al. 2016). An initial radar-only product is produced with 2-min latency. The MRMS QPE used in this study has an

TABLE 1. Threshold sources examined as a function of AI and QPE source, using the symbology of the manuscript text.

	1 h	3 h	6 h	24 h
MRMS	FT, ARI, FFG	FT, ARI, FFG	FT, ARI, FFG	FT, ARI
ST4	FT, ARI, FFG	FT, ARI, FFG	FT, ARI, FFG	FT, ARI
CCPA			FT, ARI	FT, ARI

TABLE 2. FT thresholds examined as a function of AI, where "X" indicates that the given threshold–AI combination is examined.

	1 in.	1.5 in.	2 in.	2.5 in.	3 in.	3.5 in.	4 in.	5 in.	6 in.
1 h	X	X	X	X	X	X	X		
3 h		X	X	X	X	X	X	X	
6 h		X	X	X	X	X	X	X	
24 h			X	X	X	X	X	X	X

additional gauge correction step, whereby gauges are ingested and undergo quality control, after which they are compared against the collocated radar-only estimates—incorporating aspects such as gauge network density and distance from estimate location—to develop a bias grid that is subtracted from the radar-only estimates. This gauge-corrected MRMS QPE has approximately 1 h of latency, still much less than ST4 and CCPA (Zhang et al. 2016), leading to more operational applicability in operational forecast settings. In this study, we compare how the high precipitation tail of each of these QPE sources compares with each other and with FFRs.

In addition to each of these QPE sources, two other factors are considered as well: accumulation interval (AI) length and threshold source. Regarding AIs, threshold exceedances for 1-, 3-, 6-, and 24-h QPEs are considered as flash flood proxies. All of these AIs are considered for ST4 and MRMS QPEs; only 6- and 24-h QPEs are available from CCPA. Three different sources for flash flood thresholds are considered as well: 1) a fixed threshold (FT) across the CONUS, 2) exceedances of ARI thresholds, and 3) exceedances of FFG. Each of these methods is explained further below; FT and ARI exceedances are available for all AIs, while FFG exceedances are available for 1-, 3-, and 6-h accumulations. A full summary of the threshold exceedance comparisons made for each QPE source is provided in Table 1 for reference.

FT grids are extraordinarily simple, as they are constant across the CONUS. A variety of different thresholds are considered to assess the relationship between precipitation severity and flash flooding. Applying more stringent thresholds will result in fewer false alarms but more misses, while more lenient thresholds will induce the opposite result; it is expected that an optimal balance exists between these extremes. Of course, precipitation sufficient to produce flash flooding when falling within an hour likely will not produce flash flooding when distributed across a longer period. Therefore, the exact thresholds considered must necessarily depend on the AI; the full set of thresholds evaluated is indicated in Table 2.

The ARI thresholds are generated using very similar methodology to Herman and Schumacher (2016), where CONUS-wide thresholds are produced by stitching thresholds from several sources. NOAA's Atlas 14 thresholds (Bonnin et al. 2006, 2011; Perica et al. 2011, 2013, 2015), an update from older work and currently under development, are used wherever they were available at the time this research began. For five northwestern states—Washington, Oregon, Idaho, Montana, and Wyoming—updated thresholds are not available, and derived Atlas 2 threshold estimates are used instead (Miller et al. 1973; Herman and Schumacher 2016). In Texas, which currently has Atlas 14 threshold estimate updates in progress but no finalized thresholds available, Technical Paper 40 (TP-40; Hershfield 1961) estimates are used. Everywhere else uses the Atlas 14 ARI threshold estimates. All of these threshold estimates are based on many decades of gauge data based on the availability and density of historical data in the region. While sophisticated spatial statistics are applied to derive the estimates and downscale to ungauged locations, particularly in the case of Atlas 14 (e.g., Bonnin et al. 2011; Perica et al. 2011), it is possible that undersampling from use of exclusively gauges can result in uncertain or erroneous estimates, particularly in historically rural areas, in areas of complex terrain, and areas without updated thresholds. Threshold uncertainty is quantified in Atlas 14 and increases with increasing ARI; this study uses only the best estimate values provided for the 1-, 2-, 5-, 10-, 25-, 50-, and 100-yr ARI thresholds for each different AI. Atlas 14 provides estimates for each of these AIs. TP-40 provides estimates for each of these ARIs but only for 6- and 24-h AIs. Furthermore, NOAA Atlas 2 has available in digitized form only 6- and 24-h ARI thresholds for the 2- and 100-yr ARIs. Herman and Schumacher (2016) derived thresholds for those AIs for the other ARIs. For Texas, Washington, Oregon, Idaho, Montana, and Wyoming, 1- and 3-h threshold estimates are thus not natively available and had to be derived. This was accomplished using the methodology described in appendix B and stitched with the Atlas 14 estimates to produce the CONUS-wide threshold grids of Fig. 1 (see also Fig. S1 in the online supplemental material). The grids present a stark contrast to the spatially uniform FT grids, with values spanning near an order of magnitude across the CONUS. Climatologically drier areas such as the Intermountain West have lower thresholds, while wetter regions such as the Gulf Coast have higher thresholds. As expected, thresholds are lowest for the smallest AIs and ARIs and become larger with increasing duration and rarity. However, the extent of change as a function of AI in particular is not spatially uniform and instead reflects the climatological characteristics of the types of precipitation systems associated with locally extreme precipitation in the given region. This is seen to some extent when comparing the left and right columns of Fig. 1, but especially when comparing those of Fig. 1 with those of Fig. S1. For example, while thresholds for the 24-h AI are comparable between the Gulf Coast and Pacific coastal mountains (Fig. S1), the former region has much higher thresholds at the 1-h AI (e.g., ~125 mm vs ~45 mm for the 100-yr ARI in Fig. 1m). Further, locations much farther north and more distant from an ocean, such as over Iowa and Minnesota, have appreciably lower thresholds than the Pacific mountains for 24-h accumulations but are also much higher for 1-h AIs. Over the Pacific Coast, extreme precipitation events are typically associated with long-duration atmospheric river events, which can produce moderate to heavy rain for an extended duration (e.g., Rutz et al. 2014; Herman and Schumacher 2016). In contrast, over the Southeast and Great Plains, most extreme precipitation is associated with smaller-scale convective systems, which can produce higher rain rates than their West Coast counterparts, but last for a shorter duration at any given point (e.g., Herman and Schumacher 2016). The Gulf Coast region sustains high thresholds across the spectrum of AIs; at shorter AIs, this is predominantly associated with small-scale convective storms, while high thresholds at longer durations are predominantly supported by tropical cyclone rainfall (e.g., Kunkel et al. 2012).

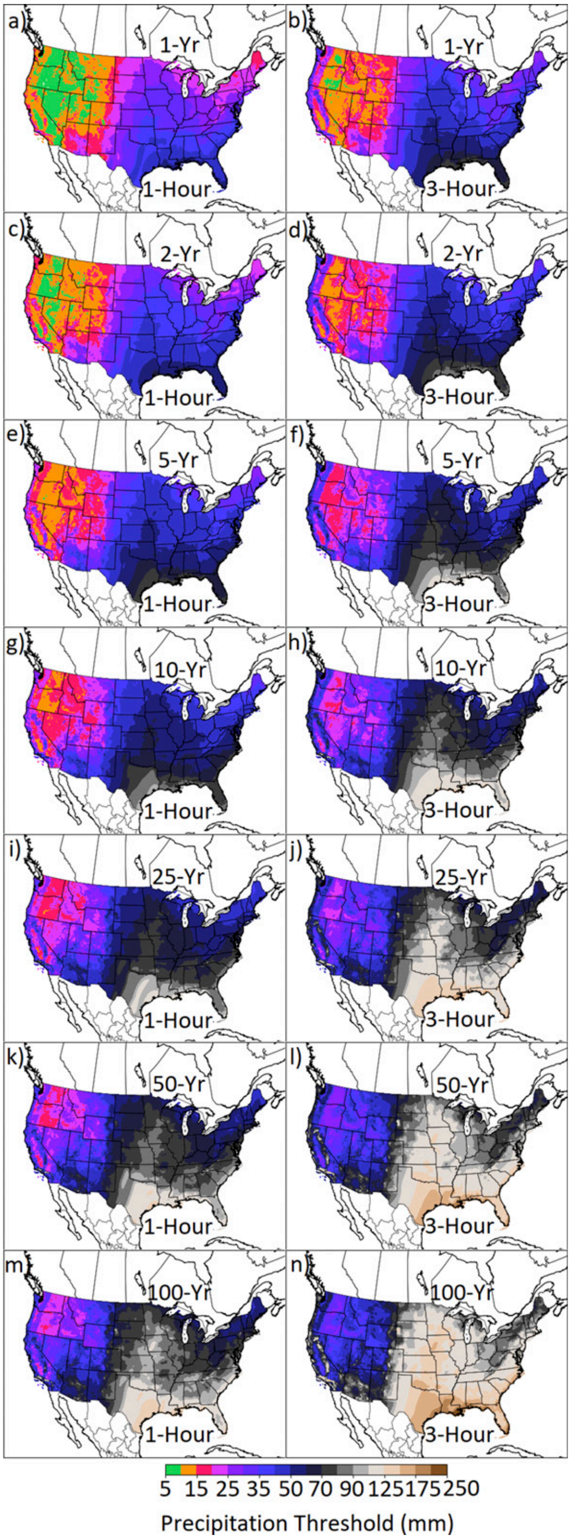


FIG. 1. ARI threshold estimates for (left) 1-h and (right) 3-h precipitation accumulations for (a),(b) 1-; (c),(d) 2-; (e),(f) 5-; (g),(h) 10-; (i),(j) 25-; (k),(l) 50-; and (m),(n) 100-yr ARIs. Threshold estimates come primarily from NOAA Atlas 14 but are supplemented from other sources as described in the text.

FFG estimates the average precipitation amount required over an area in a prescribed amount of time to initiate flooding of small streams in that area (Sweeney 1992). FFG is calculated individually by each RFC, with each office maintaining independent code and algorithms for FFG calculation (Sweeney 1992; Barthold et al. 2015). RFC-generated FFG may be assembled to form a national grid covering all of the CONUS, with the exception of Washington and Oregon west of the Cascades; the Northwest RFC does not calculate FFG for this small region of the CONUS. FFG values are based on threshold-runoff calculations, which specify the minimum amount of runoff (not precipitation) into a stream or basin over a prescribed 1-, 3-, or 6-h duration necessary to produce bank-full conditions (Sweeney 1992; Ntelekos et al. 2006). This is done offline for thousands of small basins and is independent of present conditions. These basin-specific threshold-runoff calculations are interpolated onto a ~4 km grid to providec;a unified analysis. A hydrologic model, such as the Sacramento Soil Moisture Accounting

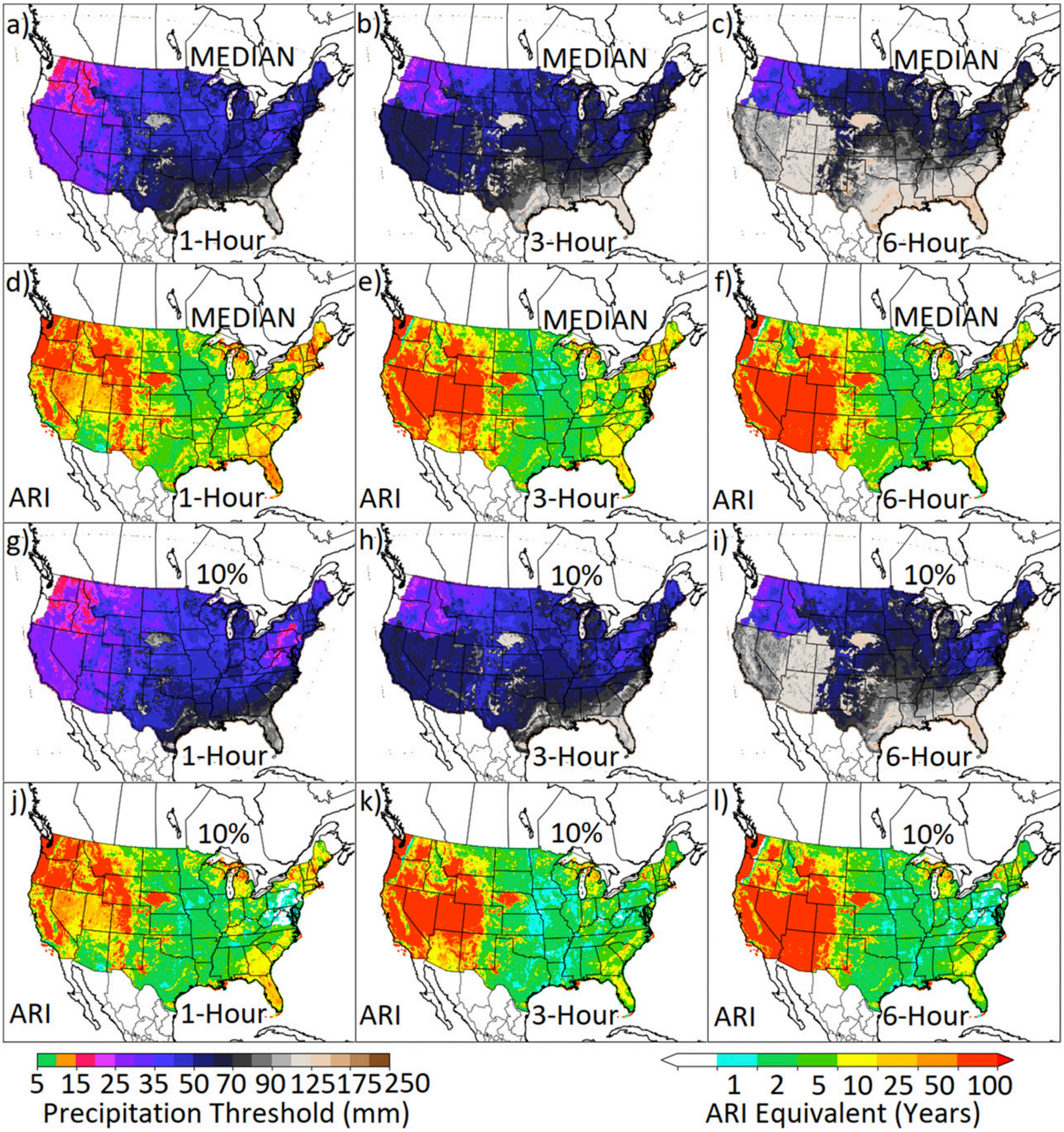


FIG. 2. (a)–(f) Median and (g)–(l) 10th percentile FFG estimates over the 2.5-yr period of record. The (left) 1-, (center) 3-, and (right) 6-h FFG values. Panels (a)–(c) and (g)–(i) correspond to the actual threshold estimates, while (d)–(f) and (j)–(l) correspond to the equivalent ARIs to those thresholds for the particular grid point.

