Evaluation of Tools Used for Monitoring and Forecasting Flash Floods in the United States

JONATHAN J. GOURLEY

NOAA/National Severe Storms Laboratory, Norman, Oklahoma

JESSICA M. ERLINGIS

*NOAA/National Severe Storms Laboratory, and Cooperative Institute for Mesoscale Meteorological Studies,
University of Oklahoma, Norman, Oklahoma*

YANG HONG

Department of Civil Engineering and Environmental Science, University of Oklahoma, Norman, Oklahoma

ERNEST B. WELLS

Office of Climate, Water, and Weather Services, National Weather Service, Silver Spring, Maryland

(Manuscript received 4 November 2010, in final form 11 July 2011)

ABSTRACT

This paper evaluates, for the first time, flash-flood guidance (FFG) values and recently developed gridded FFG (GFFG) used by the National Weather Service (NWS) to monitor and predict imminent flash flooding, which is the leading storm-related cause of death in the United States. It is envisioned that results from this study will be used 1) to establish benchmark performance of existing operational flash-flood prediction tools and 2) to provide information to NWS forecasters that reveals how the existing tools can be readily optimized. Sources used to evaluate the products include official reports of flash floods from the NWS *Storm Data* database, discharge measurements on small basins available from the U.S. Geological Survey, and witness reports of flash flooding collected during the Severe Hazards Analysis and Verification Experiment. Results indicated that the operational guidance values, with no calibration, were marginally skillful, with the highest critical success index of 0.20 occurring with 3-h GFFG. The false-alarm rates fell and the skill improved to 0.34 when the rainfall was first spatially averaged within basins and then reached 50% of FFG for 1-h accumulation and exceeded 3-h FFG. Although the skill of the GFFG values was generally lower than that of their FFG counterparts, GFFG was capable of detecting the spatial variability of reported flash flooding better than FFG was for a case study in an urban setting.

1. Introduction

In the United States, flash flooding is the number-one cause of death among all storm-related hazards, with approximately 100 lives lost each year (Ashley and Ashley 2008). Despite their considerable impacts on life and infrastructure, flash floods are poorly observed relative to other weather-related hazards (Gruntfest 2009). However, big advances in remote sensing of precipitation

have occurred in the Weather Surveillance Radar-1988 Doppler (WSR-88D) era. The Next Generation Weather Radar (NEXRAD) network has provided observations of intense rainfall rates at scales on the order of 1 km², every 5 min. It is thus possible to incorporate these high-resolution rainfall estimates into a prediction system that considers additional factors that control the initiation of flash flooding, including soil moisture states, slope of the underlying terrain, impervious surfaces in developed zones, and so on. The current method for operational flash-flood forecasting at the National Weather Service (NWS) utilizes the Flash Flood Monitoring and Prediction (FFMP) software package to compare WSR-88D rainfall estimates with flood-induced rainfall accumulation

Corresponding author address: Jonathan J. Gourley, National Weather Center, 120 David L. Boren Blvd., Norman, OK 73072-7303.

E-mail: jj.gourley@noaa.gov

thresholds, known as flash-flood guidance (FFG) values (Sweeney 1992). The success of FFMP depends on both the accuracy of the radar-estimated rainfall rates and the FFG values. Numerous studies have evaluated radar rainfall estimates [see a summary in Krajewski et al. (2010)], but quantitative verification of forecasting procedures using FFG and FFMP together is very limited. Davis (2003, 2004) reports on a small number of case studies that evaluate FFMP and FFG in Pennsylvania.

The purpose of this study is to benchmark the performance of FFG and the gridded approach developed at the Arkansas–Red Basin River Forecast Center (ABRFC) using observations of flash floods from 1) official spotter reports collected in the NWS *Storm Data* database, 2) discharge measurements at U.S. Geological Survey (USGS) stream gauges in small basins, and 3) witness reports collected during the Severe Hazards Analysis and Verification Experiment (SHAVE) at the National Severe Storms Laboratory (Ortega et al. 2009). None of these flood observation databases provides a comprehensive identification of all flash-flood events, but use of the databases in combination yields useful information. Establishing the present skill of FFG and gridded FFG (GFFG) will help to guide further development and operational implementation of improved guidance values using approaches outlined in Schmidt et al. (2007) and Reed et al. (2007). The ABRFC with its generally good radar coverage is a good choice for minimizing the errors in quantitative precipitation estimation (QPE) so that we can focus the study on the performance of FFG and GFFG. The next section provides an overview of two methods used operationally at the NWS to derive guidance values. Section 3 discusses the study domain and datasets. Guidance values are compared with each other in section 4, and respective skill scores are quantified using the assembled flash-flood observation databases. Section 5 summarizes the results and provides conclusions.

2. Operational flash-flood forecasting

Hong et al. (2011) define flash floods as follows: “A flash flood is a rapid flooding of water over land caused by heavy rain or a sudden release of impounded water (e.g., dam or levee break) in a short period of time, generally within minutes up to several hours, a time scale that distinguishes it from fluvial floods.” At the NWS, a 6-h basin response time is generally used to divide responsibilities between local Weather Forecast Offices (WFOs) and River Forecast Centers (RFCs), and as the primary defining characteristic to distinguish between flash floods and floods. This is reflected in the NWS definition of a flash flood as “a rapid and extreme flow of high water into a normally dry area, or a rapid water level rise in a stream

or creek above a predetermined flood level, beginning within six hours of the causative event (e.g., intense rainfall, dam failure, ice jam).” This study focuses on the FFG and GFFG predictors for rainfall-induced events alone.

In NWS operations, WFOs issue all flash-flood watches, warnings, and urban/small stream advisories in their respective County Warning Areas. RFCs produce stage forecasts for river locations on larger basins that respond more slowly to rainfall (i.e., >6 h). Typically, no specific forecasts are provided for locations in/on these smaller basins where the response time is less than 6 h. In the absence of point-specific forecasts for these faster-responding basins, RFCs compute the FFG thresholds and deliver them to the local WFOs within their regions of responsibility. WFOs utilize these FFG thresholds as primary decision assistance in fulfilling their responsibility to issue flash-flood watch, warning, and urban/small stream advisory products. Below, in section 2a, we describe the FFG system that followed the NWS modernization recommendations for areal FFG in Sweeney (1992) and that was implemented throughout the NWS. Recently, RFCs have developed and operationally implemented newer, higher-resolution estimates of FFG, which we refer to hereinafter as gridded FFG or GFFG. We refer to the work of Smith (2003) for the derivation of flash-flood potential maps that are used predominantly in western RFCs (e.g., Colorado Basin RFC).

a. Flash-flood guidance system

The primary indicator used by forecasters at WFOs to predict the onset of flash flooding is when radar-based rainfall estimates exceed FFG values over durations of 1, 3, or 6 h. FFG is defined as the threshold rainfall required to initiate flooding on small streams that respond to rainfall within a few hours (Georgakakos 1986; Sweeney 1992). The primary components to computing FFG are threshold runoff (Thresh-R) calculations, snow-melt calculations using the Snow-17 model, and runoff calculations using the Sacramento Soil Moisture Accounting model (SAC-SMA; Anderson 1973; Burnash 1995; Carpenter et al. 1999). Thresh-R is defined very similarly to FFG, but it is the amount of rainfall *excess* required to cause flooding at the basin outlet; rainfall *excess* is the effective volume of rainfall that is transformed into surface runoff at the basin scale. This amount is either equal to or, most often, less than the total rainfall due to infiltration, storage, interception by vegetation, and evaporation. These processes are modeled in real time using Snow-17 and SAC-SMA, whereas Thresh-R values are assumed to be constant and are thus computed once in an offline mode.

