The 23 June 2016 West Virginia Flash Flood Event as Observed through Two Hydrometeorology Testbed Experiments

STEVEN M. MARTINAITIS,[a,b] BENJAMIN ALBRIGHT,[c,d] JONATHAN J. GOURLEY,[b] SARAH PERFATER,[d,e] TIFFANY MEYER,[a,b] ZACHARY L. FLAMIG,[a,b] ROBERT A. CLARK,[a,b] HUMBERTO VERGARA,[a,b] AND MARK KLEIN[d]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma;* [b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma;* [c] *Systems Research Group, Inc., College Park, Maryland;* [d] *NOAA/NWS/Weather Prediction Center, College Park, Maryland;* [e] *I.M. Systems Group, Inc., College Park, Maryland*

(Manuscript received 27 January 2020, in final form 21 April 2020)

ABSTRACT: The flash flood event of 23 June 2016 devastated portions of West Virginia and west-central Virginia, resulting in 23 fatalities and 5 new record river crests. The flash flooding was part of a multiday event that was classified as a billion-dollar disaster. The 23 June 2016 event occurred during real-time operations by two Hydrometeorology Testbed (HMT) experiments. The Flash Flood and Intense Rainfall (FFaIR) experiment focused on the 6–24-h forecast through the utilization of experimental high-resolution deterministic and ensemble numerical weather prediction and hydrologic model guidance. The HMT Multi-Radar Multi-Sensor Hydro (HMT-Hydro) experiment concentrated on the 0–6-h time frame for the prediction and warning of flash floods primarily through the experimental Flooded Locations and Simulated Hydrographs product suite. This study describes the various model guidance, applications, and evaluations from both testbed experiments during the 23 June 2016 flash flood event. Various model outputs provided a significant precipitation signal that increased the confidence of FFaIR experiment participants to issue a high risk for flash flooding for the region between 1800 UTC 23 June and 0000 UTC 24 June. Experimental flash flood warnings issued during the HMT-Hydro experiment for this event improved the probability of detection and resulted in a 63.8% increase in lead time to 84.2 min. Isolated flash floods in Kentucky demonstrated the potential to reduce the warned area. Participants characterized how different model guidance and analysis products influenced the decision-making process and how the experimental products can help shape future national and local flash flood operations.

SIGNIFICANCE STATEMENT: Testbed environments allow for researchers to evaluate new products and techniques through structured application and feedback from end-users. Two Hydrometeorology Testbed experiments were operating during the historic flash flood event in the West Virginia region on 23 June 2016. This study investigates how experiment participants applied experimental numerical weather prediction forecasts and hydrologic model guidance in the generation of forecast products and flash flood warnings. Findings from model and product evaluations characterized the various strengths and challenges with predicting a record rainfall event while assessing participant perceptions on how new products influenced flash flood operations. Findings from this event can help shape future operational needs and the movement toward probabilistic information in the forecast and warning process.

KEYWORDS: Hydrometeorology; Numerical weather prediction/forecasting; Flood events

1. Introduction

One of the more significant weather events in the United States during the 2016 calendar year occurred on 23 June across an area from northern Kentucky to central Virginia. The greatest impacts were in West Virginia and western Virginia from flash flooding, a flood caused by excessive rainfall that leads to a rapid rise in water within a 6-h period. Record rainfall accumulations of 200–250 mm were observed in this region over a 24-h period ending 1200 UTC 24 June 2016. The event resulted in 23 fatalities along with damage or destruction of thousands of structures and over 1500 roads and bridges. The estimated cost of damage from flooding and other associated severe weather phenomena was $1.0 billion (U.S. dollars), which the National Center for Environmental Information classified as 1 of 15 events designated as a billion-dollar disasters in the United States during 2016 and 1 of 5 events mostly attributed to flooding (https://www.ncdc.noaa.gov/billions/events/US/2016).

Flash flood prediction has improved over the years; however, challenges still limit the ability to accurately forecast and detect flash floods. The skill of numerical weather prediction (NWP) model quantitative precipitation forecasts (QPFs) over the warm season (June–August) has seen some performance increases while skill over the cold season (December–February) has seen marked improvements (Fritsch and Carbone 2004; Ebert et al. 2007; Barthold et al. 2015). Variations in seasonal performance were attributed to spatiotemporal differences in precipitation events, the scale of the forcing mechanism, and the skill of NWP models to accurately depict synoptically driven events versus

Corresponding author: Steven M. Martinaitis, steven.martinaitis@noaa.gov

DOI: 10.1175/WAF-D-20-0016.1

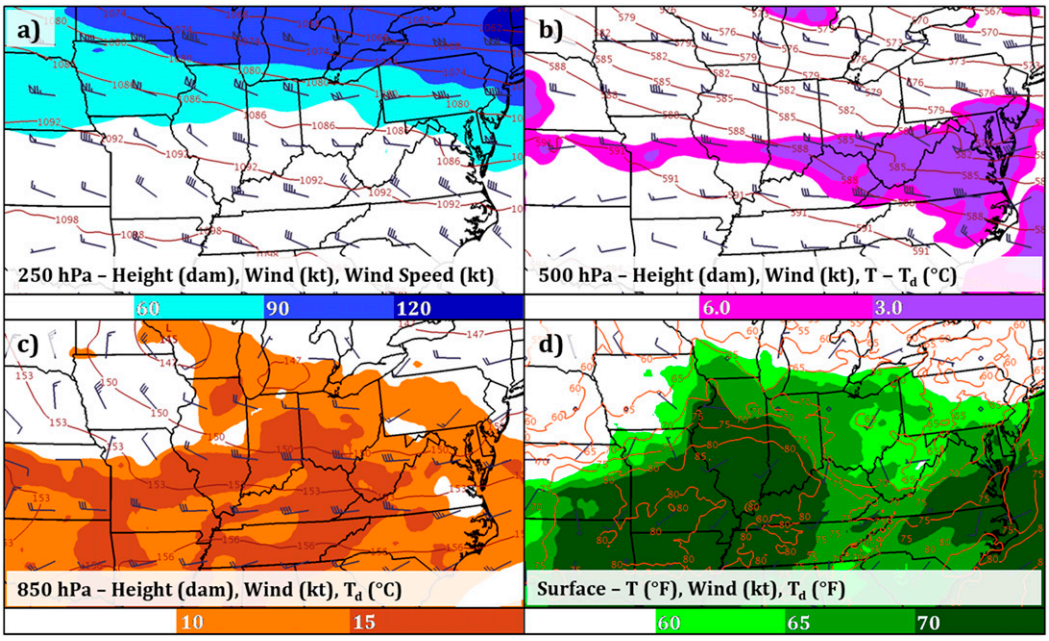

FIG. 1. Regional analysis from the Rapid Refresh model at 1200 UTC 23 Jun 2016 for the following environmental parameters: (a) 250-hPa height (contours; dam), wind (barbs; kt; 1 kt $\approx$ 0.51 m s^{-1}), and wind speed magnitude (shaded; kt); (b) 500-hPa height (contours; dam), wind (barbs; kt), and dewpoint depression (shaded; °C); (c) 850-hPa height (contours; dam), wind (barbs; kt), and dewpoint temperature (shaded; °C); and (d) surface temperature (contours; °F), wind (barbs; kt), and dewpoint temperature (shaded; °F).

localized convective events (Sukovich et al. 2014; Barthold et al. 2015).

The short-term detection and warning of flash flooding on the 0–6-h time scale has also seen negligible improvements in performance metrics from 2008 to 2014 (Martinaitis et al. 2017). National Weather Service (NWS) forecasters have typically relied on a flash flood guidance (FFG) product generated at NWS River Forecast Centers to estimate the amount of rainfall needed over various temporal periods to generate bank-full conditions on small waterways (Sweeney 1992). Regional critical success index scores of FFG over the conterminous United States (CONUS) varied from 0.00 to 0.19 when compared to NOAA StormData reports (https://www.ncdc.noaa.gov/stormevents/) and 0.00–0.44 when compared to U.S. Geological Survey stream gauge observations (Clark et al. 2014).