Model (e.g., Carpenter et al. 1999) or Antecedent Precipitation Index models (e.g., Brocca et al. 2008), are then used in conjunction with current conditions to relate rainfall amounts to runoff amounts. The minimum rainfall to yield a runoff in the hydrologic model in excess of the gridded threshold-runoff values then constitute the gridded FFG values (Ntelekos et al. 2006).

Unlike ARI thresholds, FFG thresholds vary dynamically based on the antecedent conditions. While this makes it impossible to plot a single static plot depicting the FFG thresholds across the entire period of record, the distribution of issued values can be considered. The median FFGs across the period of record (Figs. 2a–c) vary in a similar fashion to the tail of the precipitation climatologies as quantified by the ARI thresholds (Fig. 1), with very low values over the Intermountain West increasing progressively to very high values over the Gulf Coast and particularly Florida. However, while ARI thresholds reflect only the precipitation climatology and do not *directly* address the hydrologic component of flash flooding, FFG does account for these factors. This can result in large gradients in FFG climatologies in regions of rapid change in soil type or land use; one prominent example is in the Nebraska Sand Hills (e.g., Figs. 2a,b). Also evident are the large spatial discontinuities that occur even in the median across RFC boundaries. One glaring example in the 6-h median FFG (Fig. 2c) is the border between the Northwest RFC and California–Nevada RFC near the southern borders of Oregon and Idaho. The same general findings exist on the high-risk tail of the FFG climatology, as evidenced by the 10th percentile FFGs (Figs. 2g–i). In general, the difference between the 50th and 10th percentiles over the western RFC domains (cf. Figs. 2b and 2h) is small, while thresholds for the 10th percentile are appreciably—although not uniformly—lower across the central and eastern CONUS. In particular, the Middle Atlantic RFC appears to be more responsive to antecedent conditions than its neighbors, resulting in locally lower thresholds in their domain and large spatial discontinuities in the 10th percentile FFGs at their RFC boundaries; this is especially pronounced for 1-h FFG (Fig. 2g).

Both ARI and FFG exhibit strong and clearly apparent contrasts with FT methodology but are quite different from each other as well. Median FFGs are, according to ARI thresholds (Figs. 2d–f), most commonly exceeded over the Great Plains and Mississippi Valley regions. There, ARI equivalents for median FFGs can be as low as 1 year in Iowa for 3-h FFGs (Fig. 2e) and are between 2 and 5 years across most of the region. In contrast, median FFGs near and along the Atlantic Coast are generally appreciably higher, with values of 10–25 years. Higher still are typical thresholds in the West, with ARI equivalents mostly ranging from 25 to over 100 years. In the West, large differences in ARI equivalence are found depending on the AI used. In the northern Intermountain West, including Idaho, equivalent ARIs to the median 6-h FFGs (Fig. 2f) are only 2–5 years, while being mostly 10–25 years in those same areas for 1-h FFGs (Fig. 2d). The opposite, and even stronger, contrast is seen in the arid Southwest, particularly Arizona. The ARI equivalent for the median 6-h FFG (Fig. 2f) is at least 100 years, while it is 2–5 years over much of the state for 1-h FFG (Fig. 2d) and is even as low as a 1-yr ARI across the southeast portion of the state. These AI-dependent contrasts suggest, for example, that most floods in the Southwest are associated with short-lived rain events and that for 1- and 6-h rain events of equal rarity in the Southwest, the 1-h event rates have greater hydrometeorological impact. All the same general findings are also found comparing the ARI framework with the 10th percentile FFG thresholds, just with lower ARI equivalent thresholds. The one very prominent difference is again in the Middle Atlantic RFC; here, 10th percentile 1-h FFGs are below the 1-yr ARI across much of their domain (Fig. 2j), while the 10th percentile FFGs in neighboring areas largely correspond to 5–10-yr ARIs.

3. Analysis methodology

All grids are first regridded if necessary onto the ST4 Hydrologic Rainfall Analysis Project (HRAP; Fulton et al. 1998) ~4-km grid. ST4 and CCPA QPEs are already provided on this grid; MRMS QPE is regridded onto this grid using a first-order conservative scheme (Ramshaw 1985). ARI and FFG thresholds are regridded bilinearly onto this grid. FFRs and FFWs are not gridded products at all, the former being points in space and the latter polygons in space. FFRs are remapped onto the HRAP grid using a 40-km radius of influence, projecting a single report onto numerous points on the grid. Events are defined for 24-h 1200–1200 UTC "meteorological days," similar to current operational practice at the Weather Prediction Center and Storm Prediction Center (e.g., Edwards et al. 2015; NWS 2017a; Herman et al. 2018). As such, despite both having 1-min resolution, an FFR "event" for the purpose of this study is defined as one or more reports within 40 km of the point occurring anytime within the meteorological day. An FFW event is similarly defined as any FFW enclosing the given HRAP grid point valid at any time during the meteorological day.

Once this is performed and all fields are assembled on a uniform grid, slight additional quality control is performed following Herman and Schumacher (2016) to remove QPEs that are clearly nonphysical, and then binary comparison between QPE and selected thresholds is made. Comparisons are first made on the ST4 HRAP grid to generate binary exceedance grids. For subdaily AIs, there are multiple grids with valid times falling within a given 1200–1200 UTC period. In these instances, the maximum of all grids with the same AI and corresponding meteorological day is taken to form a single daily exceedance grid for the series; all subsequent analyses use these aggregated daily grids. In this way, daily grids for subdaily AIs correspond to one or more of the given type of QPE exceedance occurring at that point during the meteorological day, regardless of

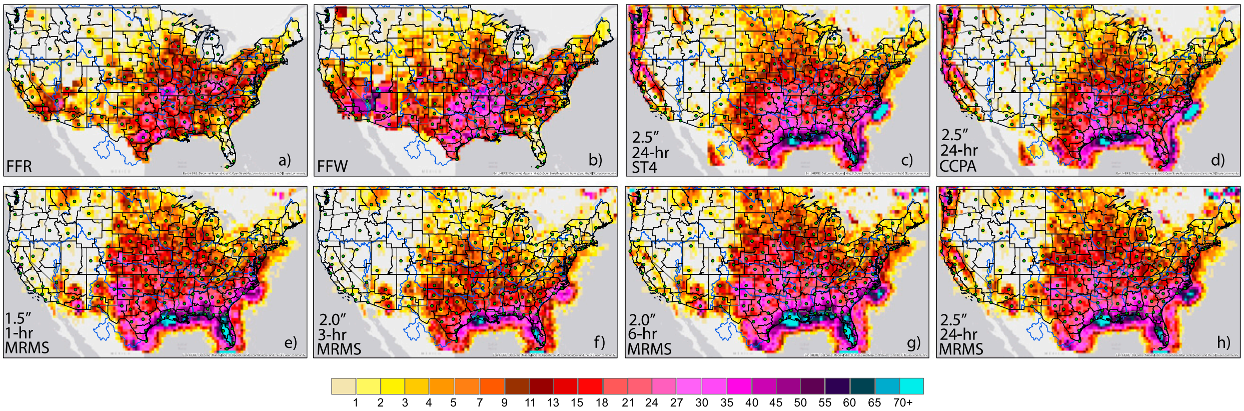


FIG. 3. Heat maps for FT exceedances, FFRs, and FFWs during the relevant period of record (see text). (a) FFRs and (b) FFWs reported and issued during the period of record, gridded as described in the manuscript text. (e)–(h) MRMS QPE FT exceedances, where columns from left to right correspond to 1-, 3-, 6-, and 24-h precipitation accumulations and 1.5, 2.0, 2.0, and 2.5 in. (38, 51, 51, 64 mm). (c),(d) As in (h), but for ST4 and CCPA, respectively. Thick black outlines depict CWA boundaries, blue lines indicate RFC domain boundaries, and green circles indicate locations of NEXRAD radar sites.

the exact number of exceedances. Tolerance to small spatial displacements is provided by using a maximum nearest neighbor upscaling from the HRAP grid to a 0.5° × 0.5° grid. For each day, all HRAP grid points are mapped to their nearest point on the 0.5° grid; at each grid point on this coarser grid, an event is recorded if any of their mapped HRAP points indicated an exceedance event for that meteorological day. QPEs are only considered for periods centered about the meteorological day and are not considered for any interval spanning across meteorological days (e.g., 24-h 0000–0000 UTC accumulations). Based on data availability, a 2.5-yr period of record spanning from 2 January 2015 to 23 June 2017 is used for most verification in this study, with a slightly truncated period beginning 19 March 2015 for MRMS QPE comparisons, again limited by data availability.

After exceedance grids have been computed, they are compared to assess how the characteristics vary as a function of threshold source, accumulation interval, threshold magnitude, and QPE source. Despite the aforementioned limitations of FFRs and FFWs, evaluation is made using each of these sources as a reference truth. Although these are not believed to completely embody ‘‘true’’ occurrences and nonoccurrences of flash floods, it is performed in order to provide a common framework for comparison between different QPE exceedances. In this framework, the reference—either FFRs or FFWs—serves as a deterministic truth, and the QPE exceedances serve as deterministic predictions. The analysis framework employed here, namely, deterministic binary predictions and binary observations, lends itself well to the use of contingency table statistics (Wilks 2011). Given the number of different thresholds, intervals, and sources considered, it is convenient to represent the comparison statistics succinctly in a single plot. One popular way to present the full dimensionality of the contingency table verification for many different forecast sets in a single plot is through the so-called performance diagram (PD; Roebber 2009). The PD succinctly places a forecast set in the context of these verification statistics on one plot, with success ratio (SR) increasing on the x axis, probability of detection (POD) increasing on the y axis, frequency bias (FB) increasing from 0 at the lower right corner to infinity at the upper left, and critical success index (CSI) increasing from 0 at the lower left corner to unity—a perfect score—at the upper right. In addition to PDs, spatial maps of CSI are assessed to provide context of where correspondence between the QPE exceedances and reference truth are better and worse across the CONUS. Finally, a single geometric mean equitable threat score (ETS) is computed between the comparisons with the two reference datasets and is so chosen over CSI to alleviate concerns that the latter exhibits with varying underlying event frequencies (Gandin and Murphy 1992; Marzban 1998) and produce a skill metric more independent of the event climatology (e.g., Jolliffe and Stephenson 2003). A geometric mean is chosen over a conventional one to more strongly penalize lack of correspondence with either observation set.

4. Results: Exceedance climatologies

Examination of simple exceedance, report, or warning counts, as the case may be, over the period of record in Fig. 3 illuminates several interesting contrasts between the datasets. A heat map of FFRs over the period

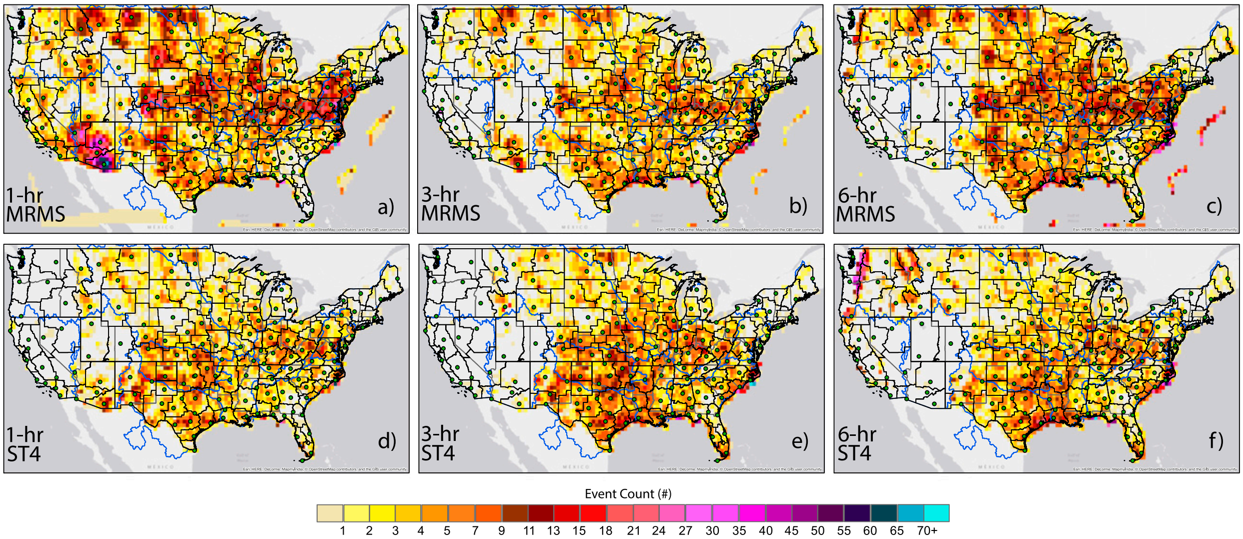


FIG. 4. Heat maps for FFG exceedances across the CONUS during the relevant period of record (see text). (a)–(c) Exceedances of FFG based on MRMS and (d)–(f) ST4 QPE exceedances for (left) 1-, (center) 3-, and (right) 6-h FFGs. Thick black outlines depict CWA boundaries, blue lines indicate RFC domain boundaries, and green circles indicate locations of NEXRAD radar sites.