Carpenter et al. (1999) discuss four possible methods for computing Thresh-R based on two ways to estimate

a catchment's threshold flow that results in flooding and two ways to estimate a catchment's linear response to rainfall. Here, we focus on the methods used operationally by the NWS to derive FFG values. Thresh-R (in length units) is computed by dividing an estimate of bankfull discharge (length cubed divided by time), which is typically assumed to be the exceedance of the 2-yr return-period flow, by the unit hydrograph peak flow (length cubed divided by time divided by length). For initial derivation of county-wide FFG values, RFCs derived Thresh-R values at selected watershed outlets with USGS stream gauges and then used manual contouring to fill in geographic regions between the selected outlets. Subsequently, GIS-based methods to estimate Thresh-R values at finer resolutions were developed (Carpenter et al. 1999; Reed et al. 2002). The GIS tools allowed estimation of Thresh-R for basins as small as 5 km^2 , which were then resampled onto a rectangular grid at Hydrologic Rainfall Analysis Project (HRAP) resolution in a polar stereographic projection. The nominal size of an HRAP grid cell is $4 \times 4 \text{ km}^2$, but the true cell size varies with latitude (Reed and Maidment 1999). These Thresh-R values are computed once and then stored at each HRAP grid cell.

In NWS operations, hydrologists at RFCs regularly run the lumped-parameter SAC-SMA model on basins with catchments on the order of 1000 km^2 under iterative rainfall scenarios to produce rainfall-runoff curves. Because the SAC-SMA is a lumped model, this method assumes the rainfall is uniformly distributed in space and time over the basin. The Snow-17 model is also run to compute the contribution to runoff from snow accumulation and subsequent melting. The curves are derived for 1-, 3-, and 6-h accumulation periods. The curves are used to determine the amount of uniform rainfall (FFG) (in length units) for each accumulation period required to exceed Thresh-R under the current soil moisture conditions. Typically, the FFG values are updated 1–3 times per day depending on specific RFC operating practices. The soil moisture states are initialized with each model run and impact the depth of rainfall required to become surface runoff and ultimately flood-producing runoff. The scale mismatch between precipitation-runoff calculations (basins on the order of 1000 km^2) and Thresh-R calculations (basins as small as 5 km^2) motivated the development of the GFFG as described in the next section. FFG values are transmitted to local WFOs and are compared with real-time, WSR-88D-based rainfall estimates to support their mission in warning the public of impending flash floods. Forecasters at WFOs utilize direct comparisons (difference and ratio) of WSR-88D rainfall estimates and the RFC-issued FFG thresholds as primary decision support when issuing flash-flood warnings.

Since the 1970s, FFG has been the primary decision assistance tool used by NWS forecasters to evaluate flash-flood threat potential (Mogil et al. 1978). Uncertainties in FFG values have been addressed by Georgakakos (2006) who computed analytical results that compare SAC-SMA surface runoff with FFG rainfall thresholds. Ntelekos et al. (2006) quantified the uncertainty in deriving Thresh-R values and in SAC-SMA predictions to demonstrate a probabilistic approach toward FFG. An alternative to using an observation-based 2-yr return flow to define the threshold flow in Thresh-R calculations was proposed by Reed et al. (2007). The threshold frequency approach relies on model-simulated return flows. Provided a long-term rainfall record, one runs a distributed hydrologic model retrospectively and computes a flood frequency analysis at interior basins where there are no discharge measurements. Moreover, the method removes model bias by relying on the relative ranking of the events rather than absolute discharge values. Norbiato et al. (2009) used this threshold frequency approach to derive FFG and compared model skill relative to FFG based on Thresh-R values from observed flows. They demonstrated an improvement in the critical success index (CSI) from 0.34 to 0.5 using the model-simulated return flows for Thresh-R.

Despite the preponderance of FFG approaches used for flash-flood warning in the NWS, Central America, and Europe, extensive verification of their skill is lacking. Schmidt et al. (2007) refer to a final report from the RFC Development Management Team in 2003 that states that there exists no verification system to provide direct feedback on the accuracy of the FFGs issued by RFCs. Recent studies in Europe give some indications of how FFG might perform in the United States. Norbiato et al. (2008) developed FFG values in a manner that was similar to what is done in the NWS but replaced the SAC-SMA hydrologic model with the probability distributed moisture model described in Moore (1985). The skill of FFG was assessed in basins in northeastern Italy and central France. The skill of FFG evaluated on the larger parent basins, where the hydrologic model had been calibrated, was quantified with a CSI of 0.43. The CSI value deteriorated to 0.22, however, when parent basin parameters and soil moisture states were transposed to interior basins. This latter skill score of 0.22 will serve as a benchmark hereinafter for evaluating NWS FFG skill.

b. Gridded flash-flood guidance

Schmidt et al. (2007) describe an alternative, raster-based method for deriving FFG, which we will refer to hereinafter as GFFG. The GFFG derivation method was developed at the ABRFC and has subsequently been implemented for operational use. The basic design of GFFG follows that of FFG in that it first requires

offline Thresh-R calculations and then computes rainfall thresholds at 1-, 3- and 6-h durations that will result in bankfull conditions given initial soil moisture conditions. To compute Thresh-R, GFFG uses a design 3-h rainfall event over the ABRFC having a 5-yr return period. The linear response of flow associated with the 5-yr rainstorm (i.e., the threshold flow that results in flooding) is computed using the Natural Resources Conservation Service (NRCS) triangular unit hydrograph method. The unit hydrograph peak is also estimated from the NRCS method. The NRCS method differs from Snyder's method in that it takes into account basin characteristics such as slope and NRCS curve number (CN) as opposed to relying on regionally optimized coefficients. The NRCS CN is an estimate of a basin's potential maximum retention, which is derived from land-use categories and four hydrologic soil groups. Once the unit hydrograph peak is estimated from basin characteristics, Thresh-R values that correspond to the design 5-yr rainstorm are computed. Resulting values of Thresh-R used in GFFG reflect the spatial variability present in the basin characteristic maps and digital elevation model-derived slope. In general, basin-averaged GFFG Thresh-R values are similar to those used in FFG, but they are lower in areas of high topographic relief.