Experimental products and modeling techniques were developed over recent years to advance flash flood prediction. NWP models have increased in spatial resolution and aimed to improve model accuracy of QPF placement and

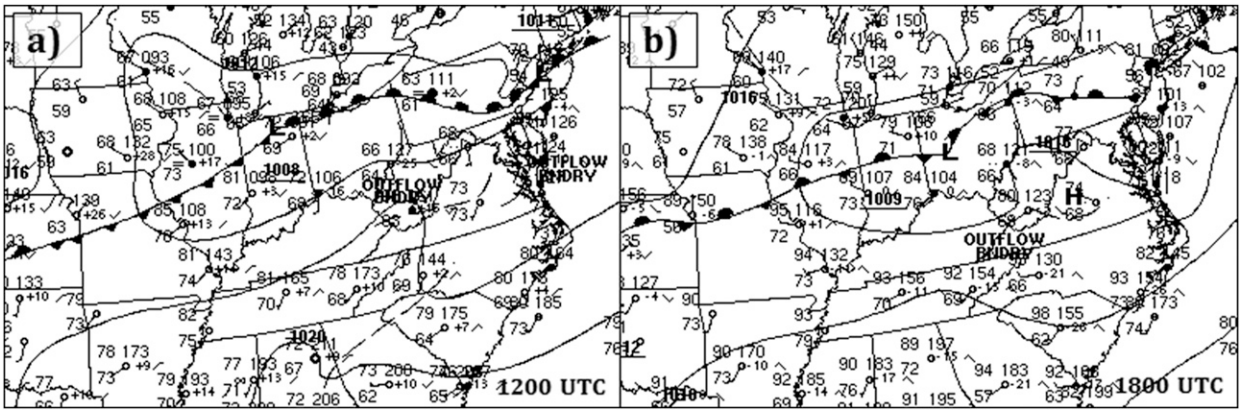

FIG. 2. Surface analysis from the NOAA/NWS Weather Prediction Center (WPC; http://www.wpc.ncep.noaa.gov/) at (a) 1200 and (b) 1800 UTC 23 Jun 2016. Depicted on each map are frontal and outflow boundaries, high and low pressure areas, and station observations.

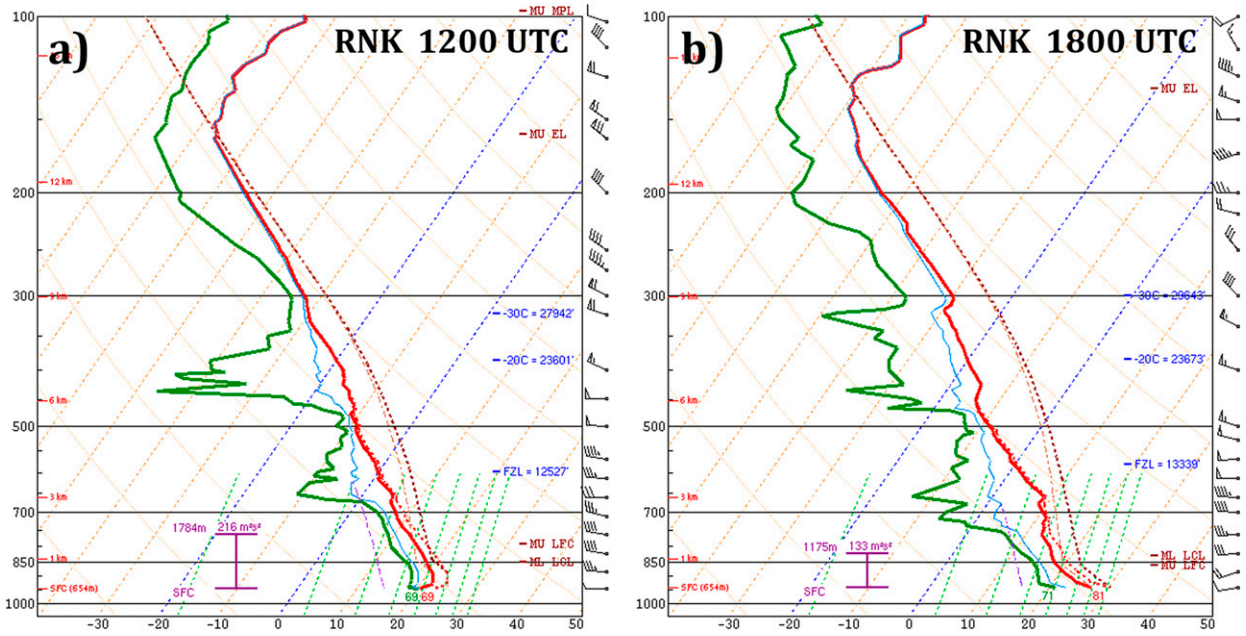

FIG. 3. Skew *T*–log*p* plot of the observed Roanoke, Virginia (RNK), upper-air sounding for (a) 1200 and (b) 1800 UTC 23 Jun 2016 from the NOAA/NWS Storm Prediction Center (SPC; https://www.spc.noaa.gov/).

magnitude, while new convection-allowing ensemble model approaches have been developed to improve overall precipitation forecasting. Ensemble models have employed techniques such as using a probability matched mean QPF to identify high magnitude QPFs that a regular ensemble mean may not accurately capture (e.g., Snook et al. 2019); moreover, new hydrologic modeling platforms (e.g., Viterbo et al. 2020) have also been in development to incorporate quantitative

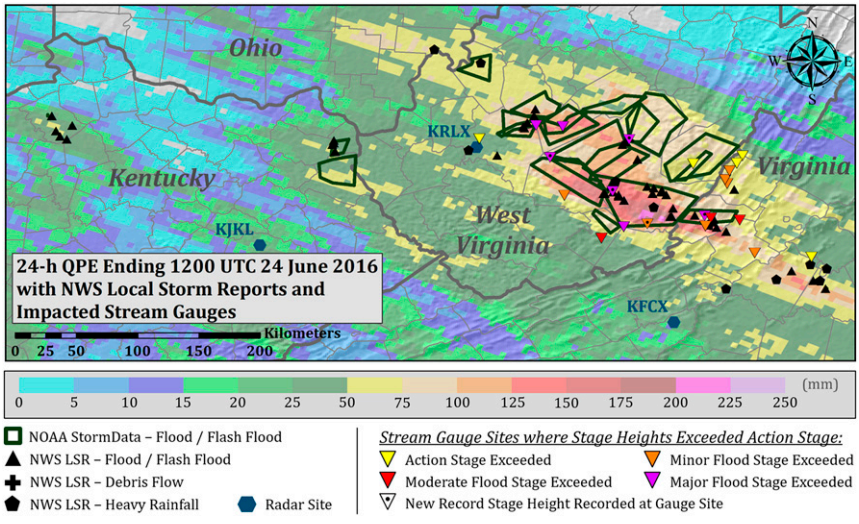

FIG. 4. The 24-h QPE from the NCEP Stage IV analysis (Lin and Mitchell 2005) ending 1200 UTC 24 Jun 2016 overlaid with radar locations and NWS flood-related reports along with stream gauge sites where stage heights exceeded action stage. NWS flood-related local storm reports (LSRs) that were recorded during the event are represented by solid black symbols. NWS flood and flash flood reports from NOAA StormData (https://www.ncdc.noaa.gov/stormevents/) are represented by green polygons. Stream gauge sites exceeding the action flood stage (i.e., the level where some mitigation action is taken in preparation for flooding) are represented by inverted triangle, color coded based on the highest level flood stage category being exceeded. Gauge sites with a dot in the symbol represent a record stage height crest observed at that location.

TABLE 1. Observed and normal precipitation totals along with percent of normal precipitation observed for the month of May 2016 and for the period 1–22 Jun 2016 for select NWS climate locations in Virginia and West Virginia.

Location	1–31 May 2016			1–22 Jun 2016		
	Observed (mm)	Normal (mm)	Percent observed	Observed (mm)	Normal (mm)	Percent observed
Blacksburg, VA	115.8	110.0	105%	41.4	72.6	57%
Lynchburg, VA	175.0	94.7	185%	44.5	67.1	66%
Roanoke, VA	155.4	103.1	151%	85.6	72.6	118%
Beckley, WV	167.4	118.4	141%	113.3	72.6	156%
Charleston, WV	129.3	121.9	106%	65.0	78.2	83%
Huntington, WV	133.9	120.8	111%	113.0	72.6	156%

precipitation estimates (QPEs) and NWP QPFs to predict flooding hazards.

The research-to-operations paradigm outlined by Nietfeld (2013) described a means for demonstrating and validating new or enhanced applications, methods, products, and services through testing and analysis for end-users. One research-to-operations platform to evaluate flash flood forecasting advancements was the annual Flash Flood and Intense Rainfall (FFaIR) experiment (Barthold et al. 2015) hosted at the Weather Prediction Center (WPC) and conducted under the Hydrometeorology Testbed (HMT) banner. The 2016 FFaIR experiment brought together participants from across the weather enterprise in a simulated pseudo-operational environment to create experimental probabilistic forecasts and evaluate emerging models, tools, and datasets. Participants utilized a combination of experimental NWP models as well as new hydrologic products to produce forecasts highlighting the probability of rainfall exceeding FFG. Short-term forecasting skill was also evaluated through the daily generation of 6-h probabilistic forecasts to define the probability of flash flooding for a limited domain.

Operating in conjunction with the FFaIR experiment was the annual HMT Multi-Radar Multi-Sensor Hydro (hereafter denoted as HMT-Hydro) experiment, which was also conducted under the HMT banner and held at the National Weather Center (Martinaitis et al. 2017). The 2016 HMT-Hydro experiment provided a real-time environment to evaluate new tools and techniques designed to assist in the 0–6-h time frame for the prediction and warning of flash floods. Participants utilized experimental products in a real-time operational

TABLE 2. List of stream gauges that exceeded action flood stage (i.e., the level where some mitigation action is taken in preparation for flooding) from the 23 Jun 2016 event. The sites presented have defined flood stage levels, and locations denoted by an asterisk only have flood stages defined up to minor flooding level. The gauges listed were from the U.S. Geological Survey (USGS), the Integrated Flood Observing and Warning System (IFLOWS), and the West Virginia Division of Homeland Security and Emergency Management (WV DHSEM). Also listed in the table are the peak crest observed (m) observed at each site, the flood stage level exceeded, and the ranking of the peak crest from the event.