(Fig. 3a) illustrates several of the aforementioned limitations of reports as representations of true flash floods. The population bias is clearly evident in some regions of the country. For example, in Texas, far more reports are observed in the Houston, Dallas, Austin, and San Antonio metro areas than in surrounding areas despite them having similar rainfall climatologies (e.g., Fig. 1). Some of this is likely a legitimate reflection of urban environments flooding more easily than rural ones due to land use and other factors, but extracting the component of legitimate spatial variation from a population-based reporting bias is challenging. Spatial variations attributable to human factors can be discerned as well, with discontinuities in report counts across CWA boundaries in places, particularly entering Florida and Michigan and to a lesser extent in Georgia and Alabama (Fig. 3a). These political effects from WFO tendencies are even more prominent in FFW issuances (Fig. 3b) in those same locations. Additionally, there are several local "hot spots," such as Las Vegas, wherein one WFO issues far more FFWs than its neighbors. Again, some of this is certainly meteorological due to different climatological flood susceptibility of neighboring CWAs and from limited sampling due to the finite period of record. For example, FFWs and especially FFRs (Fig. 3a) are more concentrated in the Southern Plains and Midwest, with far fewer events over the Northern Plains and farther west over the Rocky Mountains and Pacific Coast. These large-scale spatial variations accord with previous studies of flash flooding (e.g., Brooks and Stensrud 2000; Hitchens et al. 2013; Schumacher and Johnson 2006) and are likely legitimate rather than an artifact of the datasets. The magnitude of the differences suggest, however, that at least some contribution to these *local* spatial variations across CWA boundaries is political rather than purely hydrometeorological. FT exceedances, in contrast, do not exhibit any of these spatial discontinuities at political boundaries. For 1-h accumulations (e.g., Fig. 3e), they instead exhibit a prominent, relatively smooth gradient from almost no exceedances of 1.5 in. h^{-1} occurring over the northwestern CONUS, to being extremely common over the southeastern CONUS near the Gulf Coast. The spatial distribution of events over the central and eastern CONUS remain similar with increasing AI (cf. Figs. 3e and 3h), but the number of exceedances along the Pacific Coast increases dramatically, with almost no exceedances for 1-h accumulations (e.g., Fig. 3e) and as many exceedances as the Gulf Coast for 24-h accumulations (e.g., Fig. 3h). This largely accords with the ARI thresholds, which are similar in these two regions for 24-h thresholds (Herman and Schumacher 2016) and much higher over the Gulf Coast for 1- and 3-h accumulations (Fig. 1).

Politically attributable exceedance count discontinuities are also evident in FFG exceedance heat maps (Fig. 4), in this case primarily across RFC boundaries. For 1-h FFGs (Fig. 4a), this is most readily apparent with respect to the Middle Atlantic RFC; there are far more exceedances in their domain than either their Northeast or Southeast RFC neighbors. For 3- and 6-h FFGs (Figs. 4e,f), a very large discontinuity is seen between the Colorado Basin RFC and its neighbors to the north and east, including the Northwest, Missouri Basin, and West Gulf RFCs, with almost no exceedances of 3- or

6-h FFGs in the Colorado basin domain but numerous exceedances immediately adjacent in other RFC domains. These discrepancies are consistent with the FFG threshold climatologies (Fig. 2). There are also far fewer total FFG exceedances across the CONUS than exceedances of the FT thresholds presented in Fig. 3, with a prominent exception of 1-h MRMS QPE FFG exceedances in Arizona (Fig. 4a). There is also in general far less spatial gradient in exceedance counts of FFG compared with the FT exceedances. In places, such as the Southeast and particularly Florida, the anomalously low number of reports (Fig. 3i) and warnings (Fig. 3j) in the area compared with its surroundings is corroborated by a relatively low number of FFG exceedances (e.g., Fig. 4d), while in other areas, such as Michigan (e.g., Fig. 4c), it is not.

Aside from sampling noise associated with using a finite and relatively short period of record, ARI threshold exceedance heat maps (Fig. 5) should definitionally be uniform across the entire spatial domain. Departures from uniformity must then be attributable to either 1) sampling noise, 2) inaccurate ARI threshold estimates, or 3) systematic error in the QPE source. Comparison across different QPE sources and threshold sources can help identify root causes. Some notable spatial variations can be seen—some are consistent across QPE sources, while others are particular to one. For example, MRMS QPE (Figs. 5a,b,e,f) exhibits a glaring anomaly in exceedance counts in the West: there are far more ARI exceedances observed in the immediate vicinity of radar sites compared with their surroundings. While this is evident for all AIs and across different levels of severity, it is especially apparent for shorter intervals (e.g., Figs. 5a,b). This phenomenon, which is also seen in the FT (e.g., Fig. 3e) and FFG (e.g., Fig. 4a) exceedances, is considerably alleviated or even entirely eliminated to the east of the Rocky Mountains. ST4 QPE ARI exceedances (Figs. 5c,d,g,h) exhibit two sharp local maxima far above any other location: one in Wyoming and the other in New Mexico. This is especially prominent for the 24-h AI (Fig. 5h); it is seen to a lesser extent with MRMS QPE (Fig. 5f) as well. Interestingly, the discontinuity in FFG exceedance counts across the Colorado basin RFC boundary is replicated in the ARI exceedances—especially prominent in ST4 but evident in all QPE sources. This suggests that the discontinuity may be largely attributable to artifacts of the native QPE rather than politically based discontinuities in FFG thresholds. ST4 exceedances at short 1- and 3-h AIs (e.g., Figs. 5c,d) have a substantial reduction in ARI exceedances in the West due to the aforementioned limited production of 1-h ST4 QPE in the western RFCs. Last, for CCPA QPE (Figs. 5i,j), the maximum in Wyoming remains clearly apparent, but the maximum in New Mexico is greatly muted. The consistency of overexceedances in Wyoming suggests that the ARI threshold estimates, which are now several decades old, may be too low in this area. That the New Mexico maximum is largely removed with the bias correction applied by CCPA suggests that the New Mexico issue may be more attributable to deficiencies with ST4 and MRMS QPE in complex terrain, with small areas of large radar estimated values unable to be properly corrected due to insufficient gauge data in the region.

5. Results: Flash flood correspondence skill

PDs for CONUS-wide verification for the complete set of QPE to observation reference comparison verification in Fig. 6 (see also the online supplemental material) illustrate that, for a given QPE source, AI, and threshold method, a curve sweeps from the top-left corner of the PD to the bottom-right corner with increasing threshold magnitude. The lowest thresholds jointly exhibit a high POD, high FB, and low SR, while high thresholds possess the opposite characteristics. The curve is not, however, parallel with the curved skill (CSI) lines in the PD and instead attains a maximum CSI for some middle threshold magnitude. Surprisingly, out of all the different QPE comparisons against FFRs (Fig. 6), maximum CSI values are obtained for 2.0 in. (50.8 mm) $(6\,\mathrm{h})^{-1}$, 2.5 in. (63.5 mm) day^{-1}, and 3.0 in. (76.2 mm) day^{-1}—all FT threshold sources (warm-colored interior symbols). The maximum CSI obtained using ST4 (blue-outlined symbols) and CCPA-based (green-outlined symbols) QPE comparisons exceed that reached using MRMS QPE exceedances (red-outlined symbols), though the CSI differences are small, with maximum values all around 0.23. Across the range of threshold comparisons considered, the highest FBs extend over 5, and the lowest are well under 0.1; SRs range from 0.1 to 0.65, and PODs range from almost 0 to near 0.85. FBs for FFG exceedances are all near or below unity, consistent with the findings of Clark et al. (2014), which also found raw FFG to be too stringent and found better correspondence using fractional FFG. For a given threshold method, magnitude, and AI, there are similar common differences between QPE source comparisons. CCPA QPE consistently exhibits a lower FB, higher SR, and lower POD than ST4 or MRMS; MRMS usually exhibits the highest FB, highest POD, and lowest SR of the three. When compared against FFWs (Fig. 7), the general scatter of QPE exceedance verifications in the PD phase space remain the same, but CSIs are generally somewhat higher and differences in the specifics emerge. Specifically, while the highest CSI values are obtained for ST4 and CCPA QPEs with FFRs, correspondence with

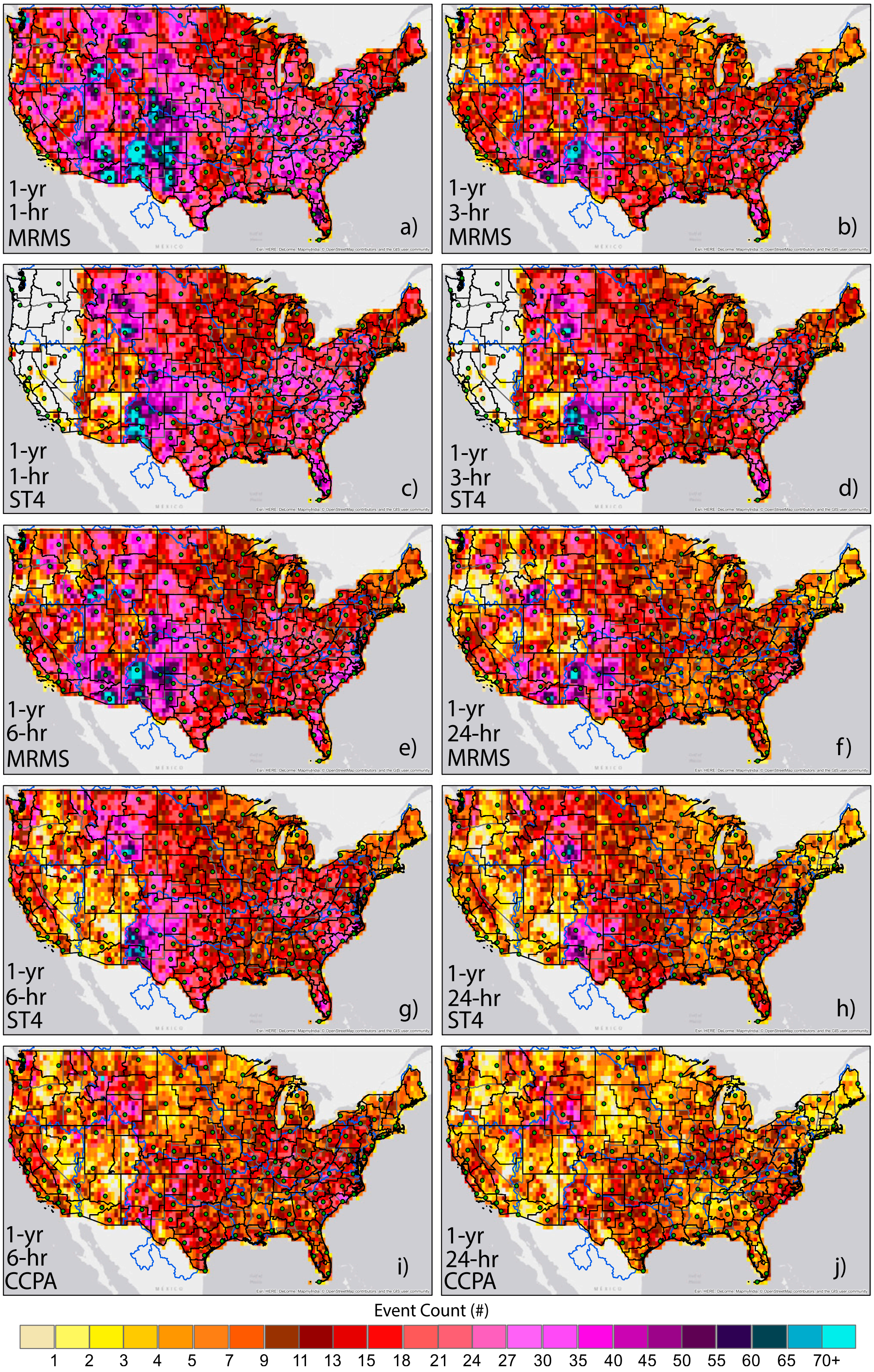


FIG. 5. Heat maps for ARI exceedances for different ARIs, AIs, and QPE sources across the period of record. MRMS QPE exceedances of 1-yr for (a) 1- and (b) 3-h ARIs; (c),(d) as in (a) and (b), but for ST4 QPE. MRMS QPE exceedances of the 1-yr ARI for (e) 6- and (f) 24-h accumulations. (g),(h) ST4 exceedances and (i),(j) CCPA for 6- and 24-h accumulations. Symbology otherwise as in Fig. 3.