The second component of GFFG accounts for initial soil moisture conditions and solves for rainfall thresholds at 1-, 3-, and 6-h durations that correspond to the aforementioned Thresh-R values. In this case, the NRCS CN [Eq. (1)] is used to compute the GFFG values in each grid cell rather than running a lumped hydrologic model under varying rainfall scenarios:

$$R = \frac{(P - 0.2S)^2}{(P + 0.8S)}, \quad (1)$$

where R is accumulated direct runoff (mm), P is the accumulated rainfall (mm), and S is the potential maximum retention (mm). Here, R is the Thresh-R value, P is GFFG, and S comes from the following CN equation:

$$S = \frac{1000}{\text{CN}} - 10, \quad (2)$$

where CN comes from a lookup table that is based on land use and hydrologic soil groups. The traditional NRCS CN method accounts for soil moisture conditions by adjusting the a priori CN on the basis of previous rainfall in the past 5 days and the time of year (to account for losses due to evapotranspiration). In the GFFG method, however, CN values are adjusted using two model states from NWS Hydrology Laboratory Research Distributed Hydrologic Model simulations (Koren et al. 2004). To be specific,

spatially distributed grids of the upper-zone free and tension water contents (UZFWC and UZTWC) are summed and normalized by their potential maximum values (UZFWM and UZTWM). These latter values are model parameters and are estimated a priori on the basis of the physical relationships described in Koren et al. (2000). The model soil moisture states thus represent the degree of saturation of the upper soil layers. This soil saturation value is then used to adjust CN in Eq. (2) in lieu of using antecedent precipitation in the traditional NRCS CN method. Then P is solved for in Eq. (1) since the values for R (Thresh-R) and for the soil-saturation-adjusted S are known. This is GFFG, and values are nominally computed for 1-, 3-, and 6-h rainfall accumulations.

3. Study domain and datasets

Heavy-rainfall events in the ABRFC domain occur primarily during the warm season as a result of landfalling tropical storms and localized intense convective storms and from mesoscale convective complexes (MCCs) traversing the region through the nighttime hours (Arndt et al. 2009; Maddox 1980). In fact, Kincer (1916) found that as much as 60% of total rainfall during the warm season occurs at night, and Maddox (1980) later identified the culprit of the nocturnal rainfall maximum as being organized MCCs. Average annual precipitation amounts vary considerably across the ABRFC, with minimum values of 219 mm occurring in the west. Rainfall amounts increase gradually to the east, where an annual maximum of 1797 mm occurs. The highest elevation of 4225 m occurs in the westernmost boundary of the ABRFC where snowmelt from the Rocky Mountains feeds the headwaters of the Arkansas River (see Fig. 1). The lowest elevation of 60 m occurs in the eastern part of the ABRFC where the Arkansas drains to the Mississippi River.

This study focuses on operational flash-flood prediction datasets produced by the ABRFC from 1 September 2006 to 22 August 2008 (Fig. 1). Figure 2 shows the average rainfall and upper-99.9th percentile of hourly QPE for the study period. The 99.9th-percentile plot is based on the top 17th hour of hourly QPE in the ranked histograms computed at each grid point. This analysis is meant to show where extreme rainfall occurred during the study period. The 99.9th percentile is shown rather than the maximum QPE because the latter product was found to be corrupted by radar-based artifacts that may have only been present in a single hourly accumulation. Figure 2b shows relatively high amounts, approximately 25 mm, in western Oklahoma. These amounts were associated with a strengthening Tropical Storm Erin, which produced record rainfall amounts and subsequent flash flooding as it traversed the state from 17 to 20 August 2007. The

FIG. 1. Study domain of the ABRFC, outlined in red. Flash-flood reports from the NWS are shown as light-blue polygons. Instrumented basins with catchments of $<250 \text{ km}^2$ are outlined in dark blue. The NWS reports and discharge measurements are used to evaluate operational flash-flood guidance thresholds from 1 Sep 2006 to 22 Aug 2008.

Oklahoma Climate Survey reported June 2007 as the wettest month on record since 1895 for four out of nine climate divisions in Oklahoma. Throughout the summer of 2007, there were 15 days of damaging flash floods, which makes this study period particularly useful for evaluating flash-flood prediction tools.

The study time period and area were selected on the basis of the availability of all three of the following datasets: 1) hourly QPEs from the NWS stage-IV product, 2) FFG produced in hindcast mode daily at 1800 UTC for 1-, 3-, and 6-h accumulation periods, and 3) GFFG produced operationally at 1200, 1800, 0000, and, on occasion, 0600 UTC for the same accumulation periods as in FFG. QPE and GFFG grids were obtained from the archive maintained by the NWS National Precipitation Verification Unit. Hindcast FFG data were generated by the ABRFC because the GFFG method has replaced FFG for use in operations. FFG values were resampled on the 4-km-resolution HRAP grid used for both the GFFG

and QPE products. The precipitation forcing is from the hourly multisensor stage-IV product, which combines rainfall estimates from WSR-88Ds and rain gauges with quality control performed manually by NWS forecasters. The technique of merging radar and rain gauge rainfall amounts has its roots in the P1 method originally developed at the ABRFC and is now implemented operationally at other RFCs. Additional details of the algorithm can be found online (<http://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage4/>).

Prior to conducting a quantitative comparison and evaluation of FFG and GFFG as flash-flood prediction tools, issues about operational product generation time and duration over which they are valid needed to be taken into consideration. For instance, following suggestions by ABRFC forecasters, the valid times of GFFG and FFG products were lagged by the duration of their accumulation period. A 6-h FFG product issued at 1800 UTC was not considered to be valid until 0000 UTC. The reason for

FIG. 2. (a) Average QPE from the hourly stage-IV rainfall product for the study period and (b) upper-99.9th-percentile quantile of QPE.