Gauge ID	Source	Location	Event peak crest (m)	Flood stage	Event crest rank
02011400	USGS	Jackson River near Bacova, VA	3.18	Minor	21
02011460	USGS	Back Creek near Sunrise, VA	1.92	Action	20
02011470	USGS	Back Creek at Sunrise, VA	2.42	Action	22
02011500	USGS	Back Creek near Mountain Grove, VA	2.53	Minor	25
02013000	USGS	Dunlap Creek near Covington, VA	5.08	Major	1
02013100	USGS	Jackson River at Covington, VA	6.81	Moderate	3
02014000	USGS	Potts Creek near Covington, VA	3.00	Minor	9
02016500	USGS	James River at Lick Run, VA	6.96	Moderate	9
02019500	USGS	James River at Buchanan, VA	6.20	Minor	24
02025500	USGS	James River at Holcomb Rock, VA	5.90	Action	54
03182500	USGS	Greenbrier River at Buckeye, WV	4.40	Action	27
03183500	USGS	Greenbrier River at Alderson, WV	6.71	Major	3
03184000	USGS	Greenbrier River at Hilldale, WV	7.76	Moderate	3
03185400	USGS	New River at Thurmond, WV	5.97	Minor	5
03187000	USGS	Gauley River at Camden-on-Gauley, WV	9.07	Major	1
03192500	USGS	Gauley River at Belva, WV	9.24	Major	1
03196800	USGS	Elk River at Clay, WV	9.24	Major	3
03197000	USGS	Elk River at Queen Shoals, WV	10.15	Major	2
03198000	USGS	Kanawha River at South Side Bridge, WV	8.81	Action	56
CSBV2*	IFLOWS	Back Creek at Cassidy, VA	2.29	Minor	2
HNEW2	IFLOWS	Meadow River at Hines, WV	5.73	Major	1
RONW2*	WV DHSEM	Greenbrier River at Ronceverte, WV	7.07	Minor	1

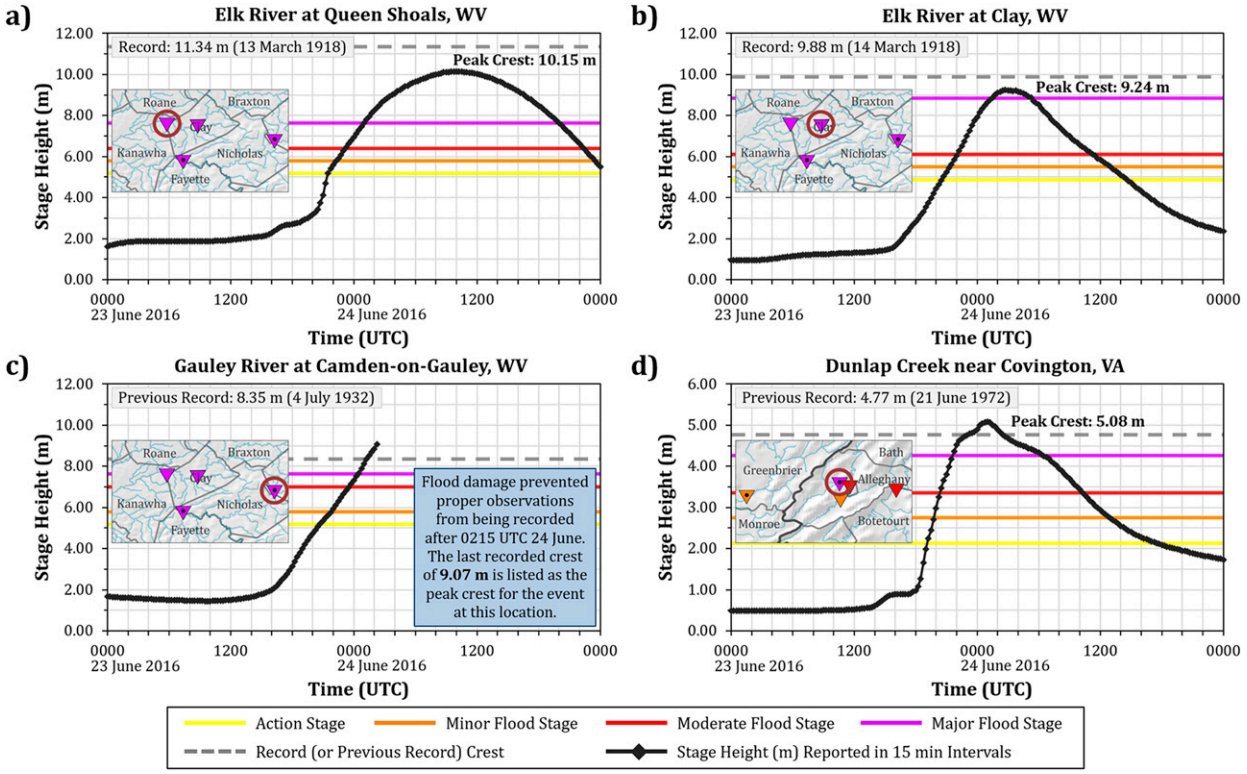

FIG. 5. Time series of stage height (m) with listed peak crest from 0000 UTC 23 Jun to 0000 UTC 25 Jun 2016 for (a) Elk River at Queen Shoals, WV (USGS ID 03197000); (b) Elk River at Clay, WV (USGS ID 03196800); (c) Gauley River at Camden-on-Gauley, WV (USGS ID 03187000); and (d) Dunlap Creek near Covington, VA (USGS ID 02013000). Also plotted are the flood stage categories (solid lines) and the record or previous record crest (dashed gray line).

environment to issue experimental flash flood warnings. Daily subjective evaluations focused on the ability of products to capture the spatial coverage and magnitude of flash flood events and on the issuance of experimental warnings. Other feedback mechanisms assessed how various products influenced the warning decision-making process.

This work describes the FFaIR and HMT-Hydro experiments and the activities surrounding the 23 June 2016 flash flood event. Numerous experimental hydrometeorological applications were accessible in real time to facilitate decision-making processes in creating short-term forecasts. Experimental product outputs were compared to operational products and local storm reports, and objective evaluations assessed the statistical skill of warning flash flood event. Experimental model guidance was subjectively evaluated to measure the perception of operational utility. The inclusion of subjective feedback allowed for an efficient mechanism to engage participant expertise in the development and training of products for operational transition.

2. Overview of 23 June 2016 flash flood event

Multiple precipitation events traversed the area extending from northern Kentucky and southern Ohio to central Virginia in response to a progressive shortwave trough and deep moisture advection. Rapid Refresh model (Benjamin et al. 2016) analysis initialized at 1200 UTC 23 June placed West Virginia within the right-entrance region of a zonally oriented 67 m s^{-1} (130 kt) jet streak at 250 hPa located over New England (Fig. 1a). Midlevel moisture originating from the eastern Pacific Ocean was abundant with <3.0°C dewpoint depressions at 500 hPa and dewpoint temperatures > 10°C at 850 hPa (Figs. 1b,c). Surface dewpoint temperatures in the region were ≥15.5°C (60°F) and originated from the Gulf of Mexico via advection around the western and northern periphery of a high pressure center over the Florida panhandle (Fig. 1d). Areas with ≥21.1°C (70°F) dewpoint temperatures were within proximity of the ongoing convection at 1200 UTC.

WPC surface analysis depicted a low pressure center over northern Indiana with a zonally oriented warm front extended into Pennsylvania at 1200 UTC (Fig. 2a). Two outflow boundaries were analyzed near the area of interest and acted as forcing mechanisms for the additional rounds of convection that occurred between 1200 and 2300 UTC. The outflow boundaries were related to the initial mesoscale convective system that traversed the region prior to 1200 UTC. This pattern persisted through 1800 UTC with a mesohigh identified over Virginia, which likely enhanced the southerly flow and subsequent forcing into any remaining outflow boundaries (Fig. 2b) analogous to the