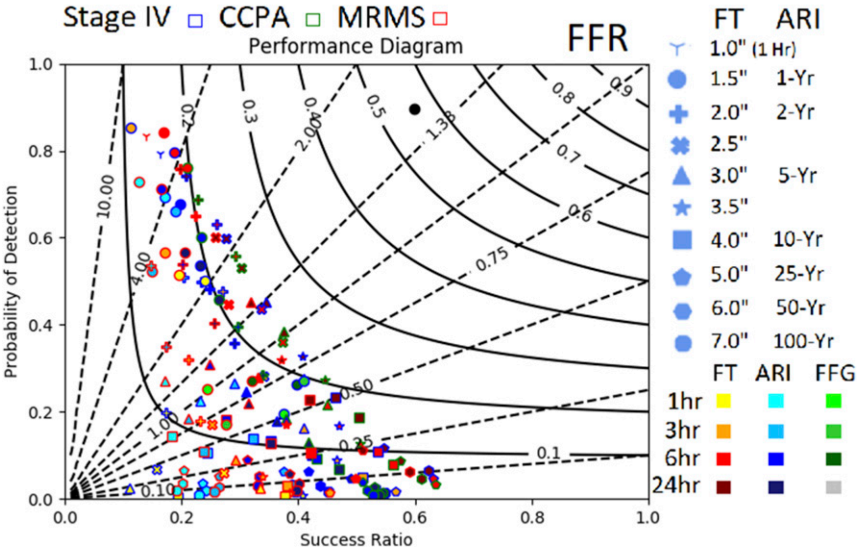


FIG. 6. Performance diagrams as per Roebber (2009) evaluated over the entire period of record and across all of the CONUS. Verification made with respect to FFRs for many different QPE threshold exceedances. The symbol shape corresponds to the threshold magnitude, as indicated in the top of the panel legend. All FFG exceedances use a circular symbol. The inner color to each symbol indicates the AI associated with the comparison, as indicated in the middle part of the figure legend. Outer edge colors indicate the QPE source of the marker, with blue corresponding to ST4, green to CCPA, and red to MRMS as indicated at the bottom of the figure legend. The black circle depicts the verification of FFWs with respect to FFRs. Further description to aid with interpretation of PDs is included in the manuscript text.

FFWs is maximized using MRMS QPE exceedances. Consistent with using FFRs for truth, the 2.5 in. day^{-1} threshold provides the maximum CSI among the various QPE exceedance comparisons evaluated.

The PD view also allows for straightforward identification of some flaws and deficiencies in the QPE products. For example, as noted above, 3-h ST4 QPE, which is acquired by summing 1-h ST4 QPE, has less quality control and thus more spurious high values than compared with 6-h ST4 QPE, which is a separate, independent grid and not necessarily equal to the sum of the six 1-h QPEs that fall within the 6-h period. This is evident in Fig. 6, for example, when comparing 3-h FT exceedances with the same magnitude 6-h FT exceedances. For the same QPE source and period of record, there should necessarily be as many or more exceedances of a given precipitation amount occurring over a 6-h period than a 3-h one. However, the orange circle surrounded by blue, denoting ST4 QPE exceedances of 1.5 in. (38.1 mm) (3 h)$^{-1}$ or higher, has a *higher* FB than the red circle surrounded by blue, denoting exceedances of the same threshold over a 6-h period. Other deficiencies and limitations of QPE sources exist on a regional basis as well, some of which are discussed below.

Maps of CSI (Fig. 8) reveal considerable spatial variability in correspondence between different QPE exceedances and FFRs. FFWs (Fig. 8g) have much better correspondences with FFRs than any QPE threshold exceedance—as evidenced by the black dot of Fig. 6—including FFG exceedances (Figs. 8a–f). Highest FFR–FFW CSI is found across much of the CONUS east of the Rocky Mountains. Exceptions include central North Dakota, southern Florida, and Michigan, where there is an overall lack of reports

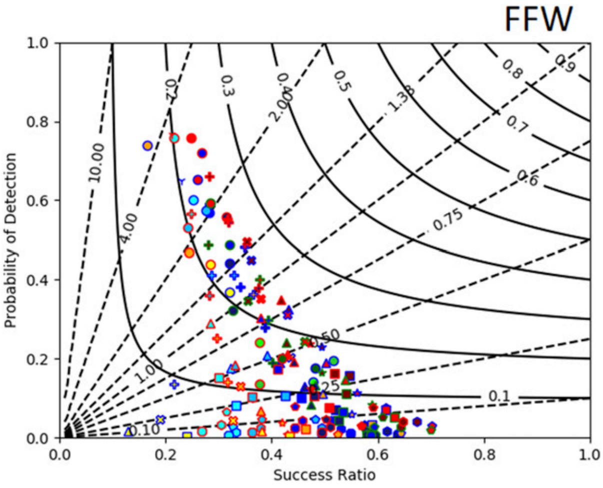


FIG. 7. As in Fig. 6, but using FFWs as reference "truth."

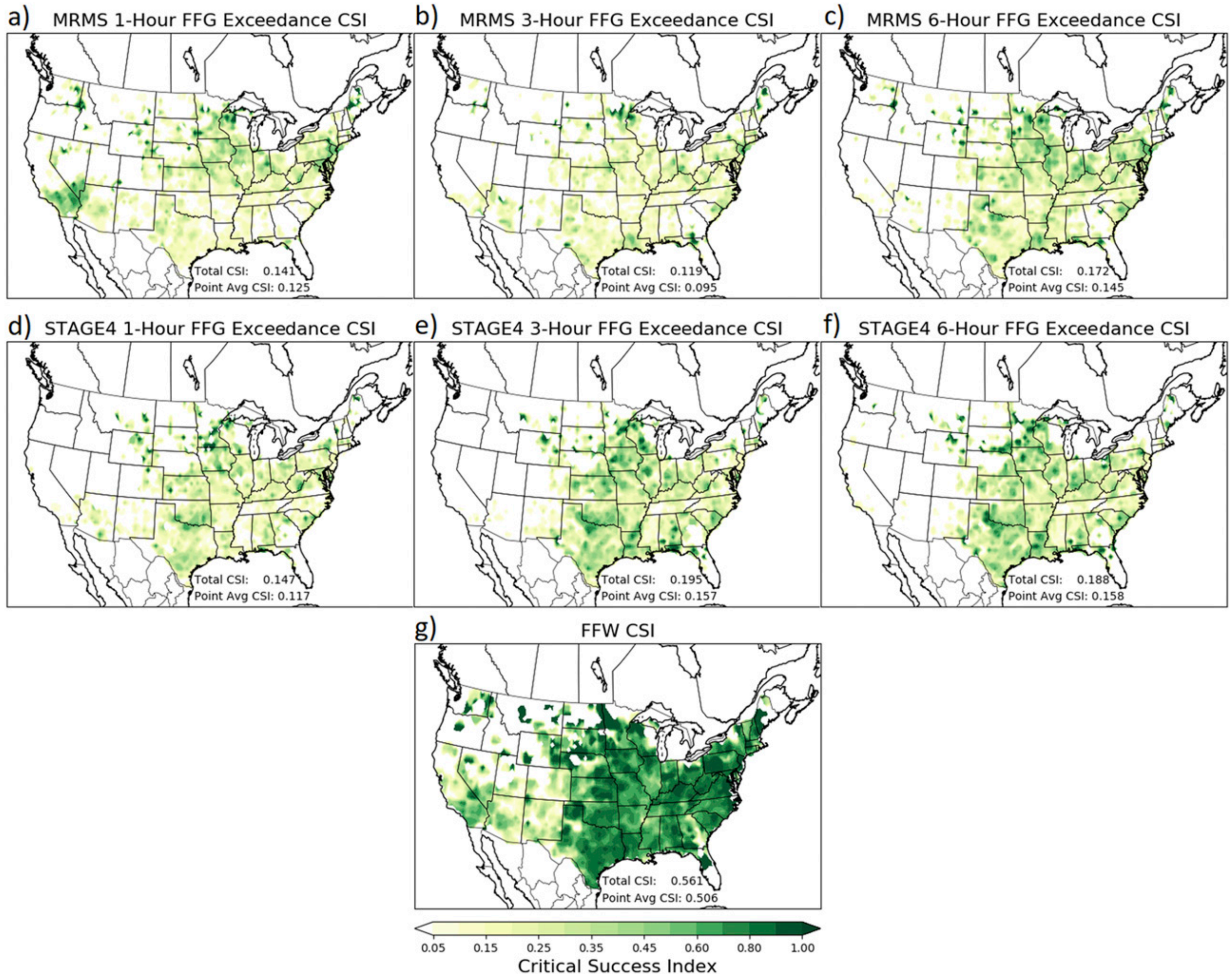


FIG. 8. CONUS-wide maps of CSI for comparisons between several sources with FFRs used as reference for "truth." The top six panels correspond to exceedances of FFG based on (a)–(c) MRMS and (d)–(f) ST4 QPE for (left) 1-, (center) 3-, and (right) 6-h FFGs. (g) Correspondence between NWS FFWs and FFRs over the period of record, again based on CSI. The top number at the bottom of each panel shows the CSI for the corresponding threshold comparison when all observations contribute to a single set of hits, misses, and false alarms. The bottom number instead shows aggregate performance when the aggregate scores over the period of record are calculated individually for each grid point, and then averaged between all grid points.

(Fig. 3a). The number of reports is similarly scarce over much of the northern Intermountain West, resulting in low CSI scores there as well; in the extreme of no reports over a grid point, the CSI is necessarily 0, since it is impossible to hit. In the West, scores are higher than surrounding areas in southern California and Nevada, where there are both more reports (Fig. 3a) and many more warnings (Fig. 3b) than in adjacent locations. FFG exceedances all exhibit somewhat similar CSI spatial distributions. The 1-h MRMS QPE exceedances of FFG (Fig. 8a) appear to perform the best of the six over the West, particularly in the southern Nevada and California vicinity. Correspondence with FFRs is generally highest over the Midwest and Southern Great Plains, with maximum correspondence for longer AIs (e.g., 6-h; Figs. 8c,f) and with ST4 QPE providing better correspondence compared with MRMS (cf. Figs. 8b and 8e). CONUS-wide FFG-based CSIs, 0.1–0.2 depending on various choices, are quantitatively consistent with past findings over different study periods (Clark et al. 2014).

Comparing QPE exceedances of FFG (Fig. 8) with a sample of evaluated FT (Fig. 9) and ARI (Fig. 10) thresholds yields several interesting findings. Overall, largely because the base QPEs are the same and the only difference is the threshold discriminator between flash flood and nonevent, the spatial character of correspondence between QPE exceedances and FFRs is broadly similar for each set of exceedances. MRMS QPE exceedances consistently yield the best correspondence with FFRs in the Southwest in southern California and

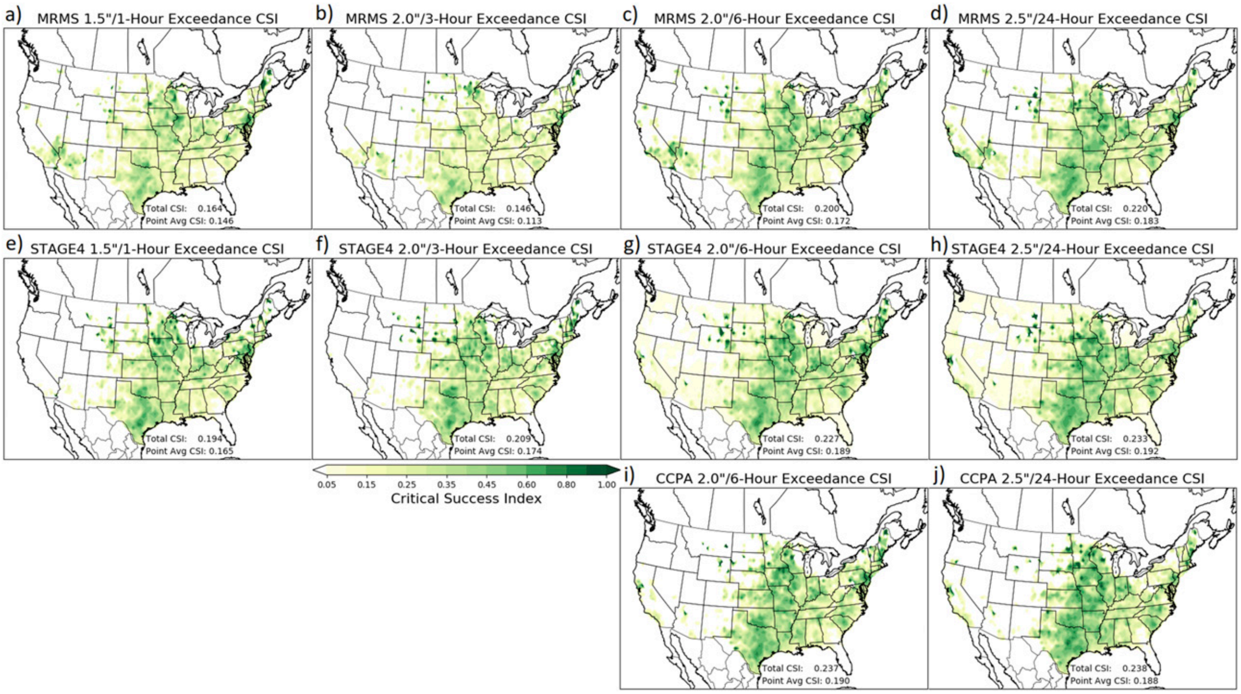


FIG. 9. Maps of CSI for comparisons between several QPE FT exceedances with FFRs used as reference for "truth." (a)–(d) MRMS-based FT exceedances, (e)–(h) ST4-based FT exceedances, and (i),(j) comparison of CCPA-based FT exceedances with FFRs. Columns from left to right correspond to 1-, 3-, 6-, and 24-h precipitation accumulations and 1.5, 2.0, 2.0, and 2.5 in. (38, 51, 51, 64 mm).

Nevada, and the highest CSI appears to be achieved for 1-yr, 24-h ARI exceedances (Fig. 10d) in that local area. ST4 QPE exceedances, particularly for longer AIs such as the 2.5 in. day^{-1} exceedances (Fig. 9h), appear to achieve the highest CSI across the broader western region. Longer AIs, particularly in the ARI framework (Fig. 10), appear to exhibit improved correspondence compared with 1- and 3-h QPEs across the central and eastern CONUS as well (cf. Figs. 10i and 10l). The lower 1-yr ARIs demonstrate superior correspondence with FFRs across the CONUS compared with more extreme 10-yr thresholds, with the one exception of 10-yr, 1-h MRMS QPE exceedances, which provide optimum correspondence in the aforementioned small region surrounding Las Vegas and vicinity (Fig. 9e).