the introduction of the lag is that the FFG (and GFFG) values are conditioned on the rainfall and antecedent soil moisture conditions leading up to the product issuance time. It is entirely possible that 3- and 6-h rainfall accumulations within 2–5 h immediately following the FFG issuance (e.g., at 1900 UTC) will contain accumulated rainfall from the time prior to the FFG issuance (e.g., 1300–1900 UTC). This could result in a situation in which it is not currently raining at 1900 UTC yet prior accumulated rainfall could have exceeded recently issued FFG and thus falsely and artificially alert on a flash flood. A second consideration was made regarding the frequency mismatch over which QPE, FFG, and GFFG products were generated. Hourly QPE grids were summed to yield 3- and 6-h accumulations at the top of each hour. FFG and GFFG products, however, were produced only once and 3–4 times per day, respectively. An interpolation scheme was developed for FFG and GFFG products to create hourly, interim products. Hourly values were computed from the most recent FFG (GFFG) values and the subsequently issued products. As opposed to linearly

interpolating the FFG and GFFG values between the issuance times, we used a weighting scheme that was based on hourly rainfall accumulations as follows:

$$\text{FFG}_t = \text{FFG}_{i+1}w_t + \text{FFG}_i(1 - w_t), \quad (3)$$

where

$$w_t = \frac{\sum_{j=1}^t \text{QPE}_j}{\sum_{j=1}^{i+1} \text{QPE}_j} \quad (4)$$

and the subscript t refers to the present hour, i and $i + 1$ correspond to the prior and next FFG (or GFFG) issuance times, and j in Eq. (4) iterates through each hour between i and $i + 1$. Note that all indices apply to the time-lagged FFG and GFFG values. This had the effect of dropping the FFG and GFFG values closer to their next-issued values in response to concurrent rainfall amounts. The length of time over which values are interpolated is greater with FFG than with GFFG. This introduces additional uncertainty in FFG and needs to be taken into consideration when directly comparing the skill of the different methods. Reproducing FFG and GFFG values in hind-cast mode at the same frequency, even hourly, would yield equitable datasets, but this capability does not presently exist.

4. Results

a. Intercomparison of FFG and GFFG values

Prior to evaluating the skill of FFG and GFFG values as flash-flood prediction tools, we first compare the spatial distribution of their time-averaged values throughout the study period and the histograms of their hourly values. Temporal averages of FFG and GFFG for 1-h accumulation duration were computed throughout the entire study period, as shown in Fig. 3. In qualitative terms, the spatial patterns of FFG and GFFG show similarities with minimum values in eastern Colorado and New Mexico and maximum values in northwestern Oklahoma. The most obvious difference between the two products is the appearance of basin-averaged values in FFG, which result from running the lumped-parameter SAC-SMA model on basins with catchments on the order of 1000 km². The GFFG values, on the other hand, have greater spatial variability owing primarily to their derivation from NRCS CNs on 4-km grid cells. The ABRFC GFFG method assigns the hydrologic soils in group A (and associated CNs) in northwestern Oklahoma relatively high values of GFFG.

FIG. 3. Average values of (a) FFG and (b) GFFG corresponding to a 1-h accumulation period. (c) Histograms of FFG (black line) and GFFG (blue line) values for all times and grid points in study domain.

Relative frequency histograms of GFFG and FFG values were computed by considering each hour and each grid point in the domain (Fig. 3c). Despite the differences in spatial resolution, as well as differences in the methods used to compute their values, the frequency histograms for 1-h accumulation from the two methods match very well. The only noteworthy feature is the unnatural relative maximum in the 85–90-mm bin in the FFG histogram. This artifact is due to a cap or threshold imposed on FFG values. The average spatial patterns for the 3-h accumulation period are very similar to those for 1 h for both FFG and GFFG, but the values for both methods are higher than their 1-h respective accumulations, as would be expected (Figs. 4a,b). The frequency histograms, however, indicate that GFFG values are higher than FFG (Fig. 4c). The mode of the 3-h GFFG distribution falls in the 65–70-mm bin whereas FFG's mode is in the 60–65-mm bin, indicating a quantitative difference of 5 mm, or 8%. The significance of this difference is assessed in the quantitative evaluation in sections 4c and 4d. The same spatial patterns shown at 1- and 3-h accumulations for FFG and GFFG are reproduced for the 6-h accumulation period (Figs. 5a,b). A comparison of the relative frequency histograms in Fig. 5c reveals that GFFG is once again higher than FFG by

approximately 6%. We also see the impact of a threshold maximum amount applied to FFG values, with an unnatural relative maximum occurring in the 150–155-mm bin. In summary, GFFG values have more spatial variability than FFG for all accumulation periods. GFFG values are higher than those for FFG by approximately 8% for 3-h rainfall accumulations and by 6% for 6-h accumulations.

b. NWS Storm Data reports

The NWS maintains an archive of severe-weather events, including flash floods, that occur throughout the United States. Collection of reports composing the database follows NWS Instruction 10-1605: “*Storm Data preparation*” (NWS 2007). Information stored in the database includes estimated beginning and ending time of event, direct and indirect injuries and fatalities, property and crop damage in U.S. dollars, cause of flooding (e.g., heavy rain or dam break), source of information (e.g., trained spotter, emergency management, or law enforcement), an event narrative, and location of the event. Prior to October of 2007, locations were nominally recorded by impacted county. County areas vary considerably across the United States and more specificity of flash-flooding impacts was deemed desirable by the NWS

FIG. 4. As in Fig. 3, but for 3-h accumulation period.

Performance Management Team. As a result, the locations of the events are now stored as polygons with as many as eight vertices (defined by latitude and longitude). Figure 1 shows the locations of the total 579 reports of flash flooding by the NWS in the ABRFC from 1 September 2006 to 22 August 2008.

The method to objectively evaluate FFG and GFFG using NWS *Storm Data* reports first involved computing every instance in which QPE exceeded FFG and GFFG at each grid point for the study period. In NWS operations, forecasters are alerted to the potential onset of a flash-flooding event when this exceedance occurs. We also realized that the utility of FFG and GFFG could potentially be maximized by considering QPE-to-guidance ratios other than 1 being reached or exceeded. For example, a given FFG value may prove to be a more skillful predictor when rainfall exceeds it by 150% rather than 100%. As such, we considered QPE-to-guidance ratios of 0.5, 0.75, 1.0, 1.25, 1.5, 2.0, 2.5, and 3.0.

A binary approach (i.e., yes/no) was taken to evaluate FFG (GFFG) using the NWS reports of flash flooding. Contingency tables were populated for each of the aforementioned QPE-to-guidance ratios at 1-, 3-, and 6-h durations, totaling 48 tables (see example in Table 1). A forecast flash flood was considered as an instance in

which the QPE-to-guidance ratio was exceeded for each accumulation period and for each ratio. To determine whether the forecast event was associated with an observed event (i.e., was a hit in Table 1), the NWS *Storm Data* event database was searched for a reported flash flood collocated in space and time. Search parameters were defined to account for spatial and temporal offsets between the occurrence of heavy rainfall and observed impacts from flash flooding as well as potential inaccuracies in the timing and location of the NWS reports. Because the NWS *Storm Data* reports contained both county locations and specific polygons of flash-flood impacts, we chose to compute the centroid of each report and then to define the spatial search domain as a 7×7 box centered on each report. The temporal search domain was defined to cover QPE-to-guidance ratio exceedance up to 8 h in advance of the reported flash flooding and 2 h following the report. Justification for the temporal search parameters was a nominal 6-h period between causative rainfall and flash flooding plus 2 h on either side to account for uncertainty in the timing of the report. The searching procedure reported the maximum QPE-to-guidance ratio associated with observed flash-flooding events. A two-pass scheme through the datasets was required so as to first find the hits and misses

FIG. 5. As in Fig. 3, but for 6-h accumulation period.

associated with observed reports and then report the false alarms associated with forecast but unobserved events. Correct negatives were computed but are not reported hereinafter because of the predominance of occurrences falling into this category (i.e., areas not receiving any rain and no reported flash flooding).