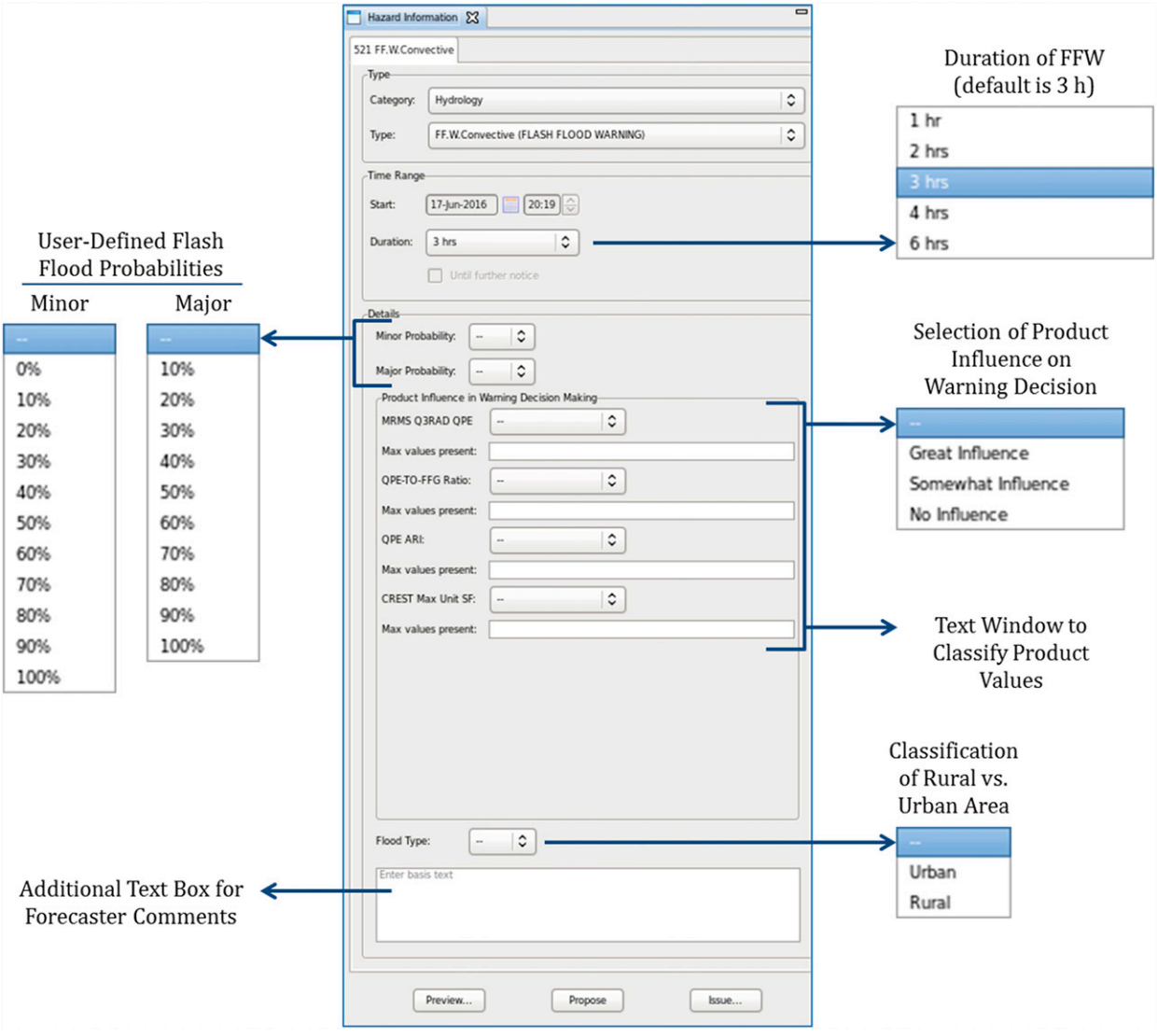

FIG. 6. Modified hazard information graphical user interface within the Hazard Services software for the issuance of FFWs. This interface was utilized by participants during experimental flash flood operations to survey the use of experimental products in their warning decision-making process.

mesohigh pattern for heavy rainfall described by Maddox et al. (1979).

The Roanoke, Virginia, observed soundings at both 1200 and 1800 UTC 23 June depicted the available moisture content throughout the vertical levels of the atmosphere (Fig. 3). The 1200 UTC sounding observed a record high surface dewpoint temperature of 20.6°C (69.1°F) and a precipitable water value of 38.4 mm that exceeded the 90th percentile value (37.3 mm) for that location and time when compared to the Storm Prediction Center sounding climatology (https://www.spc.noaa.gov/exper/soundingclimo/). The 1800 UTC sounding recorded a dewpoint temperature of 21.4°C (70.5°F) along with a precipitable water value of 33.3 mm. It is noted that the Roanoke, Virginia, sounding later recorded a precipitable water value of 45.0 mm at 0000 UTC 24 June 2016 (not shown). This meteorological setup coupled with record high moisture content resulted in widespread accumulated precipitation values of 50–150 mm with localized regions of 200–250 mm over the 24-h period ending 1200 UTC 24 June (Fig. 4).

Long-term antecedent conditions portrayed some above-normal precipitation in the region from 1 May to 22 June 2016 when compared to climatology (Table 1). This included a relatively dry period from 6 to 19 June preceding the studied event. The initial precipitation that traversed the region from 0700 to 1200 UTC 23 June increased the soil moisture content and flash flood potential. The subsequent rounds of precipitation combined with saturated grounds and hilly terrain resulting in flash flooding that occurred between 1650 UTC 23 June and 0100 UTC 24 June, which transitioned into areal flooding through 2200 UTC 24 June. Local storm reports

FFaIR Experiment	22 June 2016	HMT-Hydro Experiment
Experimental Day 1 model analysis	1200 UTC	
Day 1 ERO creation (valid 1500 UTC 22 June to 1200 UTC 23 June); Forecast briefing preparation		
Subjective model evaluation (valid 1200 UTC 21 June to 1200 UTC 22 June)	1400 UTC	
Lunch		
Day 1 PFFF creation (valid 1800 UTC 22 June to 0000 UTC 23 June); Finish forecast briefing	1600 UTC	Lunch; Evaluate FFaIR products (valid 21 June)
FFaIR daily forecast briefing	1800 UTC	FFaIR daily forecast briefing
Day 2 ERO creation (valid 1200 UTC 23 June to 1200 UTC 24 June)		Subjective product evaluation (Harlan Co., KY – 21 June)
	2000 UTC	Real-time experimental warning operations
	2200 UTC	
	0000 UTC	

FFaIR Experiment	23 June 2016	HMT-Hydro Experiment
Experimental Day 1 model analysis	1200 UTC	
Day 1 ERO creation (valid 1500 UTC 23 June to 1200 UTC 24 June); Forecast briefing preparation		
Subjective model evaluation (valid 1200 UTC 22 June to 1200 UTC 23 June)	1400 UTC	
Lunch		
Day 1 PFFF creation (valid 1800 UTC 23 June to 0000 UTC 24 June); Finish forecast briefing	1600 UTC	Lunch; Evaluate FFaIR products (valid 22 June)
FFaIR daily forecast briefing	1800 UTC	FFaIR daily forecast briefing
Day 2 ERO creation (valid 1200 UTC 24 June to 1200 UTC 25 June)		
	2000 UTC	Real-time experimental warning operations
	2200 UTC	
	0000 UTC	Subjective product evaluation (Kanawha Co., WV – 23 June)

FFaIR Experiment	24 June 2016	HMT-Hydro Experiment
Experimental Day 1 model analysis	1200 UTC	
Day 1 ERO creation (valid 1500 UTC 24 June to 1200 UTC 25 June); Forecast briefing preparation		
Subjective model evaluation (valid 1200 UTC 23 June to 1200 UTC 24 June)	1400 UTC	
Lunch	1600 UTC	Subjective product evaluation (Boyd Co., KY – 23 June)
Day 1 PFFF creation (valid 1800 UTC 23 June to 0000 UTC 24 June); Finish forecast briefing		"Tales from the Testbed" webinar preparation
FFaIR daily forecast briefing	1800 UTC	Lunch; Best practices discussion
Day 2 ERO creation (valid 1200 UTC 25 June to 1200 UTC 26 June)		"Tales from the Testbed" weekly webinar
	2000 UTC	Feedback survey; Group photo

Experiment Activity | Experiment Activity Related to 23 June 2016 Event | No Activities

FIG. 7. Timeline containing all activities from 22 to 24 Jun 2016 for both the FFaIR and HMT-Hydro experiments. Experiment activities related to the 23 Jun 2016 flash flood event are highlighted in green.

TABLE 3. Featured operational deterministic and ensemble NWP model guidance available during the 2016 FFaIR Experiment. Listed are the various models utilized, the model provider, the horizontal grid spacing, the forecast hour period, and additional notes describing the model.

Model	Provider	Grid spacing	Forecast period	Notes
NAM	NCEP	12 km (parent), 4 km (nested)	60 h (nested domain) and 84 h (parent domain); available at 0000, 0600, 1200, and 1800 UTC	Operational NAM with 12-km parent model with 4-km CONUS nested domain
HRRR	NCEP	3 km	15 h	High-resolution, hourly updated convective-allowing nest of Rapid Refresh (RAP) model
HRRRv2	NCEP	3 km	18 h	NCO parallel of HRRR; hourly updated and convective allowing
NMMB	EMC/NSSL	4 km	48 h	High-resolution, convective-allowing CONUS model
ARW	EMC/NSSL	4 km	48 h	High-resolution, convective-allowing CONUS model
WRF-NSSL	EMC/NSSL	4 km	36 h	High-resolution, convective-allowing CONUS model

(LSRs) of flooding and flash flooding were documented in 19 counties from northern Kentucky to central Virginia per NOAA StormData (Fig. 4). The most significant impacts occurred in Greenbrier and Kanawha Counties in West Virginia, where 22 of the 23 fatalities occurred along with flood-related damage initially estimated at $102 million. Numerous stream gauge sites exceeded minor flood stage, and seven stream gauge sites exceeded major flood stage (Table 2). Five gauge sites surpassed previous record crests. Stream gauge sites located in the regions of greatest impacts observed sustained stage increases $>0.60\,\mathrm{m\,h^{-1}}$ for periods of 4–9 h (e.g., Fig. 5). The stream gauge site along the Elk River at Queen Shoals, West Virginia, recorded a maximum hourly stage increase of 1.76 m between 2030 and 2130 UTC 23 June (Fig. 5a).