CSI maps illustrate that conclusions from aggregate nation-scale statistics do not always hold when zoomed to regional scales; PDs subsetted to particular regions allow for more quantitative analysis across the full spectrum of threshold comparisons. For example, using the region definitions of Herman and Schumacher (2018b) shown in Fig. 11, the Northeast (NE; Fig. 12a) and Southeast (SE; Fig. 12b) regions exhibit appreciably different verification results to the national total. In particular, in both of these regions, ARIs clearly outperform the use of either FT or FFG thresholds, and ST4 outperforms MRMS and to a lesser extent CCPA, as evidenced by the blue interior and exterior symbols, respectively, placed farther toward the upper-right corner of each panel. However, the exact thresholds that obtain optimal skill vary between the two regions; in the NE (Fig. 12a), the 2-yr ARI achieves optimum CSI, while in the SE region (Fig. 12b), the 1-yr ARI produces better results, with the 2-yr exceedances being negatively biased. Moreover, while 6-h AI exceedances produce maximum skill in comparison with FFRs across the NE, 24-h accumulations are more predictive across the SE region. In the SE region, the 1-yr, 24-h ARI exceedances are nearly equally skillful using ST4 and CCPA QPE, but ST4 is positively biased while CCPA is negatively biased. In both regions, FFWs correspond very well to FFRs, with CSIs of 0.76 and 0.63 in the NE (Fig. 12a) and SE (Fig. 12b) regions, respectively.

Over the Great Plains regions, Northern Great Plains (NGP; Fig. 13a) and Southern Great Plains (SGP; Fig. 13b), some mixed signals are found. In NGP, the 2.5 in. (63.5 mm) day^{-1} threshold using CCPA QPE attains the highest CSI score of all of the QPE threshold exceedance comparisons using FFRs as a reference. However, this scores very similar to 3.0 in. (76.2 mm) day^{-1} and 2.5 in. (63.5 mm) (6 h)$^{-1}$ thresholds for ST4 QPE, with the 2.5 in. day^{-1} threshold suffering from too many false alarms. While the 2-yr ARI produces the best results among the ARI thresholds considered, all ARI comparisons lag the best FT CSI values. FFG exceedances, while

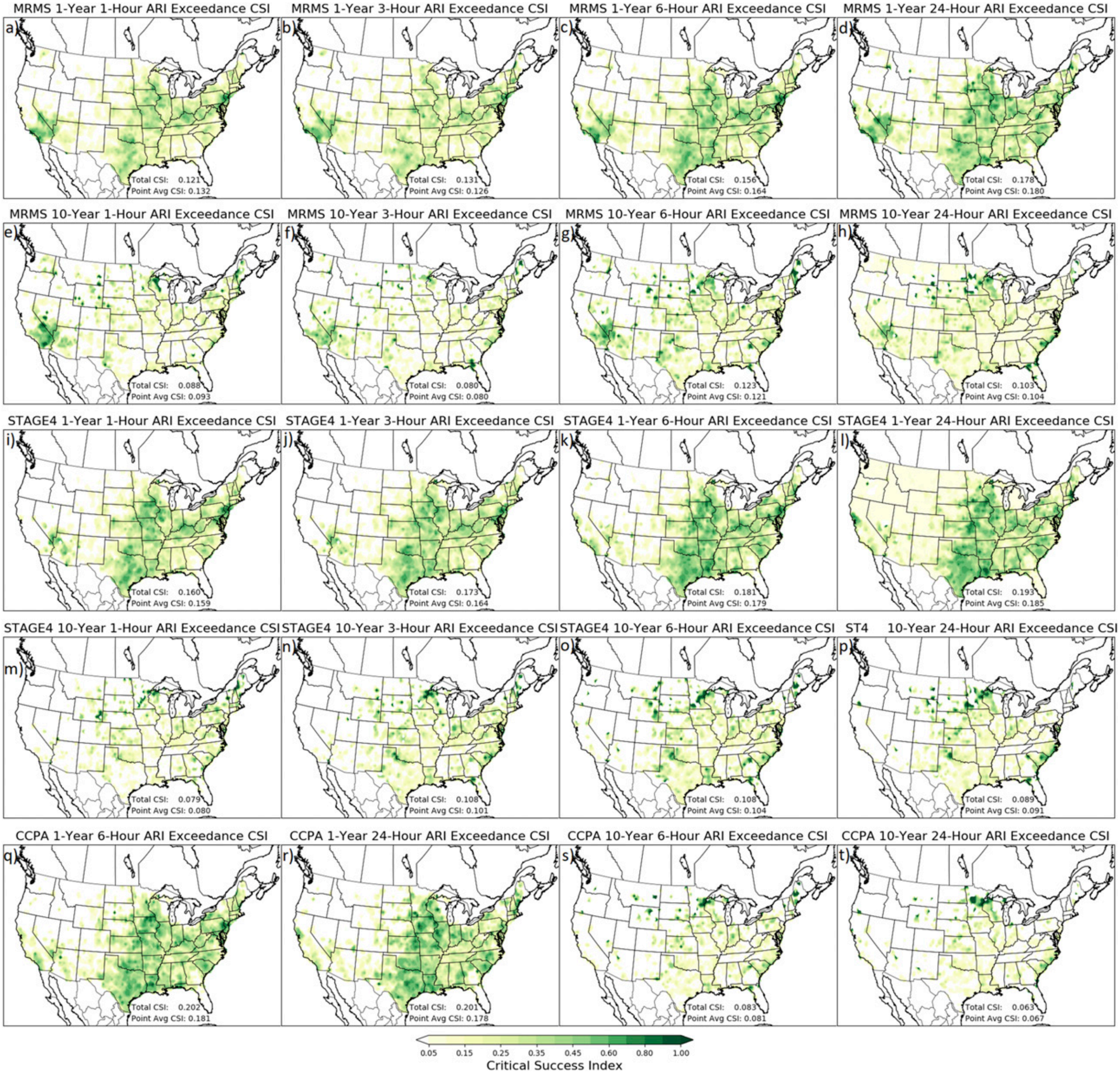


FIG. 10. Maps of CSI for comparisons between several QPE ARI exceedances with FFRs used as reference for "truth." (a)–(h) MRMS-based ARI exceedances, (q)–(t) CCPA-based ARI exceedances, and (i)–(o) ST4-based ARI exceedances. Panels (a)–(d), (i)–(l), and (q)–(r) correspond to 1-yr ARI exceedances, while (e)–(h), (m)–(p), and (s)–(t) are for 10-yr exceedances. Columns from left to right correspond to 1-, 3-, 6-, and 24-h precipitation accumulations, except for (q) and (r), which correspond to 6- and 24-h accumulations, respectively.

more competitive than in the eastern regions (Fig. 12), still lag FT exceedances in NGP (Fig. 13a). In SGP (Fig. 13b), the 1-yr, 24-h ARI exceedances using CCPA QPE attain the highest CSI of any QPE comparison. The 3.5 in. (88.9 mm) day^{-1} with ST4 QPE performs almost equally well—a higher threshold than in NGP owing to the wetter precipitation climatology in the region (e.g., Fig. 1). FFG is again somewhat competitive, especially for 6-h accumulations, but lags the other methods.

The Southwest region (SW; Fig. 11) displays very different verification characteristics (Fig. 14) to both the CONUS-wide perspective and the other individual regions examined above. Correspondence between all QPE exceedances and the reference truth are much worse than across the nation as a whole with both FFWs (Fig. 14b) and especially FFRs (Fig. 14a) serving as reference. The correspondence between the two reference truths is also particularly poor (Fig. 14a), with a CSI

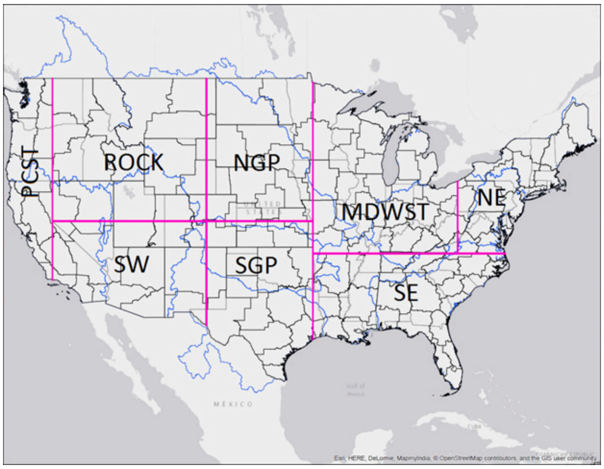

PCST
ROCK
NGP
MDWST
NE
SW
SGP
SE
MEXICO
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community

FIG. 11. Map depicting the regional partitioning of the CONUS used in this study, and the labels ascribed to each region. Adapted from Herman and Schumacher (2018b).

of only 0.3. The relative verification of the QPE exceedances also contrasts sharply with the results of other regions. Unlike over the East and Great Plains, MRMS QPE provides much better correspondence with both FFRs and FFWs than ST4 or CCPA QPE. Like in the East (Fig. 12), the ARI exceedances demonstrably outperform the FT and FFG exceedances in correspondence with both FFRs and FFWs (Fig. 14). But unlike other regions, especially for correspondence with FFRs (Fig. 14a), maximum CSI is attained for much shorter 1-h AIs, and also for much higher ARIs, with maximum correspondence for the 10-yr, 1-h ARI exceedances. Correspondence with FFWs (Fig. 14b) is higher across all comparisons. Furthermore, 3- and 6-h exceedances attain similar CSI scores with 1-h exceedances, and the highest CSI value is achieved with a much lower 2-yr ARI. Much of this is attributable to the fact that there are many more—nearly 3 times as many (Fig. 14a)—FFWs than reports in this region, with a frequency bias of near 3 (Fig. 14a). Overall, each region exhibits unique verification characteristics, with optimum QPE sources, AIs, and threshold levels all varying by region.

Synthesizing every comparison into a single ETS for each QPE exceedance set (Fig. 15) yields concrete quantitative conclusions largely consistent with the CONUS-wide findings discussed above. Higher ETSs are found for longer 6- and 24-h AIs for ST4 and MRMS QPEs. The highest score using a fixed threshold attains a higher ETS than the maximum comparison with ARIs, which in turn outperforms the best corresponding FFG exceedance set with the reference truths. Overall, despite very different characteristics, when averaged across the CONUS, each of the three QPE sources evaluated achieved similar overall verification scores, all achieving a maximum ETS of almost 0.24. ST4 and MRMS achieved maximum ETS for a 2.5 in. day^{-1} threshold, while CCPA's maximum ETS was for 2.0 in. (6 h)$^{-1}$ exceedances. Among ARIs, best correspondence was obtained for 1-yr, 24-h exceedances for MRMS and ST4 QPEs and 1-yr, 6-h exceedances for CCPA QPEs.

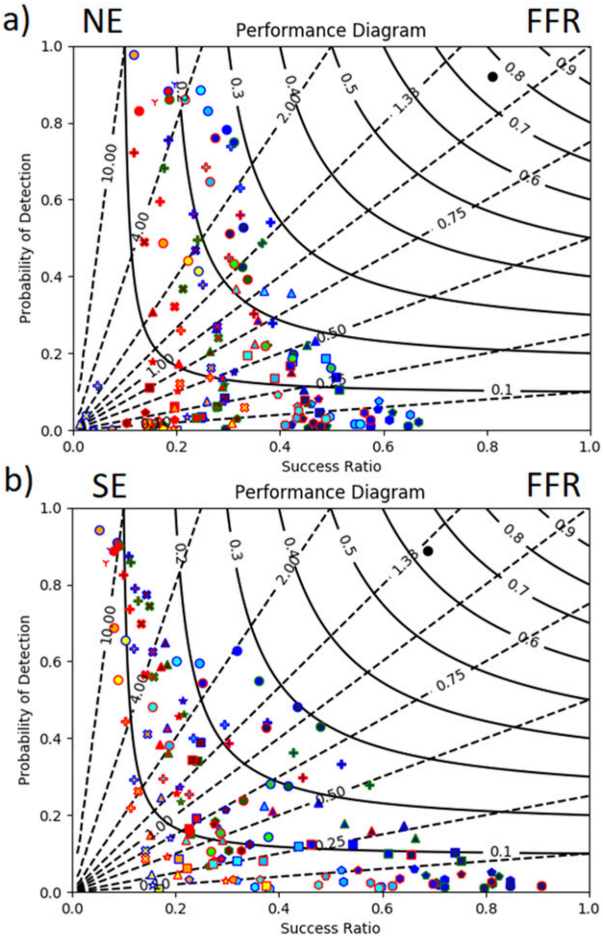


FIG. 12. As in Fig. 6, but with verification restricted to the (a) NE and (b) SE region, with region definitions as depicted in Fig. 11.