After the contingency tables were populated, three statistics were computed from the hits, misses, and false alarms in each of the contingency tables. The probability of detection (POD) describes the fraction of observed flash floods that were correctly forecast:

$$\text{POD} = \frac{\text{hits}}{\text{hits} + \text{misses}}. \quad (5)$$

A POD of 1 indicates that all flash floods were correctly forecast; 0 indicates no flash floods were detected by the forecast tools. The POD must be considered along with a false-alarm ratio (FAR), which describes the fraction of forecast events that were not associated with observed events:

$$\text{FAR} = \frac{\text{false alarms}}{\text{hits} + \text{false alarms}}. \quad (6)$$

Similar to POD, FAR ranges from 0, indicating no forecast events went unobserved, to 1, indicating all forecast

flash floods were not associated with an observed event. A third statistic, the aforementioned CSI or threat score, was computed:

$$\text{CSI} = \frac{\text{hits}}{\text{hits} + \text{misses} + \text{false alarms}}. \quad (7)$$

CSI combines both aspects of POD and FAR and thus describes the skill of a forecast system. CSI ranges from 0, indicating no skill, to 1, for perfect skill.

Table 2 reports the contingency-table statistics for FFG and GFFG for all three accumulation periods by considering QPE-to-guidance ratios of 1.0, that is, what is nominally used to alert NWS forecasters on imminent flash flooding. Hereinafter, we refer to FFG and GFFG for an exceedance ratio of 1.0 as the uncalibrated performance. The POD ranges from 0.41 with 1-h GFFG to

TABLE 1. Contingency table used to evaluate binary (yes/no) flash-flood forecasts.

		Obs	
		Yes	No
Forecast	Yes	Hit	False alarm
	No	Miss	Correct negative

TABLE 2. Contingency table statistics for exceedance ratio of 1 using NWS *Storm Data* reports.

		POD	FAR	CSI
FFG	1 h	0.45	0.84	0.13
	3 h	0.66	0.95	0.05
	6 h	0.61	0.97	0.03
GFFG	1 h	0.41	0.86	0.12
	3 h	0.65	0.94	0.06
	6 h	0.52	0.96	0.04

0.66 for 3-h FFG, which indicates that approximately one-half of the events were detected with the uncalibrated tools. The values for FAR, however, range from 0.84 to 0.97, indicating overforecasting, incomplete reporting of flash-flood events in *Storm Data*, or both. Inadequacies of individual flash-flood observation databases are discussed in Gruntfest (2009) and Gourley et al. (2010), with the latter study providing an assessment of the population representativeness of NWS reports. It is undoubted that flash-flood events occurring in sparsely populated regions in the ABRFC, especially the western portion, will not be reported and thus will lead to an artificially inflated FAR. The NWS reports can thus be used to assess the relative performance of the flash-flood warning tools rather than their absolute skill. Because of the high FAR, the CSI values for all uncalibrated tools are very low, especially in comparison with the 0.22 benchmark reported in Norbiato et al. (2008). The highest CSI values of 0.13 and 0.12 are associated with the 1-h accumulation period for uncalibrated FFG and GFFG, respectively, and CSI values worsen with longer accumulation periods because of higher false alarms. Nonetheless, the relative differences between uncalibrated FFG and GFFG are slight and indistinguishable.

Because there are numerous assumptions in the derivation of FFG and GFFG that are prone to uncertainties, we computed the CSI for FFG and GFFG for various QPE-to-guidance exceedance ratios (see Fig. 6). First, we see that the 1-, 3-, and 6-h FFG skill curves have higher CSI values at all exceedance ratios relative to GFFG. Results from both FFG and GFFG indicate that skill decreases with increasing accumulation period. Moreover, Figs. 4 and 5 indicate that FFG values were lower than GFFG for 3- and 6-h accumulations. Figure 6 shows how the skill of FFG for these longer accumulation periods can be optimized by considering rainfall exceeding FFG by a factor of 2 rather than 1. In essence, the apparent negative (low) bias present in the FFG values can be accommodated by merely considering higher rainfall exceedances when using the NWS *Storm Data* reports. However, the analysis using NWS *Storm Data* reports is useful for relative FFG and GFFG comparisons, whereas the next

FIG. 6. The skill of flash-flood forecasting tools for a variety of exceedance ratios. Observations are from reports of flash flooding in the NWS *Storm Data* database over the study region from 1 Sep 2006 to 22 Aug 2008.

section is complementary and provides more insights to their skill in an absolute sense.

c. USGS stream-discharge measurements

The major advantage of using USGS discharge data to evaluate FFG and GFFG is the continuous nature of their measurements. As such, there will be very few unobserved events that artificially inflate FAR, which was the case in the prior analysis. On the other hand, USGS stations do not explicitly indicate the occurrence of flash flooding as a binary event. Instead, we must set a threshold to approximate flows that likely exceed bankfull conditions. Another downfall of the USGS dataset is the lack of instrumented basins, especially for small catchments. With NWS *Storm Data* reports, we assumed that a flash-flooding event could be possible anywhere within the ABRFC; the limiting factor was population density. In this case, we can only consider forecasts of flash flooding that occurred within USGS-gauged basins. The sample size is thus decreased by this restriction. Within the ABRFC, we found 19 stream gauges with contributing basin areas of $<260 \text{ km}^2$, that is, those that collect data at the flash-flood scale. It is noted that the definition of basin scale susceptible to flash flooding is ambiguous, but it approximately corresponds to a basin concentration time of 6 h, which is a time scale used to divide responsibilities between RFCs and local forecast offices at the NWS (Reed et al. 2007).

Table 3 shows the characteristics of the USGS dataset employed in this study. Stations with contributing basin areas of $<260 \text{ km}^2$ and with a minimum of 10 yr of data record were considered. This period of record was required to proceed with a flood frequency analysis. The

TABLE 3. Description of USGS discharge data used in the study. Two-year return flows were considered as observed flash floods. Their values were computed using a log Pearson type-III distribution applied to the annual maximum series; i.e., the Water Resources Council method discussed in Chow et al. (1988).