3. Daily operations and methodologies

a. FFaIR experiment framework

FFaIR experiment activities focused on the development of two probabilistic forecast products: the excessive rainfall outlook (ERO) and probability of flash flood forecast (PFFF). The ERO was fashioned after the WPC operational ERO and defined as the probability of precipitation exceeding FFG within 40 km of a point. Contours were drawn at probability values of 2% (marginal risk), 5% (slight risk), 10% (moderate risk), and 30% (high risk). Experiment participants utilized the combination of experimental NWP guidance, probabilistic ensemble output, and hydrologic model products. Two experimental EROs were issued each day. The experimental Day 1 ERO was issued around 1415 UTC and was valid from 1500 to 1200 UTC the following day. The experimental Day 2 ERO was issued around 2000 UTC and was valid for the 24-h period starting at 1200 UTC the following day. Short-term forecasting skill was evaluated through the development of an experimental 6-h PFFF. The PFFF was defined as the probability of flash flooding occurring within 40 km of a point using 10% (slight risk), 30% (moderate risk), and 50% (high risk) probability contours. The PFFF product was issued around 1745 UTC and was valid between 1800 and 0000 UTC over a limited domain area defined by a greater potential for flash flooding and/or forecasting challenges if multiple threat areas existed.

The day following each Day 1 ERO and PFFF issuance contained a formal subjective evaluation session to measure the perceptive skill of the various models, ensembles, tools, and experimental forecasts. NWP and hydrologic model evaluations focused on the spatial coverage and magnitude of the various model outputs. Model QPFs were compared against the Multi-Radar Multi-Sensor (MRMS) radar-only QPE (Zhang et al. 2016). The Day 1 and Day 2 EROs valid for the previous day along with the associated 1800–0000 UTC PFFF were assessed based on the distribution of flood and flash flood reports, operational flash flood warnings (FFWs), and stream gauge observations compared to the contoured threat areas. Participants assigned a score ranging from 1 (very poor) to 10 (very good) for each model, tool, and experimental forecast evaluated.

TABLE 4. As in Table 3, but for the experimental deterministic and ensemble NWP model guidance.

Model	Provider	Grid spacing	Forecast period	Notes
NAMRR	EMC	12 km (parent), 3 km (nested)	18 h (hourly forecast); 60 h (nested domain) and 84 h (parent domain); available at 0000, 0600, 1200, and 1800 UTC	Features hourly forecast and assimilation cycle for 12-km North American parent domain and 3 km for nested CONUS domain; uses hybrid 3DEnVar and incorporates radar reflectivity into assimilation system via complex cloud analysis approach
ESRL HRRRv2	ESRL/GSD	3 km	18 h (hourly forecast); 36 h (hourly forecast available every 3 h)	Experimental version of HRRR; hourly updating; convective allowing; refined boundary layer and land surface schemes using WRF-ARW v3.7.1 to reduce warm/dry biases
HRRR-TLE	ESRL/GSD	3 km	24 h	Neighbor ensemble approach calculated over 3-km time-lagged HRRRv2 deterministic members; probabilities at a point refer to chance of exceeding a given threshold somewhere with 40-km radius of a given point with 80-km data spatial filter to increase ensemble spread
WPC-SSEO	SPC/ESRL/WPC	4 km	24 h	Multimodel, multiphysics convection-allowing model modified from original SSEO; removes SPCWRF and incorporates HRRR; seven members; 40-km neighborhood probabilities
WRF-ARW SSEFX	OU/CAPS	3 km	60 h	15-member (14 ARW + 1 NMMB) ensemble forecast starting at 0000 UTC
HREFX	EMC	5 km	36 h	Experimental version of HREF with eight members; produces ensemble mean precipitation in three different forms and probabilistic guidance through neighborhood probabilities and Gaussian smoothing
GEFSX	EMC	111 km	384 h	Beta version of operational GEFS using new frequency match calibration technique; 20 ensemble members and one control member (operational GEFS)

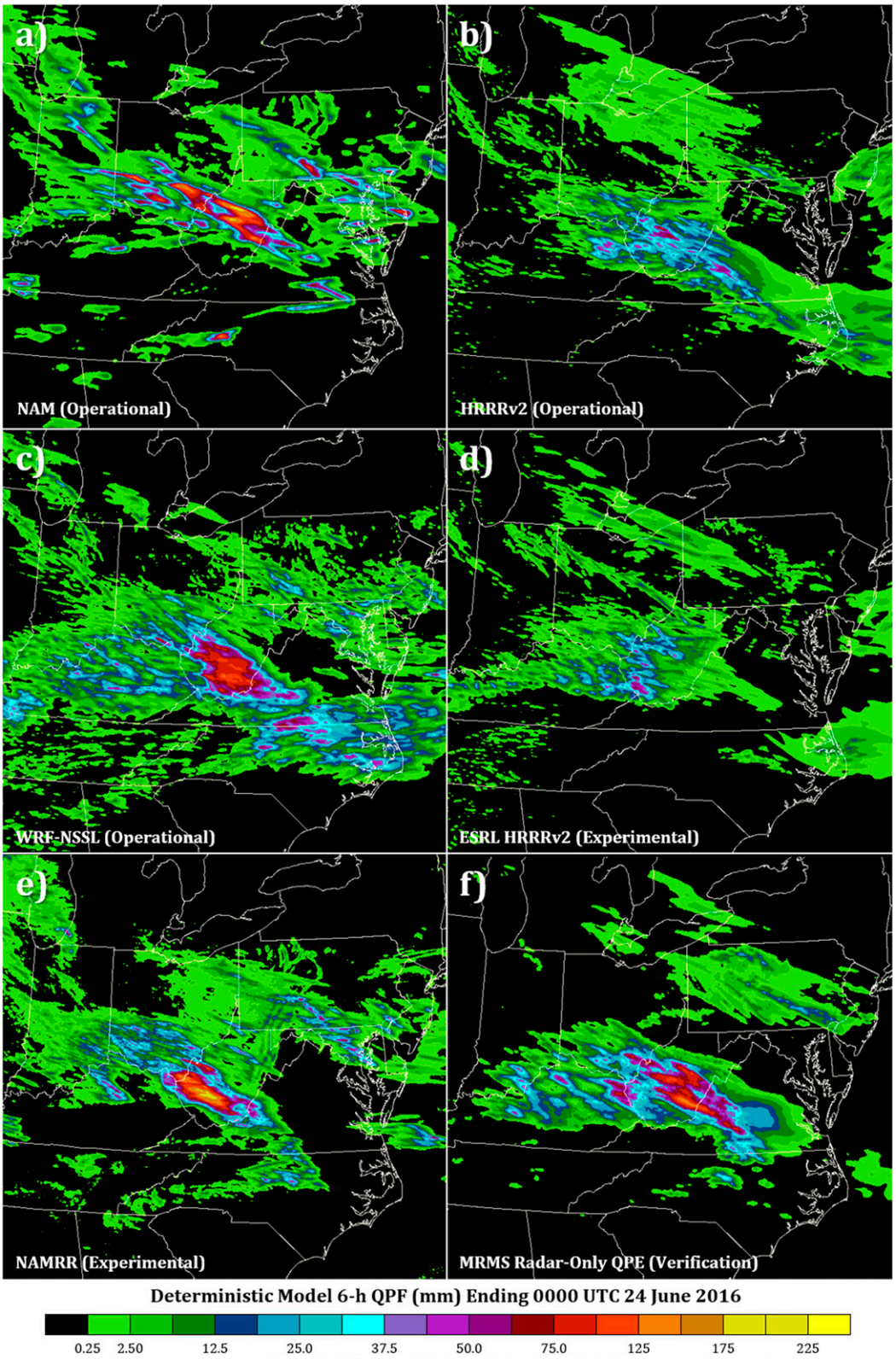

FIG. 8. Deterministic 6-h QPF (mm) ending 0000 UTC 24 Jun 2016 for the operational (a) NAM, (b) HRRRv2, and (c) WRF-NSSL models and the experimental (d) ESRL HRRRv2 and (e) NAMRR models compared to the (f) MRMS radar-only QPE used for verification.