6. Summary and conclusions

This study performed an expansive comparison using different QPE-based threshold exceedances as a proxy for flash flooding, as quantified through flash flood reports and NWS flash flood warnings. Comparisons were conducted across the CONUS for an evaluation period spanning from January 2015 through mid-June 2017. Many different factors were considered, including the QPE accumulation interval, with 1-, 3-, 6-, and 24-h accumulations evaluated; the QPE source, with three leading QPE sources—ST4, MRMS, and CCPA—each

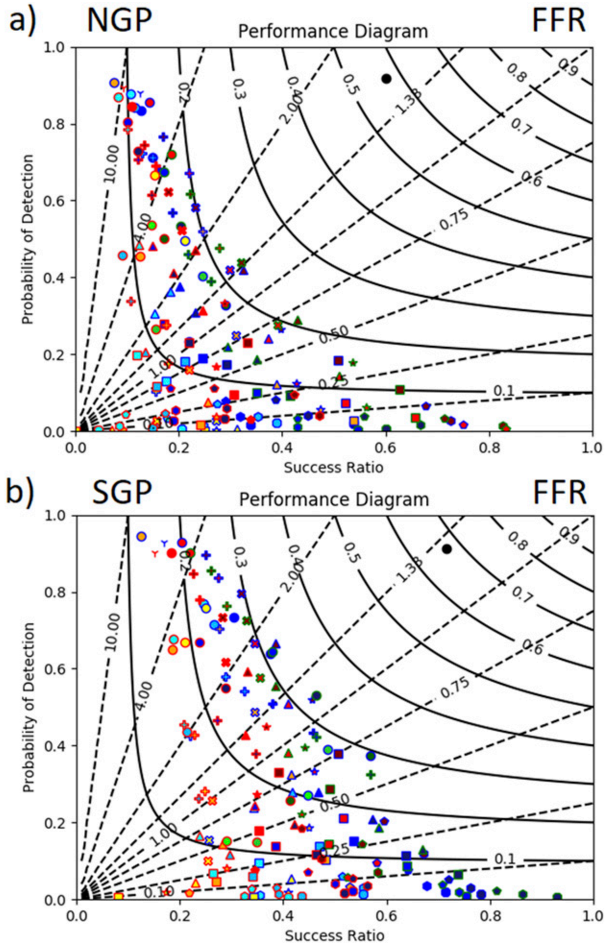


FIG. 13. As in Fig. 6, but with verification restricted to the (a) NGP and (b) SGP region, with region definitions as depicted in Fig. 11.

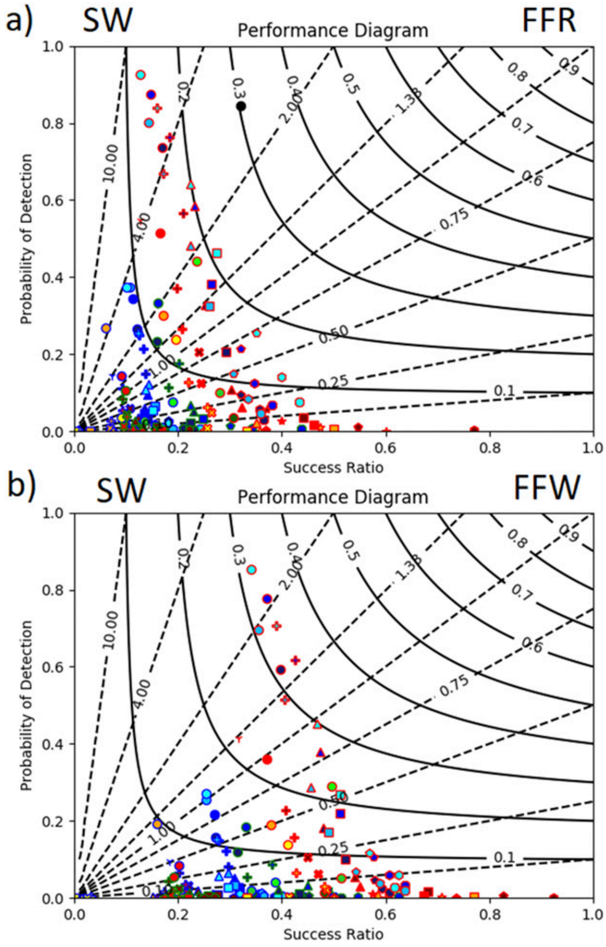
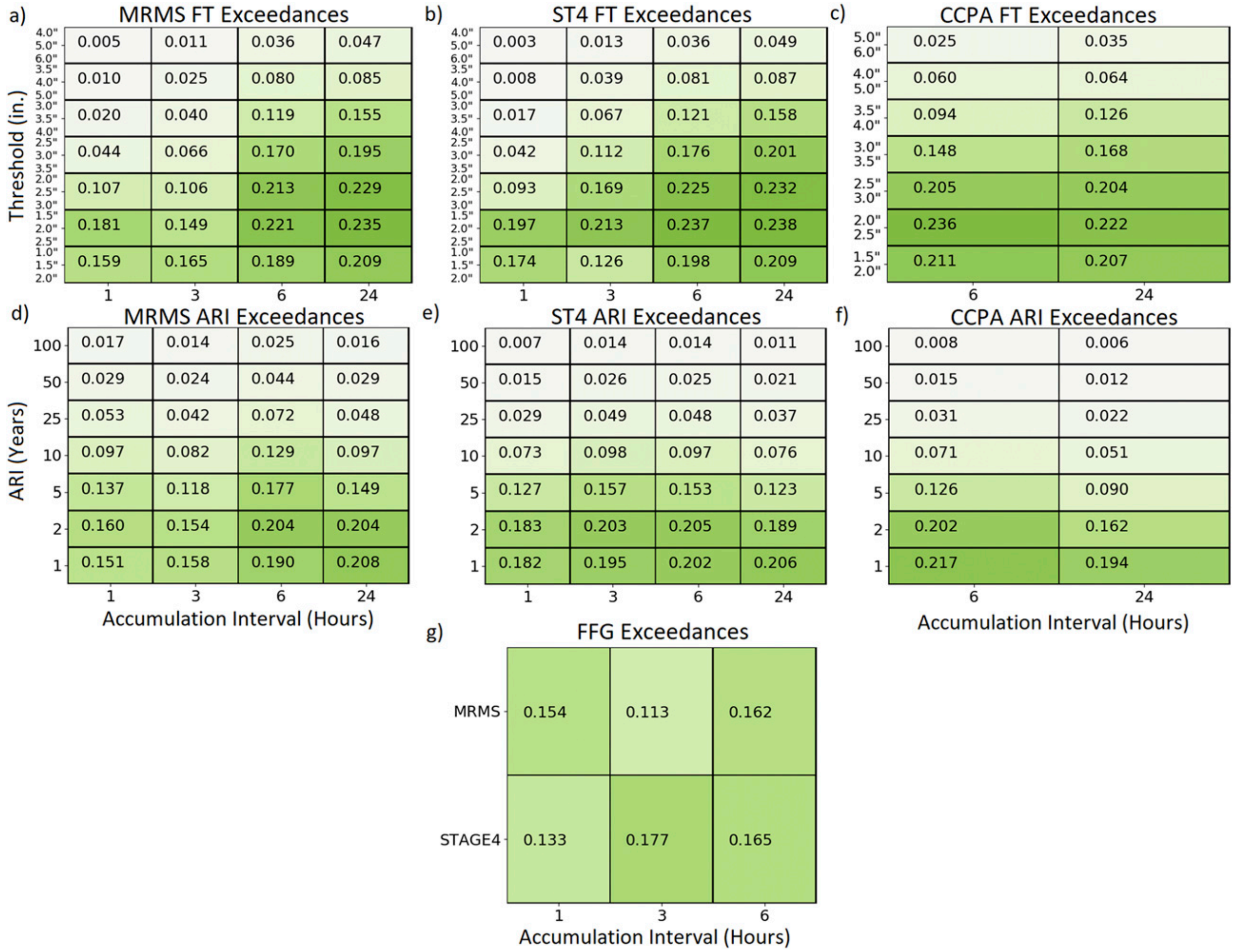


FIG. 14. (a) As in Fig. 6 and (b) as in Fig. 7, but with verification restricted to the SW region as defined in Fig. 11.

compared; and the method for deriving local QPE thresholds. In addition to considering FT exceedances, exceedances of ARIs ranging from 1 to 100 years are considered, as well as exceedances of NWS RFC-generated FFG. For each of these binary observation sets, climatologies based on the study period were constructed, and skill in correspondence between the threshold exceedances and FFRs and/or FFWs was assessed on both national and regional scales. Ultimately, the study investigates the characteristics of the high tail of the probability distribution of QPEs, and the relation these QPEs have with observed flash flood impacts across the CONUS.

Some of the findings from the study confirm prior knowledge of the hydrometeorological community, while in other areas, they introduce surprising and somewhat counterintuitive results. Even in the former case, this study gives concrete, quantitative numbers to some of these differences previously known or described only qualitatively. The principal findings of this study can be summarized as follows:

- The question of whether a flash flood has occurred is much more involved than a simple binary comparison between local QPE and a flash flood threshold. No QPE threshold exceedance corresponded well with either FFRs or FFWs.
- While the aggregate skill statistics across the CONUS were similar for each QPE source, significant regional differences emerged, with diminishing correspondence from east to west across the CONUS. MRMS does outperform ST4 and CCPA in FFR and FFW correspondence in the Southwest, while ST4 performs best in the East and CCPA the best over the central CONUS. With significant regional dependence, identification of existing deficiencies and areas for future product improvements can require regional, rather than purely national, analysis.
- Each QPE source has recurring deficiencies and biases. ST4 systematically reports heavy QPEs too frequently over much of the Intermountain West, but particularly in New Mexico and Wyoming, much more than other precipitation climatologies such as ARIs



a) MRMS FT Exceedances

Threshold (in.)	1	3	6	24
4.0" 5.0" 6.0"	0.005	0.011	0.036	0.047
3.5" 4.0" 5.0"	0.010	0.025	0.080	0.085
3.0" 3.5" 4.0"	0.020	0.040	0.119	0.155
2.5" 3.0" 3.5"	0.044	0.066	0.170	0.195
2.0" 2.5" 3.0"	0.107	0.106	0.213	0.229
1.5" 2.0" 2.5"	0.181	0.149	0.221	0.235
1.0" 1.5" 2.0"	0.159	0.165	0.189	0.209

b) ST4 FT Exceedances

	1	3	6	24
4.0" 5.0" 6.0"	0.003	0.013	0.036	0.049
3.5" 4.0" 5.0"	0.008	0.039	0.081	0.087
3.0" 3.5" 4.0"	0.017	0.067	0.121	0.158
2.5" 3.0" 3.5"	0.042	0.112	0.176	0.201
2.0" 2.5" 3.0"	0.093	0.169	0.225	0.232
1.5" 2.0" 2.5"	0.197	0.213	0.237	0.238
1.0" 1.5" 2.0"	0.174	0.126	0.198	0.209

c) CCPA FT Exceedances

	6	24
5.0" 6.0"	0.025	0.035
4.0" 5.0"	0.060	0.064
3.5" 4.0"	0.094	0.126
3.0" 3.5"	0.148	0.168
2.5" 3.0"	0.205	0.204
2.0" 2.5"	0.236	0.222
1.5" 2.0"	0.211	0.207

d) MRMS ARI Exceedances

ARI (Years)	1	3	6	24
100	0.017	0.014	0.025	0.016
50	0.029	0.024	0.044	0.029
25	0.053	0.042	0.072	0.048
10	0.097	0.082	0.129	0.097
5	0.137	0.118	0.177	0.149
2	0.160	0.154	0.204	0.204
1	0.151	0.158	0.190	0.208

e) ST4 ARI Exceedances

	1	3	6	24
100	0.007	0.014	0.014	0.011
50	0.015	0.026	0.025	0.021
25	0.029	0.049	0.048	0.037
10	0.073	0.098	0.097	0.076
5	0.127	0.157	0.153	0.123
2	0.183	0.203	0.205	0.189
1	0.182	0.195	0.202	0.206

f) CCPA ARI Exceedances

	6	24
100	0.008	0.006
50	0.015	0.012
25	0.031	0.022
10	0.071	0.051
5	0.126	0.090
2	0.202	0.162
1	0.217	0.194

g) FFG Exceedances

	1	3	6
MRMS	0.154	0.113	0.162
STAGE4	0.133	0.177	0.165

FIG. 15. Mean ETSs for each QPE exceedance method compared against both FFRs and FFWs, calculated as described in the manuscript text. Scores for FT QPE exceedance verifications for the (a) MRMS, (b) ST4, and (c) CCPA QPE sources. Like accumulation intervals are organized by column, while thresholds are organized by row. For (a) and (b), the top number of each row label corresponds to the threshold for the 1-h QPE exceedances, the middle number applies to the 3- and 6-h accumulation comparisons, and the bottom number to the 24-h QPEs. In (c), the top number corresponds to 6-h QPEs and the bottom number to 24-h QPEs. (d)–(f) As in (a)–(c) but for scores for QPE exceedances of ARI thresholds. Rows of these tables have a common ARI value, labeled in years; columns are again organized by accumulation interval. Panel (g) shows scores for FFG exceedances, with 1-, 3-, and 6-h FFGs in the leftmost, center, and right columns, respectively, and comparisons with MRMS and ST4 QPEs in the top and bottom rows, respectively.

would indicate. It also suffers from numerous spurious very large values in its 1-h QPEs that are not removed during quality control, occurring especially but not exclusively in the West. CCPA corrects for many of these issues, but in its linear calibration, resultantly removes many legitimate extreme events. It consequently has a much lower frequency bias than either ST4 or MRMS for a given QPE threshold set. MRMS also experiences many of the biases observed with ST4 in the West, but to a lesser degree. However, it additionally exhibits a strong sensitivity to radar location in that region, with many more QPE exceedance events occurring near radar sites compared with more distant locations.

- Regardless of the threshold framework, very high thresholds often employed in extreme rainfall studies and analyses are too stringent to provide optimal correspondence with FFRs or FFWs owing to too many missed flash flood events. In general, the least severe thresholds examined had among the best correspondence with the reference records.
- Contrary to expectations given the definition of a flash flood, correspondence between QPE exceedances and the reference records tended to improve with increased accumulation interval. Minimum correspondence was generally obtained for threshold exceedances of 1-h QPEs and maximum correspondence for 24-h QPE exceedances. On a regional basis, there

were exceptions where shorter accumulations provided more skillful predictions, particularly in the arid Southwest.