USGS station No.	Basin area (km ²)	Period of record	2-yr return flow (m ³ s ⁻¹)
7165565	14.1	1988–2008	35.1
7328180	19.0	1990–2008	3.8
7177800	19.5	1989–2008	65.7
7189540	20.5	1992–2008	19.4
7177650	21.2	1988–2008	45.8
7164600	31.6	1990–2008	177.0
7165562	46.1	1988–2008	86.3
7170700	95.8	1958–2008	41.5
7335700	103.9	1966–2008	260.9
7196900	105.2	1958–2008	204.3
7329852	114.2	1990–2008	111.4
7185095	115.8	1984–93, 2005–08	83.4
7325850	154.4	1970–78, 2005–08	28.7
7195855	154.9	1997–2008	49.6
7327447	160.3	1993–2008	26.0
7160350	182.1	1997–2008	98.4
7299890	191.7	1998–2008	8.6
7247250	192.7	1993–2008	319.7
7340300	232.1	1961, 1968–2008	410.0
7197360	232.1	1998–2008	121.5

method employed was the Water Resources Council method discussed in Chow et al. (1988), which uses a logarithmic Pearson type-III distribution applied to the annual maximum series. Annual exceedance probabilities were computed and converted to return flows. We defined an observed flash flood to be an instance in which the 2-yr

return-period flow was exceeded at each station shown in Table 3. The selection of this return-period flow is meant to approximate bankfull conditions, as discussed in Carpenter et al. (1999). The 2-yr return-period flows are reported in Table 3. Figure 7 shows five instances at USGS station 7165565 in which the streamflow exceeded the basin flash-flood threshold. For the study period, overall there were 72 USGS-observed flash floods, that is, instances in which the 2-yr return-period flow was exceeded.

The evaluation of FFG and GFFG was carried out in a manner that is similar to that in the previous section. That is, we populated contingency tables for 1-, 3-, and 6-h FFG and GFFG for a variety of exceedance ratios. For each observed flash flood, we searched for the QPE-to-guidance ratio being reached or exceeded 8 h prior and 2 h following the event to account for 6 h of basin concentration time plus 2 h of uncertainty in the specific timing of the event. We searched for the maximum exceedance ratio throughout this time window for all grid points contained in the basin. Recall that in the previous analysis a spatial search domain was set to a 7×7 box centered on each report to account for spatial uncertainty in the NWS reports. In this case, we only searched for grid points that were completely contained in the basins; those were the only ones that could have contributed to basin runoff.

Contingency-table statistics for uncalibrated FFG and GFFG are shown in Tables 4 and 5. The values in Table 4 correspond to instances in which a single grid point in the basin exceeded the QPE-to-guidance ratio of 1. In Table 5,

FIG. 7. Observed discharge time series for USGS station 7165565 for the study period. The 2-yr return-period flow is shown as a dashed line.

TABLE 4. As in Table 2, but flash-flood observations are from exceedance of 2-yr return-period flows at USGS stations listed in Table 3 and a forecast of flash flooding is considered when the basin-maximum QPE-to-guidance ratio exceeds 1.

		POD	FAR	CSI
FFG	1 h	0.24	0.67	0.16
	3 h	0.69	0.79	0.19
	6 h	0.67	0.85	0.14
GFFG	1 h	0.25	0.60	0.18
	3 h	0.55	0.76	0.20
	6 h	0.41	0.81	0.15

values are shown for instances in which the *basin-averaged* QPE-to-guidance ratio exceeded 1. When considering the basin maximum, POD values are low at 0.24 for FFG and 0.25 for GFFG at the 1-h accumulation period whereas POD values for the longer accumulation periods are similar to those found using the NWS *Storm Data* reports. The FAR values range from 0.60 to 0.85, which is much lower than those shown in Table 2. This result indicates that a portion of the high false alarms in the prior analysis was due to inadequacies in the observational database. The skill scores with the uncalibrated tools when considering basin-maximum exceedance range from 0.14 with 6-h FFG to 0.20 with 3-h GFFG. The latter score is comparable to the 0.22 CSI benchmark reported in Norbiato et al. (2008). Also, the relative differences between uncalibrated FFG and GFFG are insignificant but show that GFFG has slightly higher CSI scores. We see better results when evaluating the basin-mean QPE-to-guidance ratio of 1 being exceeded (see Table 5). In comparison with the basin-maximum statistics in Table 4, the POD is reduced with the basin-mean exceedance values, but the FAR is reduced even more. This results in relatively higher CSI values ranging from 0.08 with 1-h GFFG to 0.34 with 3-h FFG. The best skill with the uncalibrated FFG and GFFG values, when considering the basin-mean exceedance, occurs with the 3-h accumulation periods. Also, FFG outperforms GFFG for all accumulation periods.

The absolute skill of the forecast tools for a variety of QPE-to-guidance ratios is shown in Fig. 8. In considering basin-maximum ratios in Fig. 8a, we see that higher CSI values occur with FFG relative to GFFG for the corresponding accumulation periods. The highest CSI is with 3-h FFG at 0.28 (when considering a basin-maximum ratio of 1.5 being exceeded), and GFFG skill generally worsens with increasing accumulation period. The shapes of the curves in Fig. 8b are similar to those in Figs. 8a and 6 but are shifted to the left and have higher CSI values. The highest level of skill occurs with 1- and 3-h FFG values (CSI = 0.34) for basin-mean QPE-to-guidance ratios of 0.5 and 1.0, respectively. Despite the analyses

TABLE 5. As in Table 2, but flash flood observations are from exceedance of 2-yr return-period flows at USGS stations listed in Table 3 and a forecast of flash flooding is considered when the basin-mean QPE-to-guidance ratio exceeds 1.

		POD	FAR	CSI
FFG	1 h	0.09	0.25	0.09
	3 h	0.49	0.47	0.34
	6 h	0.43	0.67	0.23
GFFG	1 h	0.08	0.33	0.08
	3 h	0.31	0.58	0.22
	6 h	0.23	0.73	0.14

shown in Figs. 6 and 8 being based on independent observations of flash flooding, they share striking similarities. The analysis reveals the impact of lower FFG values relative to GFFG at the 3- and 6-h accumulation periods and also shows that FFG performs better than the corresponding GFFG. These same features were found using the NWS *Storm Data* dataset, thus lending confidence when interpreting the results. Overall, the tools used to alert forecasters to the onset of flash flooding have optimum skill when considering a basin-mean QPE-to-guidance ratio of 0.5 being reached or exceeded for a 1-h accumulation period and a ratio of 1.0 being reached or exceeded for 3 h. A case study is examined in the next section to assess the potential advantages in spatial resolution offered by GFFG.

d. Severe Hazards Analysis and Verification Experiment witness reports

The third dataset used to evaluate FFG and GFFG comes from data collected during SHAVE (Ortega et al. 2009). Gourley et al. (2010) compared SHAVE reports with NWS *Storm Data* reports and found them to be representative of the same population density and proximity of sampled residents to streams. The SHAVE flash-flood observations differ from those reported by the NWS because the former are point specific, dense, and storm targeted, whereas the latter are specified by bounding polygons, and the NWS *Storm Data* database is meant to encompass all flash-flooding events throughout the United States. The storm-targeted nature of SHAVE meant that we could not assess false alarms in computing FAR and CSI. Because the data were collected for specific events, we chose to compare FFG with GFFG on a case-study basis rather than for a continuous time period as was done previously.