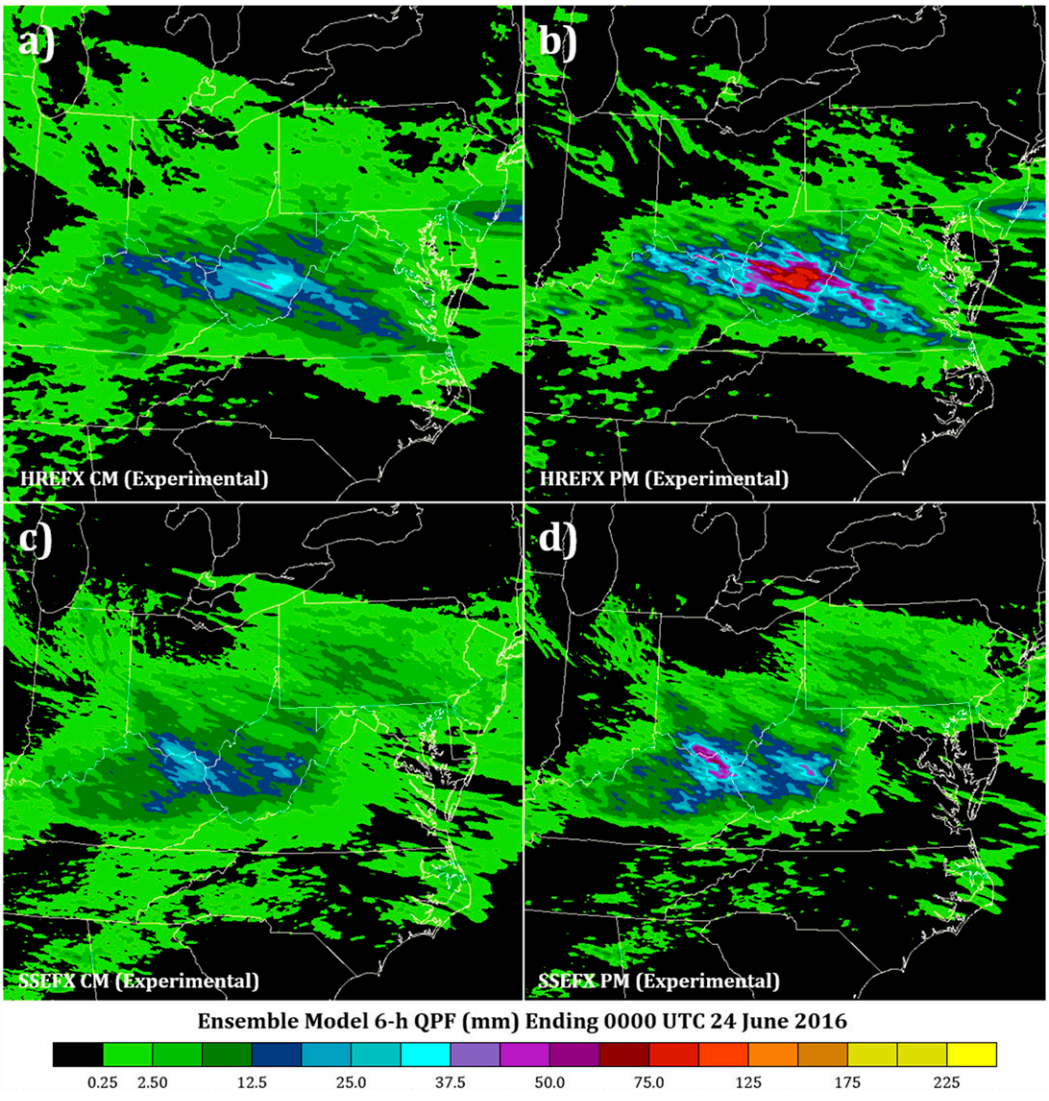

FIG. 9. Ensemble 6-h QPF (mm) ending 0000 UTC 24 Jun 2016 for the experimental HREFX model using the (a) conventional ensemble mean (CM) and (b) probability matched mean (PM) methodologies along with the experimental SSEFX model using the (c) CM and (d) PM methodologies.

b. HMT-Hydro experiment framework

HMT-Hydro experiment real-time operations focused on the issuance of experimental flash flood warnings. Real-time warning operations were generally conducted from 2000 to 0100 UTC. Participants utilized the Hazard Services software (https://esrl.noaa.gov/gsd/eds/hazardservices/) to generate experimental FFWs. Hazard Services software package was designed to integrate existing NWS forecasting functionality and hazard-related product generation (e.g., severe weather warnings) into a single platform (Argyle et al. 2017). The Hazard Information graphical user interface was modified to survey participants about their decision-making process for each experimental FFW issued and allowed participants to assign their own probabilities for minor and major flash flood impacts within the warned area. Contained within the hazard information graphical user interface were prompts asking how specific products influenced the warning decision and the product values at the issuance time regardless of the influence on the decision-making process (Fig. 6).

Subjective formal evaluations were generally conducted the following day. Each subjective formal evaluation focused on a single flash flood event that occurred the previous day. The selection of the flash flood event was based on several factors if multiple flash flood areas occurred. This allowed for the assessment of various flash flood scenarios (e.g., rural versus urban events, minor versus significant events, etc.) throughout each week. Subjective evaluations focused on the spatial coverage and magnitudes of products compared to the flash flood areal extent based on LSRs and stream gauge observations. Products were evaluated during the period from warning

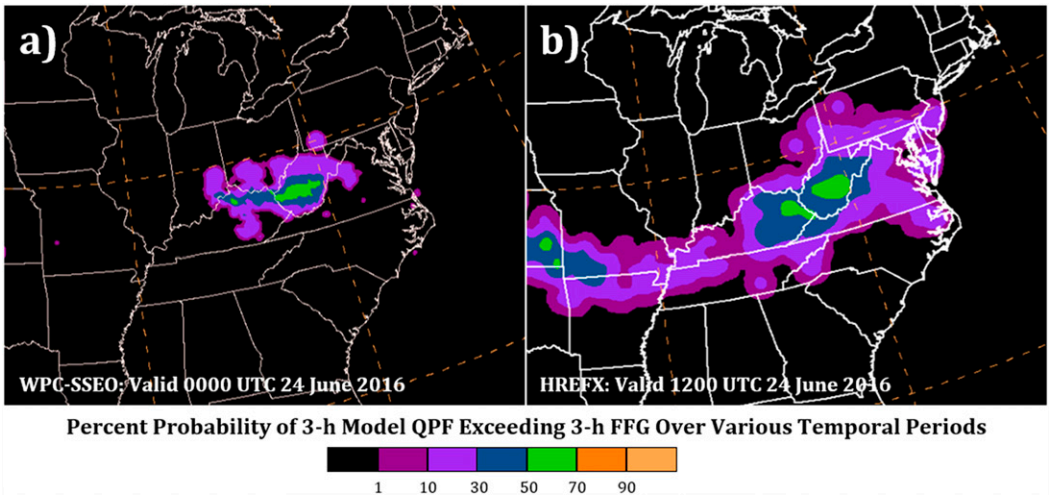

FIG. 10. The percent probability of 3-h model QPF exceeding 3-h FFG values for the (a) 6-h forecast period ending 0000 UTC 24 Jun 2016 from the 0600 UTC 23 Jun 2016 WPC-SSEO model run and (b) 24-h period ending 1200 UTC 24 Jun 2016 from the 0000 UTC 23 Jun 2016 HREFX model run.

issuance to the time of reported flash flooding. Each participant assigned a score ranging from 0 (very poor) to 100 (very good) in 10-point intervals for each product characteristic. Objective verification metrics were also conducted for experimental FFWs after the conclusion of the HMT-Hydro experiment. Emphasis was placed on flash flood detection, warning lead time, and warned area.

c. Scheduled activities surrounding the 23 June 2016 event

Testbed experiment activities related to the 23 June 2016 event began the previous day in the FFaIR experiment with the analysis of NWP models for the creation of the Day 2 ERO valid from 1200 UTC 23 June to 1200 UTC 24 June (Fig. 7). Both testbed experiments focused the majority of 23 June 2016 on the 0–24-h period surrounding the flash flood event presented in this study. Event-related activities with the FFaIR experiment began at 1200 UTC and culminated with the daily FFaIR experiment forecast briefing at 1800 UTC. The FFaIR experiment forecast briefing discussed various experimental NWP and hydrologic model outputs and concluded with the presentation of the experimental Day 1 ERO and 1800–0000 UTC PFFF products. Related HMT-Hydro experiment activities began with the FFaIR experiment forecast briefing at 1800 UTC and continued throughout the day until 0100 UTC 24 June (Fig. 7). The ongoing and forecasted event described in the FFaIR experiment forecast briefing necessitated a change in the HMT-Hydro experiment schedule to allow for immediate experimental real-time operations.

Complicating the start of the HMT-Hydro experiment real-time warning operations were two technical issues. The failure of a 20-ton Liebert cooling unit between 0530 and 0600 UTC forced the shutdown of numerous servers until 1040 UTC overnight due to excessive heat. A secondary network communications issue that impacted the MRMS server around 1630 UTC disrupted the transmission of MRMS QPEs to experimental products, resulting in a temporary degradation of products from no new precipitation forcing. Participants were notified of these issues prior to the start of real-time operations. All operational capabilities for the HMT-Hydro experiment were reestablished after 1900 UTC. Product degradations were resolved over the course of the experimental operations (mostly within the 1900–2000 UTC period). Experimental operations concluded at 0000 UTC to accommodate a rescheduled product evaluation period.