- Also surprisingly, FT exceedances provided slightly superior correspondence to FFRs and FFWs compared with FFG or ARI exceedances when a uniform threshold method was applied across the CONUS. Overall, 2.5 in. day^{-1} provided the best correspondence with FFRs and FFWs of any threshold QPE exceedance examined, although 2.0 in. $(6\,\text{h})^{-1}$ and others provided nearly equal ETS.
- In some regions of the CONUS, FFG and/or ARI exceedances outperformed any FT exceedance, but the optimal ARI varied between 1 and 5 years, and occasionally higher for certain subregions, such as Florida and New Mexico.
- Among ARIs, the 1-yr ARI provided the best predictions of FFRs across the CONUS. For ST4 and MRMS, 24-h accumulations performed best; for CCPA, 6-h accumulations performed better. Among FFGs, 6-h FFGs provided the best correspondence, but agreement was appreciably worse than with ARIs, which were in turn worse than the FT exceedances when applied uniformly across the CONUS.
- Via their warnings, the NWS is able to add substantial value over automated QPE exceedances in projecting where heavy rain will produce reported flash flooding.

There are several limitations worthy of reemphasizing. Ultimately, this study has examined how well different QPE threshold exceedances correspond with flash flood *reports* (or warnings) and not true flash floods. FFRs have numerous nonphysical reporting biases, including a tendency to underreport flash floods in rural areas and at night. FFR frequency also varies by encoding practices of local WFOs, with some preferring to encode as a flood what another office may encode as a flash flood. A perfect ‘‘truth’’ does not exist, a fact which also serves as much of the motivation for conducting this comparative analysis. Compared with true incidences of flash flooding, FFRs likely underrepresent the true flash flooding climatology due to the aforementioned reporting and encoding practices. In particular, FFRs likely have few false alarms—most reports are indeed true events—but have numerous missed events. As a result, while verification is traditionally treated symmetrically, as it is also in this study, there is reason to believe those comparisons when evaluated against FFRs with frequency biases above unity likely have better correspondence with true flash flooding than those with biases below unity for the same CSI or ETS. It may be desirable in future work to incorporate the uncertainty of observations in the evaluation framework, penalizing nonhits in densely populated areas more than those in rural ones where ‘‘truth’’ is more uncertain, similar to that suggested in Weijs and Van De Giesen (2011) and elsewhere. Relatedly, some of the improved correspondence to FFRs illustrated by the FFWs over QPE threshold exceedances is likely artificial. WFOs have different proclivities to warn flash floods and can, for example, choose to not warn storms that are likely to produce unreported flash flooding, such as those confined to highly remote areas. They can also adopt different practices on encoding reports and adopt different verification practices for warned and unwarned events (e.g., Barnes et al. 2007).

Nevertheless, there are several important implications from this analysis. Several prominent deficiencies are observed for each QPE source—some deficiencies are found in common between data sources, while others are unique to a particular source. In particular, all sources struggle with QPEs in the West. ST4 suffers from spurious very high values in areas of complex terrain, particularly in its 1-h QPEs. This is especially prominent in the complex terrain of New Mexico. This phenomenon is seen, albeit to a lesser extent, across the rest of the CONUS as well. MRMS has the same spurious high values over New Mexico, most prominently seen with ARI exceedances. While the root issues likely share commonalities with the ST4 deficiencies, MRMS appears to suffer from another major issue. Across the West, extreme QPEs occur with much larger frequency near radar sites compared with more distant locations—surely an artifact of the QPE derivation rather than a true spatially varying climatology given the number of sites exhibiting this behavior and the extent of correspondence. CCPA alleviates many of these problems but removes many extreme events correctly identified by both ST4 and MRMS. This deficiency is prominent across all of the CONUS. The lack of 1-h CCPA QPEs also limit its utility in identifying flash flood scenarios in the SW and other regions where the best QPE-FFR correspondence was identified for shorter accumulation intervals. Developers of QPE products may wish to further investigate some of these identified issues and adopt methods to alleviate them in order to generate more accurate and operationally useful products. A number of measures may assist with this, including improved quality control, particularly in the West, and statistical corrections tailored specifically for extremes, perhaps as a function of radar distance for MRMS, and to counteract the linear corrections made that necessarily and undesirably regress toward the climatological mean in the case of CCPA. Last, the verification in this study was limited to daily 1200–1200 UTC time scales. While this does not directly harm the verification performance of shorter AI exceedances compared with longer ones, the verification framework does not account for the fact that shorter AI QPE exceedances may provide additional information about the timing of flash flooding that the longer AI QPE exceedances cannot. Flash flood timing can be an important component of flash flood analysis and gives the shorter AI QPE exceedances an advantage unaccounted for in this study's verification framework.

The analysis also lends some insight into current deficiencies with the ARIs. For example, there being nearly an order of magnitude more ARI exceedances in all three QPE sources over Wyoming and to a lesser extent in Montana, when ostensibly they should be the same everywhere over an infinitely long period of record, indicates that the ARI threshold estimates from the old NOAA Atlas 2 are likely too low in this region. Many of these places are very rural—and even more so prior to Miller et al. (1973)—and the threshold estimates were likely inaccurate and highly uncertain in these areas at the time the threshold estimates were derived due to lack of sufficient robust observational data in these areas. It is expected that updated estimates through the NOAA Atlas 14 project will likely increase over the prior NOAA Atlas 2 estimates in these areas. In contrast, CCPA did not exhibit the high bias in New Mexico seen with MRMS and ST4 QPEs and were also seen for other threshold sources, suggesting that the issues encountered in that region are more likely attributable to QPE limitations in those two sources rather than a fundamental ARI threshold estimate issue in this area.

The low frequency biases observed comparing FFG exceedances with FFRs seen across much of the CONUS suggests that FFGs may be too high in many situations. It may be advisable to consider FFG calculation practices and evaluate whether any revisions can be made to increase spatial homogeneity across RFC boundaries and lower thresholds when appropriate to improve bias characteristics with respect to reported flash floods, similar to recommendations made in recent decades (Sweeney 1992; Carpenter et al. 1999). The findings of this study also raise implications about operational flash flood forecasting across a range of time scales. On longer time scales, for example, the Weather Prediction Center issues Excessive Rainfall Outlooks providing probabilistic guidance across the CONUS for the current day out to 2 days ahead that sufficiently heavy rainfall will occur to produce flash flooding. Currently, forecast probabilities are defined with respect to exceedances of FFG. This provides a concrete framework for evaluating their outlooks and avoids many of the pitfalls associated with directly using FFRs or similar observational sources. The findings of this study suggest that, contrary to conventional wisdom, the product may have more utility in relating directly to precipitation impacts if instead one defines the outlook with respect to longer accumulation intervals, such as 6- or 24-h QPE exceedances compared with the 1- and 3-h FFG exceedances used in operations, and perhaps even using a homogeneous threshold, such as 2.5 in. (63.5 mm) day^{-1}. At shorter time scales, the findings further suggest that in assessing flash flood potential from a warning perspective, operational forecasters may wish to rely more heavily on one QPE source than another depending on their location. In the East, forecasters, researchers, and model developers may wish to employ ST4 QPE more heavily, while relying more on MRMS QPE in the West and CCPA QPE across the Great Plains. Last, the findings shed insight into how QPF verification (e.g., Herman and Schumacher 2016) and heavy precipitation forecast product development (e.g., Herman and Schumacher 2018a,b) may be conducted to be more physically relevant toward the impacts of heavy rainfall. Because of latency in the generation of some analysis products evaluated in this study, particularly ST4 and CCPA, the largest operational benefits may come from improvements to precipitation forecast and analysis products rather than direct changes to forecaster practices.

New flash flood analysis tools such as those described in Gourley et al. (2017), which use hydrologic models to provide additional insights, are becoming available in forecast operations. These tools have the potential to instill hydrometeorological insights beyond what can be gleaned from a simple inspection of QPE with respect to a threshold or thresholds. However, even in this framework, hydrologic guidance is only useful to the extent that its QPE input is accurate. It is hoped that the findings from this study helped to identify specific issues and areas the QPE products can be improved to alleviate recurring errors and biases, resulting in more representative outputs from analysis tools based on QPEs. In the meantime, knowledge and quantification of these deficiencies can improve human interpretation of derived analysis products by increasing (decreasing) confidence in areas that QPE is (not) skillful and damping (raising) perceived risk in areas that systematically have QPEs that are too high (low).

More investigation is required to further validate and constrain these findings. In addition, this work has not attempted to combine information from different sources to provide better correspondence between QPE exceedances and flash flood observations. Future work should examine these joint distributions to ascertain whether the full suite of QPE information can be used more effectively for flash flood forecasting and analysis. This study also did not attempt to recalibrate QPEs with the specific focus of removing apparent systematic biases and improving their overall accuracy in heavy precipitation scenarios. Producing a CCPA-like correction to ST4 QPE, but employing different methodology geared toward the tail of the QPE distribution rather than the entire distribution, would likely be a worthwhile and fruitful endeavor.

Acknowledgments. The authors greatly thank Erik Nielsen for assisting in the creation of Figs. 3, 4, 5, and 11. We also wish to thank Diana Stovern, Sarah Perfater, Benjamin Albright, Mark Klein, Michael Erickson, and James Nelson at the Weather Prediction Center whose input helped improve the operational utility of this research. Discussions with Scott Lincoln, Tom Hamill, Kelly Mahoney, Rob Cifelli, and Janice Bytheway on flash flood reporting practices have also greatly benefited this research. The authors also wish to thank Joe Turk and three anonymous reviewers for many constructive comments which helped improve the quality of this manuscript. Funding for this research was supported by NOAA Award NA16OAR4590238.

APPENDIX A

Acronyms

AI	Accumulation interval
ARI	Average recurrence interval
CCPA	Climatology Calibrated Precipitation Analysis
CSI	Critical success index
CWA	County warning area
ETS	Equitable threat score
FB	Frequency bias
FFG	Flash flood guidance
FFR	Flash flood report
FFW	Flash flood warning
FT	Fixed threshold
GIS	Geographical information system
HRAP	Hydrologic Rainfall Analysis Project
IEM	Iowa Environmental Mesonet
LSR	Local storm report
MRMS	Multi-Radar Multi-Sensor
PD	Performance diagram
POD	Probability of detection
QPE	Quantitative precipitation estimate
RFC	River forecast center
SR	Success ratio
ST4	Stage IV Precipitation Analysis
WFO	Weather forecast office

APPENDIX B

Handling of Missing ARI Threshold Estimates

In NOAA Atlas 14, a generalized extreme value distribution is fit to an annual maximum series for each duration independently; the results are related to the extent that the underlying data are the same (e.g., a 3-h accumulation is composed of 1-h accumulations), but ARI thresholds for different AIs are not directly computed in tandem (e.g., Bonnin et al. 2011). Here, however, where Atlas 14 estimates have not yet been officially generated and the original underlying data had insufficient temporal resolution, a relationship must be derived between the threshold estimates that are available and the desired, unknown thresholds at shorter durations. Accordingly, an analytic equation is derived to relate the 6- and 24-h thresholds for a given ARI to 3- and 1-h estimates. The formula is composed of two components. One term is designed to exactly obey desired mathematical properties; the second, tunable term alters the formula to match the known relationships—where Atlas 14 estimates are available for all AIs—as well as possible while obeying the mathematical properties to the extent possible. Desired mathematical properties include 1) threshold estimates go to zero in the limit as AI goes to zero; 2) threshold estimates go to infinity in the limit as AI goes to infinity; 3) the formula is valid for any positive AI; 4) the rate of change of threshold magnitude with increasing AI decreases with increasing AI; 5) when the ratio of known threshold estimates for two different AIs is exactly equal the ratio of those AIs, the threshold estimate for an AI with an equal ratio with one of the AIs with a known threshold should exactly preserve the same ratio with the threshold estimate corresponding to that AI (e.g., if a 6-h estimate is 10 and 24-h estimate is 20, a 1.5-h estimate should be 5); 6) using the formula to derive thresholds for one of the two known AIs being used returns those same threshold estimates; 7) the formula is reversible: it can be used to derive a third "known" estimate, and the use of any two can then be used to exactly recover the third; and 8) an arbitrary number of intermediate threshold estimates can be derived without altering the estimate for a given AI (e.g., deriving a 3-h estimate from 6- and 24-h estimates, and then using 3- and 6-h estimates to derive a 1-h estimate will produce the same result as deriving 1-h estimates from the 6- and 24-h values). It can be easily shown that for shorter AI S and longer AI L with known threshold estimates Θ_S and Θ_L, an equation for deriving an unknown estimate Θ_N for AI N that satisfies all of these properties is

$$\Theta_N = \Theta_S\left[\left(\frac{\Theta_L}{\Theta_S}\right)^{\log_{L/S}(S/N)}\right]^{-1} . \qquad \text{(B1)}$$

The tunable term is constructed as

$$\left[\log_{L/S}\left(\frac{\sqrt{SN}}{\frac{S\sqrt{SL}}{L}}\right)-1\right]\log_{L/S}\left(\frac{S}{N}\right)\alpha\,, \qquad \text{(B2)}$$

for tunable parameter α. That term is further decomposed into

$$\alpha=\beta\left(\frac{\Theta_S}{\overline{\Theta}_S}\right)^{\kappa_S}\left(\frac{\Theta_L}{\overline{\Theta}_L}\right)^{\kappa_L} . \qquad \text{(B3)}$$

Tuning in areas where Atlas 14 estimates are available and thus "truth" is known yielded

$$\beta=\frac{4}{3}(1-\log_{10}\text{ARI});\quad \kappa_S=1.7;\quad \kappa_L=0.6\,. \qquad \text{(B4)}$$

REFERENCES

AghaKouchak, A., A. Behrangi, S. Sorooshian, K. Hsu, and E. Amitai, 2011: Evaluation of satellite-retrieved extreme precipitation rates across the central United States. *J. Geophys. Res.*, **116**, D02115, https://doi.org/10.1029/2010JD014741.