Figures 9a and 9b show 3-h FFG and GFFG at 2100 UTC 17 June 2008 for an urban flash-flood case near Oklahoma City, Oklahoma. The resolution differences in the two products are readily apparent, with a broader spectrum of values present in GFFG. The QPE-to-FFG ratio product in Fig. 9c shows that the spatial scale of basin-averaged FFG

FIG. 8. As in Fig. 6, but flash-flood observations are from exceedance of 2-yr return-period flows at USGS stations listed in Table 3 for (a) basin-maximum QPE/guidance and (b) basin-mean QPE/guidance.

is coarser than the SHAVE flash-flood observations. QPE-to-GFFG ratios, on the other hand, show two separate regions with values >2.5 that have nearby SHAVE-indicated severe flash flooding (Fig. 9d). There is an indication that details present in the higher-spatial-resolution GFFG, which is a defining characteristic of the product, are capable of identifying specific locations of observed flash flooding. These differences were not elucidated in the prior analyses that only considered maximum and mean exceedances within their respective search domains. Although the NRCS CN method incorporates information about land use and thus the degree of urbanization, the resulting mean GFFG values shown in Figs. 3b, 4b, and 5b show little to no sensitivity to urbanization. Thus, neither FFG nor GFFG adequately account for higher runoff ratios expected in urban settings. Because this is only a single flash-flood case, the qualitative improvements in GFFG over FFG due to higher spatial resolution must be considered anecdotally rather than on the basis of a complete analysis.

5. Summary and conclusions

This study completed a multitiered evaluation of the skill of tools used to operationally monitor and predict flash floods in the NWS. Using values from the legacy, basinwide flash-flood guidance method (FFG) and newer, gridded flash-flood guidance (GFFG), originally developed at the Arkansas–Red Basin River Forecast Center and now used operationally at the ABRFC and the Lower Mississippi, Southeast, and West Gulf RFCs, comparisons with quantitative precipitation estimates from the stage-IV product were made over the ABRFC from 1 September 2006 to 22 August 2008. A flash-flood forecast was defined when QPE-to-guidance ratios were reached or exceeded and then evaluated using observations of flash flooding from NWS *Storm Data* flash-flooding reports, exceedance of 2-yr return period flows in USGS-gauged basins with drainage areas of less than 260 km², and high-density witness reports collected during the Severe Hazards Analysis and Verification Experiment. The quantitative metrics reported herein will serve a valuable purpose by establishing benchmarks so that future flash-flood prediction tools can be developed and evaluated based on their relative skill improvements.

The following summarizes the primary findings from this study:

- When examining QPE-to-guidance exceedance at individual grid points using flash-flood observations from NWS *Storm Data* reports and USGS stream gauges, the FFG and GFFG methods were only marginally skillful, with the highest CSI of 0.20 occurring with 3-h GFFG.
- The same USGS and NWS flash-flood observations, however, both indicated that FFG performed better than GFFG when a *variety* of QPE-to-guidance exceedance ratios were considered.
- The best skill, as indicated with a CSI of 0.34, occurred from stream gauge verification with 1- and 3-h FFG. This score was achieved only when basin-mean rainfall reached 50% of FFG for 1-h accumulation and exceeded it for 3 h.
- Spatial averaging of QPE and thus QPE-to-guidance exceedance ratios on basins reduced the number of false alarms sufficiently to improve CSI. This spatial averaging also reduced the threshold exceedance ratios required to optimize CSI.
- A subjective comparison using high-resolution flash-flood observations from SHAVE indicated GFFG was capable of detecting the spatial variability of reported flash flooding better than FFG for an urban case study.

Local experience and knowledge from NWS forecasters will yield improvements to the skill of the tools

FIG. 9. Values of (a) 3-h FFG, (b) 3-h GFFG, (c) QPE-to-FFG ratio, and (d) QPE-to-GFFG ratio at 2100 UTC 17 Jun 2008 for an urban flash-flood case near Oklahoma City. Witness reports collected during SHAVE were classified into null, nonsevere, and severe classes and are shown as color-filled circles, as indicated in the legend.

when issuing flash-flood warnings. This study considered QPE-to-guidance exceedance ratios other than 1.0 and found that FFG can be very simply optimized if 1) a basin-mean rainfall is considered and 2) the basin-mean QPE reaches 50% of FFG for a 1-h accumulation period, exceeds FFG for a 3-h period, or exceeds FFG by 125% for a 6-h accumulation period. Users of GFFG should also consider basin-mean QPE and will optimize its skill for situations in which basin-mean rainfall approaches 50% of GFFG for 1 h or 75% of GFFG for both 3- and 6-h accumulation periods. NWS forecasters can use this information immediately when considering issuance of flash-flood warnings. Although the skill of averaged GFFG products was slightly lower than that from FFG using basin-integrated USGS data and NWS *Storm Data* polygon reports, there was an indication that GFFG was able to identify potential threat areas at scales that are

smaller than the resolution of FFG and the first two evaluation sources. Thus, the combined use of FFG for basinwide monitoring and GFFG for specific impacts is a rational approach.

Limitations of the results reported herein include the representativeness of the 2-yr period of coincident FFG, GFFG, and QPE observations and specificity to the ABRFC domain. It is unlikely that these results apply directly to other regions such as the semiarid Intermountain West where routing plays a major role or to northern-tier regions where frozen surface conditions impact surface runoff generation. Uncertainty in the stage-IV QPE products was not considered in this study but is believed to have a negligible impact on the FFG and GFFG evaluations. FFG values were produced retrospectively at daily frequency by the ABRFC, whereas GFFG values were generated operationally 3–4 times per day. A temporal

interpolation scheme based on rainfall occurrence was developed to account for the differences at which the products were updated. However, it is possible that the skill of more frequently generated, operational FFG values will be different than what is reported here. Nonetheless, these results can be used as the first benchmarks established for FFG and GFFG used for operational flash-flood monitoring and prediction in the NWS. Future work will continue to develop and improve flash-flood observation databases and then will expand the FFG and GFFG evaluation to other regions over the conterminous United States. Other future directions involve the development and evaluation of distributed modeling approaches to flash-flood prediction, such as the threshold frequency method in Reed et al. (2007).