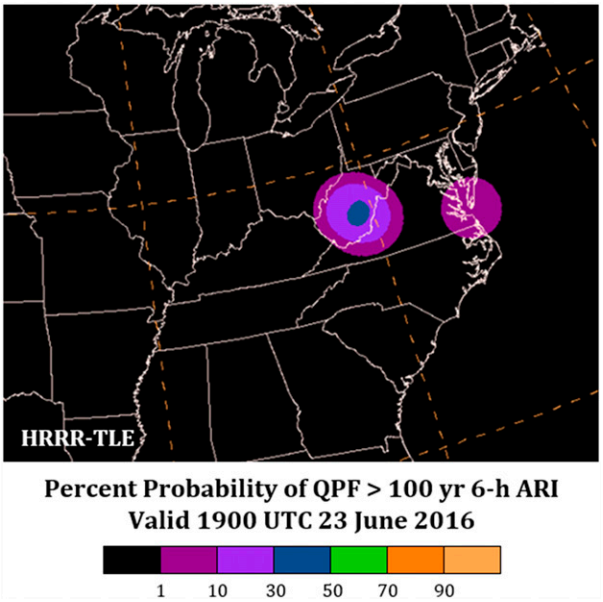

FIG. 11. The percent probability of 6-h QPF exceeding the 100-yr 6-h ARI value from the 1000 UTC 23 Jun 2016 HRRR-TLE model run valid for the time 1900 UTC 23 Jun 2016.

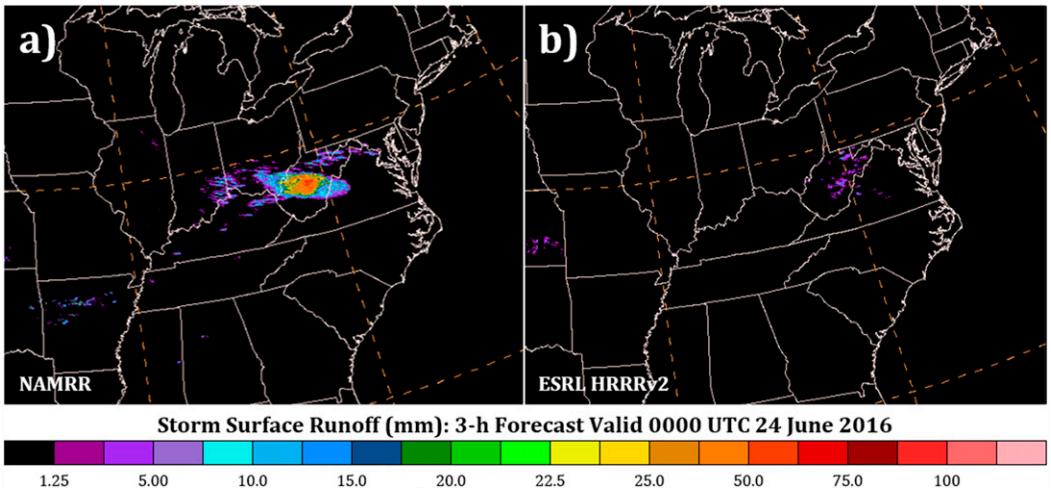

FIG. 12. Storm surface runoff forecast (mm) generated from the (a) 0000 UTC 23 Jun 2016 initialization of the NAMRR model and the (b) 0900 UTC 23 Jun 2016 initialization of the ESRL HRRRv2 model that were valid for the 3-h period ending 0000 UTC 24 Jun 2016.

4. FFaIR experiment

a. Models and guidance for short-term forecasts

Several operational and experiment modeling systems were presented to 2016 FFaIR experiment participants. Deterministic baseline NWP model guidance included the North American Mesoscale Forecast System (NAM; Janjić 2003) along with five Weather Research and Forecasting (WRF) Model–based convection-allowing models (Table 3): two versions of the High-Resolution Rapid Refresh (HRRR and HRRRv2; Benjamin et al. 2016) model along with the Nonhydrostatic Multiscale Model on the B Grid (NMMB; Janjić et al. 2001; Janjić 2003), Advanced Research version of the WRF Model (ARW; Skamarock et al. 2005), and the National Severe Storms Laboratory version of the WRF Model (WRF-NSSL; Skamarock et al. 2008).

Experimental deterministic high-resolution guidance systems (Table 4) included an experimental version of the HRRRv2

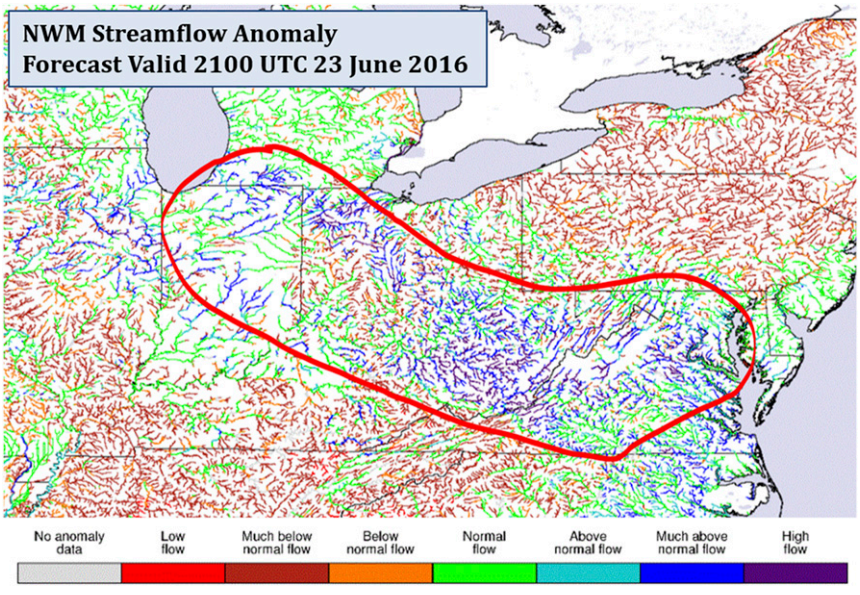

FIG. 13. Short-range forecast of the NWM streamflow anomaly product from the 1200 UTC 23 Jun 2016 run of the NWM valid at 2100 UTC 23 Jun. The red contour highlights the region that was impacted by the initial overnight mesoscale convective system along with the forecast streamflow anomalies over West Virginia and northern Virginia.

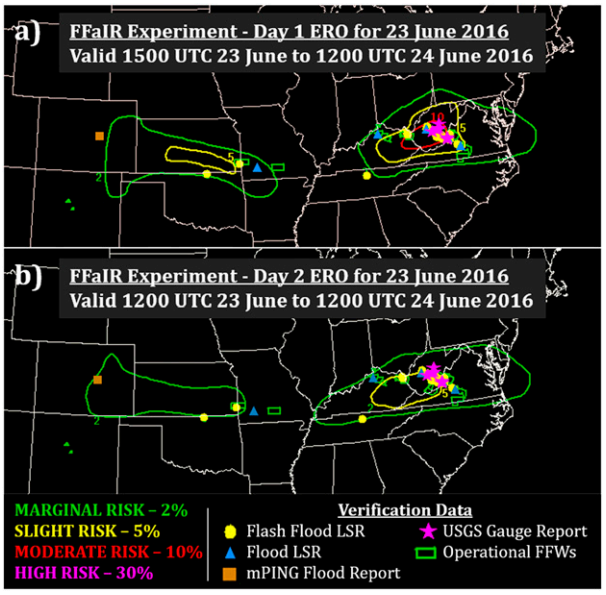

FIG. 14. The FFaIR experiment (a) Day 1 ERO issued at 1415 UTC and valid from 1500 UTC 23 Jun to 1200 UTC 24 Jun 2016 and (b) Day 2 ERO issued at 2000 UTC 22 Jun 2016 valid from 1200 UTC 23 Jun to 1200 UTC 24 Jun 2016. The colored contours designated the various risk levels with corresponding probabilities. Verification for the same period consisted of NWS flash flood LSRs (yellow circles), NWS flood LSRs (blue triangles), mPING flood reports (brown squares), USGS gauge reports exceeding flood stage (pink stars), and NWS operational FFWs issued (green polygons).

provided by the Earth Systems Research Laboratory (ESRL HRRRv2; Benjamin et al. 2016) and the North American Mesoscale Model Rapid Refresh (NAMRR; Carley et al. 2015; Rogers et al. 2009). The NAMRR in 2016 differed from the then operational NAM and NAM CONUS Nest by featuring hourly forecast and assimilation cycles for its 3-km CONUS nest domain.