Ashley, S. T., and W. S. Ashley, 2008: Flood fatalities in the United States. *J. Appl. Meteor. Climatol.*, **47**, 805–818, https://doi.org/10.1175/2007JAMC1611.1.

Barnes, L. R., E. C. Gruntfest, M. H. Hayden, D. M. Schultz, and C. Benight, 2007: False alarms and close calls: A conceptual model of warning accuracy. *Wea. Forecasting*, **22**, 1140–1147, https://doi.org/10.1175/WAF1031.1.

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experiment. *Bull. Amer. Meteor. Soc.*, **96**, 1859–1866, https://doi.org/10.1175/BAMS-D-14-00201.1.

Bonnin, G. M., D. Martin, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2006: Version 3.0: Delaware, District of Columbia, Illinois, Indiana, Kentucky, Maryland, New Jersey, North Carolina, Ohio, Pennsylvania, South Carolina, Tennessee, Virginia, West Virginia. Vol. 2, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 295 pp., http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume2.pdf.

——, ——, ——, ——, ——, and ——, 2011: Version 5.0: Semiarid Southwest (Arizona, Southeast California, Nevada, New Mexico, Utah). Vol. 1, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, Vol. 1, 265 pp., http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume1.pdf.

Brocca, L., F. Melone, and T. Moramarco, 2008: On the estimation of antecedent wetness conditions in rainfall–runoff modelling. *Hydrol. Processes*, **22**, 629–642, https://doi.org/10.1002/hyp.6629.

Brooks, H. E., and D. J. Stensrud, 2000: Climatology of heavy rain events in the United States from hourly precipitation observations. *Mon. Wea. Rev.*, **128**, 1194–1201, https://doi.org/10.1175/1520-0493(2000)128<1194:COHREI>2.0.CO;2.

Calianno, M., I. Ruin, and J. J. Gourley, 2013: Supplementing flash flood reports with impact classifications. *J. Hydrol.*, **477**, 1–16, https://doi.org/10.1016/j.jhydrol.2012.09.036.

Carpenter, T., J. Sperfslage, K. Georgakakos, T. Sweeney, and D. Fread, 1999: National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *J. Hydrol.*, **224**, 21–44, https://doi.org/10.1016/S0022-1694(99)00115-8.

Castillo, V., A. Gomez-Plaza, and M. Martınez-Mena, 2003: The role of antecedent soil water content in the runoff response of semiarid catchments: A simulation approach. *J. Hydrol.*, **284**, 114–130, https://doi.org/10.1016/S0022-1694(03)00264-6.

Clark, R. A., J. J. Gourley, Z. L. Flamig, Y. Hong, and E. Clark, 2014: CONUS-wide evaluation of National Weather Service flash flood guidance products. *Wea. Forecasting*, **29**, 377–392, https://doi.org/10.1175/WAF-D-12-00124.1.

Davis, R. S., 2001: Flash flood forecast and detection methods. *Severe Convective Storms*, Springer, 481–525.

Doswell, C. A., III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581, https://doi.org/10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

Edwards, R., G. W. Carbin, and S. F. Corfidi, 2015: Overview of the Storm Prediction Center. *13th History Symp.*, Phoenix, AZ, Amer. Meteor. Soc., 1.1, https://ams.confex.com/ams/95Annual/webprogram/Paper266329.html.

Ferree, J. T., J. Looney, and K. Waters, 2006: NOAA/National Weather Services' storm-based warnings. *23rd Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., 11.6, https://ams.confex.com/ams/23SLS/techprogram/paper_115513.htm.

Fritsch, J. M., and R. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85**, 955–965, https://doi.org/10.1175/BAMS-85-7-955.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. O'Bannon, 1998: The WSR-88D rainfall algorithm. *Wea. Forecasting*, **13**, 377–395, https://doi.org/10.1175/1520-0434(1998)013<0377:TWRA>2.0.CO;2.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370, https://doi.org/10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2.

Gourley, J. J., J. M. Erlingis, Y. Hong, and E. B. Wells, 2012: Evaluation of tools used for monitoring and forecasting flash floods in the United States. *Wea. Forecasting*, **27**, 158–173, https://doi.org/10.1175/WAF-D-10-05043.1.

——, and Coauthors, 2013: A unified flash flood database across the United States. *Bull. Amer. Meteor. Soc.*, **94**, 799–805, https://doi.org/10.1175/BAMS-D-12-00198.1.

——, and Coauthors, 2017: The FLASH project: Improving the tools for flash flood monitoring and prediction across the United States. *Bull. Amer. Meteor. Soc.*, **98**, 361–372, https://doi.org/10.1175/BAMS-D-15-00247.1.

Hapuarachchi, H., Q. Wang, and T. Pagano, 2011: A review of advances in flash flood forecasting. *Hydrol. Processes*, **25**, 2771–2784, https://doi.org/10.1002/hyp.8040.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, https://doi.org/10.1175/WAF-D-16-0093.1.

——, and ——, 2018a: ‘‘Dendrology’’ in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.*, **146**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1.

——, and ——, 2018b: Money doesn’t grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

——, E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of Storm Prediction Center convective outlooks. *Wea. Forecasting*, **33**, 161–184, https://doi.org/10.1175/WAF-D-17-0104.1.

Hershfield, D. M., 1961: Rainfall frequency atlas of the United States: For durations from 30 minutes to 24 hours and return periods from 1 to 100 years. U.S. Weather Bureau Tech. Paper 40, 65 pp., http://www.nws.noaa.gov/oh/hdsc/PF_documents/TechnicalPaper_No40.pdf.

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and temporal characteristics of heavy hourly rainfall in the United States. *Mon. Wea. Rev.*, **141**, 4564–4575, https://doi.org/10.1175/MWR-D-12-00297.1.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, https://doi.org/10.1175/JHM-D-11-0140.1.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner’s Guide in Atmospheric Science*. John Wiley & Sons, 292 pp.

Kunkel, K. E., D. R. Easterling, D. A. Kristovich, B. Gleason, L. Stoecker, and R. Smith, 2012: Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States. *J. Hydrometeor.*, **13**, 1131–1141, https://doi.org/10.1175/JHM-D-11-0108.1.

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, https://ams.confex.com/ams/Annual2005/techprogram/paper_83847.htm.

Marjerison, R. D., M. T. Walter, P. J. Sullivan, and S. J. Colucci, 2016: Does population affect the location of flash flood reports? *J. Appl. Meteor. Climatol.*, **55**, 1953–1963, https://doi.org/10.1175/JAMC-D-15-0329.1.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753–763, https://doi.org/10.1175/1520-0434(1998)013<0753:SMOPIR>2.0.CO;2.

Meierdiercks, K. L., J. A. Smith, M. L. Baeck, and A. J. Miller, 2010: Analyses of urban drainage network structure and its impact on hydrologic response. *J. Amer. Water Resour. Assoc.*, **46**, 932–943, https://doi.org/10.1111/j.1752-1688.2010.00465.x.

Miller, J., R. Frederick, and R. Tracey, 1973: Montana. Vol. 1, Precipitation-Frequency Atlas of the Western United States, NOAA Atlas 2, 33 pp., http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas2_Volume1.pdf.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, https://doi.org/10.1175/WAF-D-14-00112.1.

Nielsen, E. R., G. R. Herman, R. C. Tournay, J. M. Peters, and R. S. Schumacher, 2015: Double impact: When both tornadoes and flash floods threaten the same place at the same time. *Wea. Forecasting*, **30**, 1673–1693, https://doi.org/10.1175/WAF-D-15-0084.1.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504, https://doi.org/10.1175/WAF-D-13-00066.1.

Ntelekos, A. A., K. P. Georgakakos, and W. F. Krajewski, 2006: On the uncertainties of flash flood guidance: Toward probabilistic forecasting of flash floods. *J. Hydrometeor.*, **7**, 896–915, https://doi.org/10.1175/JHM529.1.

NWS, 2017a: Service change notice 17-100. National Centers for Environmental Prediction, Weather Prediction Center, http://www.nws.noaa.gov/os/notification/scn17-100wpc_excessive_rainfall.htm.

——, 2017b: Summary of natural hazard statistics in the United States. National Weather Service, Office of Climate, Weather, and Water Services, http://www.nws.noaa.gov/om/hazstats.shtml.

Ogden, F., H. Sharif, S. Senarath, J. Smith, M. Baeck, and J. Richardson, 2000: Hydrologic analysis of the Fort Collins, Colorado, flash flood of 1997. *J. Hydrol.*, **228**, 82–100, https://doi.org/10.1016/S0022-1694(00)00146-3.

Perica, S., and Coauthors, 2011: Version 2.3: California. Vol. 6, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 233 pp., http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume6.pdf.

——, and Coauthors, 2013: Version 2.0: Southeastern States (Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi). Vol. 9, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 163 pp., http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume9.pdf.

——, S. Pavlovic, M. S. Laurent, C. Trypaluk, D. Unruh, D. Martin, and O. Wilhite, 2015: Version 2.0: Northeastern States (Connecticut, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Vermont). Vol. 10, Precipitation-Frequency Atlas of the United States, NOAA Atlas 14, 4 pp., http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume10.pdf.

Pielke, R. A., Jr., M. W. Downton, and J. Z. Barnard Miller, 2002: Flood damage in the United States, 1926–2000: A reanalysis of National Weather Service estimates. UCAR Rep., 96 pp., http://sciencepolicy.colorado.edu/flooddamagedata/full_report.html.

Ramshaw, J. D., 1985: Conservative rezoning algorithm for generalized two-dimensional meshes. *J. Comput. Phys.*, **59**, 193–199, https://doi.org/10.1016/0021-9991(85)90141-X.

Reed, S., J. Schaake, and Z. Zhang, 2007: A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *J. Hydrol.*, **337**, 402–420, https://doi.org/10.1016/j.jhydrol.2007.02.015.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Rutz, J. J., W. J. Steenburgh, and F. M. Ralph, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142**, 905–921, https://doi.org/10.1175/MWR-D-13-00168.1.

Schmidt, J. A., A. Anderson, and J. Paul, 2007: Spatially-variable, physically-derived flash flood guidance. *21st Conf. on Hydrology*, San Antonio, TX, Amer. Meteor. Soc., 6B.2, https://ams.confex.com/ams/87ANNUAL/techprogram/paper_120022.htm.

Schroeder, A. J., and Coauthors, 2016: The development of a flash flood severity index. *J. Hydrol.*, **541**, 523–532, https://doi.org/10.1016/j.jhydrol.2016.04.005.

Schumacher, R. S., 2017: Heavy rainfall and flash flooding. Natural Hazard Science, Oxford Research Encyclopedias, https://doi.org/10.1093/acrefore/9780199389407.013.132.

——, and R. H. Johnson, 2005: Organization and environmental properties of extreme-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **133**, 961–976, https://doi.org/10.1175/MWR2899.1.

——, and ——, 2006: Characteristics of U.S. extreme rain events during 1999–2003. *Wea. Forecasting*, **21**, 69–85, https://doi.org/10.1175/WAF900.1.

Smith, J. A., A. J. Miller, M. L. Baeck, P. A. Nelson, G. T. Fisher, and K. L. Meierdiercks, 2005: Extraordinary flood response of a small urban watershed to short-duration convective rainfall. *J. Hydrometeor.*, **6**, 599–617, https://doi.org/10.1175/JHM426.1.

Stevenson, S. N., and R. S. Schumacher, 2014: A 10-year survey of extreme rainfall events in the central and eastern United States using gridded multisensor precipitation analyses. *Mon. Wea. Rev.*, **142**, 3147–3162, https://doi.org/10.1175/MWR-D-13-00345.1.

Sweeney, T. L., 1992: Modernized areal flash flood guidance. NOAA Tech. Memo NWS HYDRO 44, 37 pp., https://repository.library.noaa.gov/view/noaa/13498.

Versini, P.-A., E. Gaume, and H. Andrieu, 2010: Application of a distributed hydrological model to the design of a road inundation warning system for flash flood prone areas. *Nat. Hazards Earth Syst. Sci.*, **10**, 805, https://doi.org/10.5194/nhess-10-805-2010.

Villarini, G., W. F. Krajewski, A. A. Ntelekos, K. P. Georgakakos, and J. A. Smith, 2010: Towards probabilistic forecasting of flash floods: The combined effects of uncertainty in radar-rainfall and flash flood guidance. *J. Hydrol.*, **394**, 275–284, https://doi.org/10.1016/j.jhydrol.2010.02.014.

Waters, K., and Coauthors, 2005: Polygon weather warnings: A new approach for the National Weather Service. *21st Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, Amer. Meteor. Soc., 14.1, https://ams.confex.com/ams/Annual2005/techprogram/paper_86326.htm.

Weijs, S. V., and N. Van De Giesen, 2011: Accounting for observational uncertainty in forecast verification: An information-theoretical view on forecasts, observations, and truth. *Mon. Wea. Rev.*, **139**, 2156–2162, https://doi.org/10.1175/2011MWR3573.1.

Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action and collaboration. *Bull. Amer. Meteor. Soc.*, **88**, 503–511, https://doi.org/10.1175/BAMS-88-4-503.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wolff, A., 2013: Simulation of pavement surface runoff using the depth-averaged shallow water equations. Ph.D. thesis, University of Stuttgart, 174 pp., https://elib.uni-stuttgart.de/handle/11682/505.

Wood, E. F., 1976: An analysis of the effects of parameter uncertainty in deterministic hydrologic models. *Water Resour. Res.*, **12**, 925–932, https://doi.org/10.1029/WR012i005p00925.

Yatheendradas, S., T. Wagener, H. Gupta, C. Unkrich, D. Goodrich, M. Schaffner, and A. Stewart, 2008: Understanding uncertainty in distributed flash flood forecasting for semiarid regions. *Water Resour. Res.*, **44**, W05S19, https://doi.org/10.1029/2007WR005940.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.