Acknowledgments. Funding was provided by the NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA17RJ1227, the U.S. Department of Commerce, the National Severe Storms Laboratory's Director's Discretionary Research Funds, and the National Weather Service's Advanced Hydrologic Prediction System funds. Hindcast values of daily flash-flood guidance were produced and provided by James Paul (ABRFC). Useful discussions with John Schmidt (SERFC) and Seann Reed (OHD) helped in the synthesis of this study. Their assistance is greatly appreciated. We also appreciate the NWS National Precipitation Verification Unit for making the QPE and GFFG grids available to us.

REFERENCES

- Anderson, E. A., 1973: National Weather Service River Forecast System—Snow accumulation and ablation model. NOAA Tech. Memo. NWS HYDRO-17, 217 pp.
- Arndt, D. S., J. B. Basara, R. A. McPherson, B. G. Illston, G. D. McManus, and D. B. Demko, 2009: Observations of the overland reintensification of Tropical Storm Erin (2007). *Bull. Amer. Meteor. Soc.*, **90**, 1079–1093.
- Ashley, S. T., and W. S. Ashley, 2008: Flood fatalities in the United States. *J. Appl. Meteor. Climatol.*, **47**, 806–818.
- Burnash, R. J. C., 1995: *Computer Models of Watershed Hydrology*. Water Resources Publications, 1144 pp.
- Carpenter, T. M., J. A. Sperkslage, K. P. Georgakakos, T. Sweeney, and D. L. Fread, 1999: National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *J. Hydrol.*, **224**, 21–44.
- Chow, V. T., D. R. Maidment, and L. W. Mays, 1988: *Applied Hydrology*. McGraw-Hill, 572 pp.
- Davis, R. S., 2003: Some practical applications of flash flood monitoring and prediction. Preprints, *17th Conf. on Hydrology*, Long Beach, CA, Amer. Meteor. Soc., J4.7. [Available online at <http://ams.confex.com/ams/pdfpapers/55695.pdf>.]
- , 2004: Locally modifying flash flood guidance to improve the detection capability of the Flash Flood Monitoring and Prediction Program. Preprints, *18th Conf. on Hydrology*, Seattle, WA, Amer. Meteor. Soc., J1.2. [Available online at <http://ams.confex.com/ams/pdfpapers/68922.pdf>.]
- Georgakakos, K. P., 1986: A generalized stochastic hydrometeorological model for flood and flash-flood forecasting. 1. Formulation. *Water Resour. Res.*, **22**, 2083–2095.
- , 2006: Analytical results for operational flash flood guidance. *J. Hydrol.*, **317**, 81–103.
- Gourley, J. J., J. M. Erlingis, T. M. Smith, K. L. Ortega, and Y. Hong, 2010: Remote collection and analysis of witness reports on flash floods. *J. Hydrol.*, **394**, 53–62, doi:10.1016/j.jhydrol.2010.05.042.
- Gruntfest, E., 2009: Editorial. *J. Flood Risk Manage.*, **2**, 83–84.
- Hong, Y., P. Adhikari, and J. J. Gourley, 2011: Flash flood. *Encyclopedia of Natural Hazards*, P. Bobrowsky, Ed., Springer, in press.
- Kincer, J. B., 1916: Daytime and nighttime precipitation and their economic significance. *Mon. Wea. Rev.*, **44**, 628–633.
- Koren, V. I., M. Smith, D. Wang, and Z. Zhang, 2000: Use of soil property data in the derivation of conceptual rainfall–runoff model parameters. Preprints, *15th Conf. on Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 103–106.
- , S. Reed, M. Smith, Z. Zhang, and D.-J. Seo, 2004: Hydrology Laboratory Research Modeling System (HL-RMS) of the U.S. National Weather Service. *J. Hydrol.*, **291**, 297–318.
- Krajewski, W. F., G. Villarini, and J. A. Smith, 2010: Radar rainfall uncertainties. *Bull. Amer. Meteor. Soc.*, **91**, 87–94.
- Maddox, R. A., 1980: Mesoscale convective complexes. *Bull. Amer. Meteor. Soc.*, **61**, 1374–1387.
- Mogil, H. M., J. C. Monro, and H. S. Groper, 1978: NWS's flash flood warning and disaster preparedness programs. *Bull. Amer. Meteor. Soc.*, **59**, 690–699.
- Moore, R. J., 1985: The probability-distributed principle and runoff production at point and basin scales. *Hydrol. Sci. J.*, **30**, 273–297.
- Norbiato, D., M. Borga, S. Degli Esposti, E. Gaume, and S. Anquetin, 2008: Flash flood warning based on rainfall depth–duration thresholds and soil moisture conditions: An assessment for gauged and ungauged basin. *J. Hydrol.*, **362**, 274–290.
- , —, and R. Dinale, 2009: Flash flood warning in ungauged basins by use of the flash flood guidance and model-based runoff thresholds. *Meteor. Appl.*, **16**, 65–75.
- Ntelekos, A. A., W. F. Krajewski, and K. P. Georgakakos, 2006: On the uncertainties of flash flood guidance: Toward probabilistic forecasting of flash floods. *J. Hydrometeorol.*, **7**, 896–915.
- NWS, 2007: *Storm Data* preparation. National Weather Service Instruction 10-1605, 97 pp. [Available online at <http://www.nws.noaa.gov/directives/sym/pd01016005curr.pdf>.]
- Ortega, K. L., T. M. Smith, K. L. Manross, K. A. Scharfenberg, A. Witt, A. G. Kolodziej, and J. J. Gourley, 2009: The Severe Hazards Analysis and Verification Experiment. *Bull. Amer. Meteor. Soc.*, **90**, 1519–1530.
- Reed, S. M., and D. R. Maidment, 1999: Coordinate transformations for using NEXRAD data in GIS-based hydrologic modeling. *J. Hydrol. Eng.*, **4**, 174–182.
- , D. Johnson, and T. Sweeney, 2002: Application and national geographic information system database to support two-year flood and threshold runoff estimates. *J. Hydrol. Eng.*, **7**, 209–219.

- , 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.
- Schmidt, J. A., A. J. Anderson, and J. H. Paul, 2007: Spatially-variable, physically-derived flash flood guidance. Preprints, *21st Conf. on Hydrology*, San Antonio, TX, Amer. Meteor. Soc., 6B.2. [Available online at <http://ams.confex.com/ams/pdfpapers/120022.pdf>.]
- Smith, G., 2003: Flash flood potential: Determining the hydrologic response of FFMP basins to heavy rain by analyzing their physiographic characteristics. Report to the NWS Colorado Basin River Forecast Center, 11 pp. [Available online at http://www.cbrfc.noaa.gov/papers/ffp_wpap.pdf.]
- Sweeney, T. L., 1992: Modernized areal flash flood guidance. NOAA Tech. Rep. NWS HYDRO 44, Hydrologic Research Laboratory, National Weather Service, NOAA, Silver Spring, MD, 21 pp. and an appendix.