Five experimental ensemble systems were included in the evaluated experimental model product suite. The HRRR time lagged ensemble (HRRR-TLE; Alexander et al. 2011) consisted of forecasts from multiple deterministic HRRR runs initialized at different hours but valid for the same time. The WPC Storm-Scale Ensemble of Opportunity (WPC-SSEO; Jirak et al. 2012) was a high-resolution, multimodel, multiphysics convection-allowing ensemble with 7 ensemble members, and the experimental Storm-Scale Ensemble Forecast (SSEFX; Snook et al. 2019) produced a 15-member ensemble forecast. The experimental High-Resolution Ensemble Forecast (HREFX; Jirak et al. 2018) utilized multiple cycles of operational convective allowing models with various probabilistic outputs. The beta version of the Global Ensemble Forecast System (GEFSX; Lewis et al. 2017) utilized a new frequency match calibration for QPF generation. More details on each experimental model are provided in Table 4.

Some convective allowing models generated simulated forecast radar reflectivity values, while all deterministic models created QPFs at various temporal accumulations. Emphasis was placed on the 3-, 6-, and 24-h accumulation periods. Two different QPF methodologies within the ensemble HREFX and SSEFX were evaluated. The conventional ensemble mean (CM) averaged all ensemble members over a point. The probability matched mean (PM) methodology combined the spatial pattern of the ensemble mean QPF with the frequency distribution of rainfall rates to improve the ensemble rainfall intensity forecast (Ebert 2001).

Participants evaluated the potential impacts and frequency of forecasted precipitation through comparisons to two NOAA-derived datasets. FFG values generated every 6 h across the 12 CONUS-based NWS River Forecast Centers were compiled by WPC to create a CONUS 5-km resolution

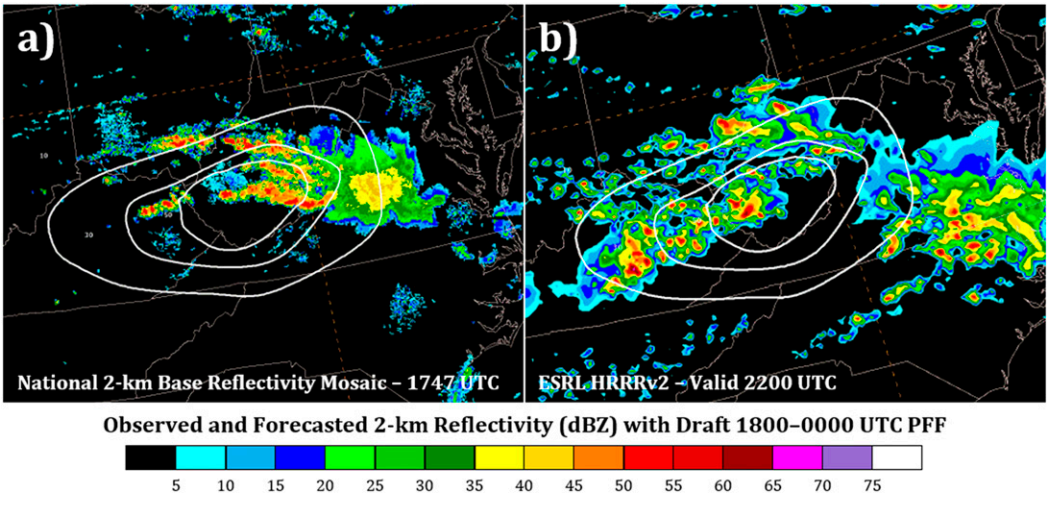

FIG. 15. The 2-km base reflectivity (dBZ) from (a) a national mosaic of radars at 1747 UTC 23 Jun 2016 and (b) the 1500 UTC 23 Jun 2016 model run of the ESRL HRRRv2 valid for 2200 UTC 23 Jun. The white contours are a draft version of the PFFF valid from 1800 UTC 23 Jun 2016 to 0000 UTC 24 Jun 2016.

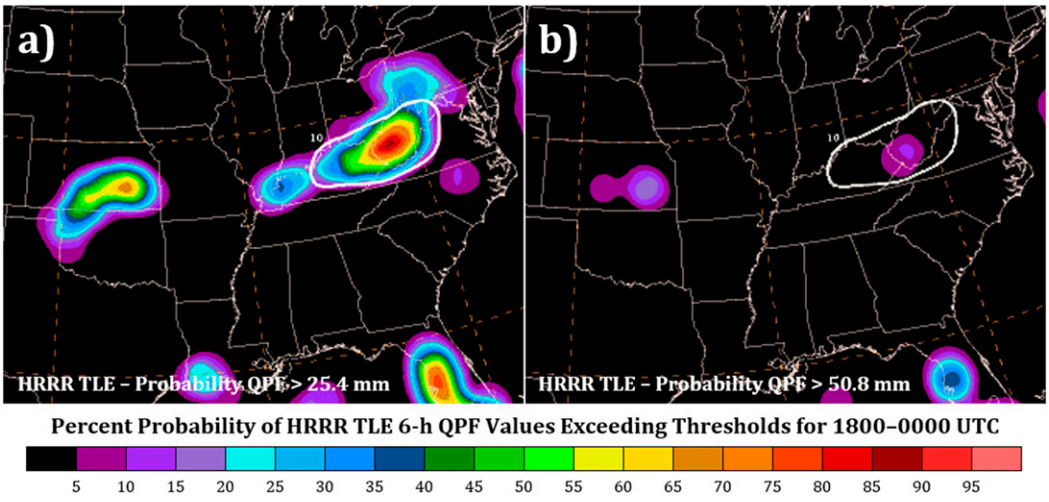

FIG. 16. The percent probability of the 6-h QPF values exceeding (a) 25.4 mm (1.00 in.) and (b) 50.8 mm (2.00 in.) from the 1500 UTC 23 Jun 2016 model run of the HRRR TLE for the period 1800 UTC 23 Jun to 0000 UTC 24 Jun. The white 10% contour is from a draft version of the PFFF valid from 1800 UTC 23 Jun to 0000 UTC 24 Jun 2016.

FFG mosaic. Various experimental ensemble models generated probabilities of exceeding various FFG thresholds. Gridded precipitation average recurrence intervals (ARIs) values from NOAA Atlas 14 climatological precipitation frequency estimates (Perica et al. 2013). ARI values indicate the approximate time between events of a given magnitude when averaged over a long period (Lincoln et al. 2017) and were available for intervals of 2–1000 years during the FFaIR experiment. Ensemble-based models generated probabilities of QPFs exceeding various ARI frequencies to determine the potential rarity of the forecasted precipitation accumulation.

The featured hydrologic model during the 2016 FFaIR experiment was the National Water Model (NWM; e.g., Viterbo et al. 2020). The NWM analysis and forecast system was based on the NCAR-supported community WRF-Hydro hydrologic model (Gochis et al. 2015) that was configured to use the Noah-MP Land Surface Model to simulate land surface processes (Ek et al. 2003). Separate water routing modules performed diffusive wave surface routing and saturated subsurface flow routing on a 250-m grid and Muskingum–Cunge channel routing down National Hydrography Dataset (https://nhd.usgs.gov/) stream reaches. Participants evaluated both the short-range and medium-range soil moisture and streamflow anomaly products from the NWM. Products related to soil moisture availability and runoff through a land surface model within the ESRL HRRRv2 and NAMRR were also utilized as supplemental assessments of antecedent conditions and flood vulnerability forecasting.

b. Day 1 ERO creation

The consensus of multiple operational and experimental deterministic NWP models was forecasting the potential for a significant rainfall event over the area from eastern Kentucky to Virginia during the 6-h period ending 0000 UTC 24 June 2016 (Fig. 8). The NAMRR was the most aggressive experimental model with >175 mm of precipitation accumulation over southwest West Virginia, while the operational and experimental HRRRv2 models were forecasting modest accumulations of 25–75 mm. The different HREFX and SSEFX ensemble model iterations varied in both precipitation location and magnitudes (Fig. 9), with only the HREFX PM forecasting accumulations > 75 mm.

Model QPFs comparisons to NOAA-derived datasets to characterize the potential precipitation impacts and rarity also highlighted the West Virginia and eastern Kentucky region. The WPC-SSEO and HREFX projected a 50%–70% probability of the 3-h FFG being exceeded over West Virginia

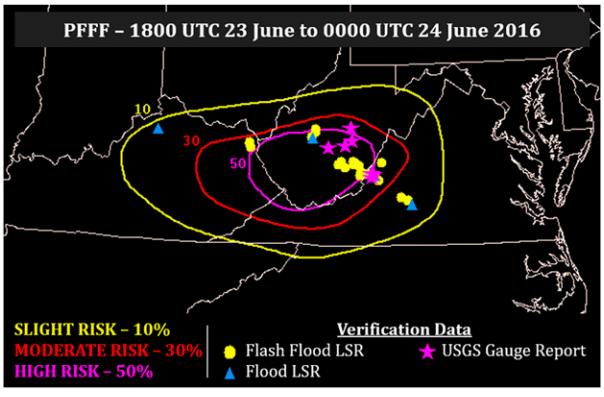

FIG. 17. The PFFF issued at 1745 UTC and valid from 1800 UTC 23 Jun 2016 to 0000 UTC 24 Jun 2016 over West Virginia and the surrounding region. The colored contours designated the various risk levels with corresponding probabilities. Verification for the same period consisted of NWS flash flood LSRs (yellow circles), NWS flood LSRs (blue triangles), and USGS gauge reports exceeding flood stage (pink stars).

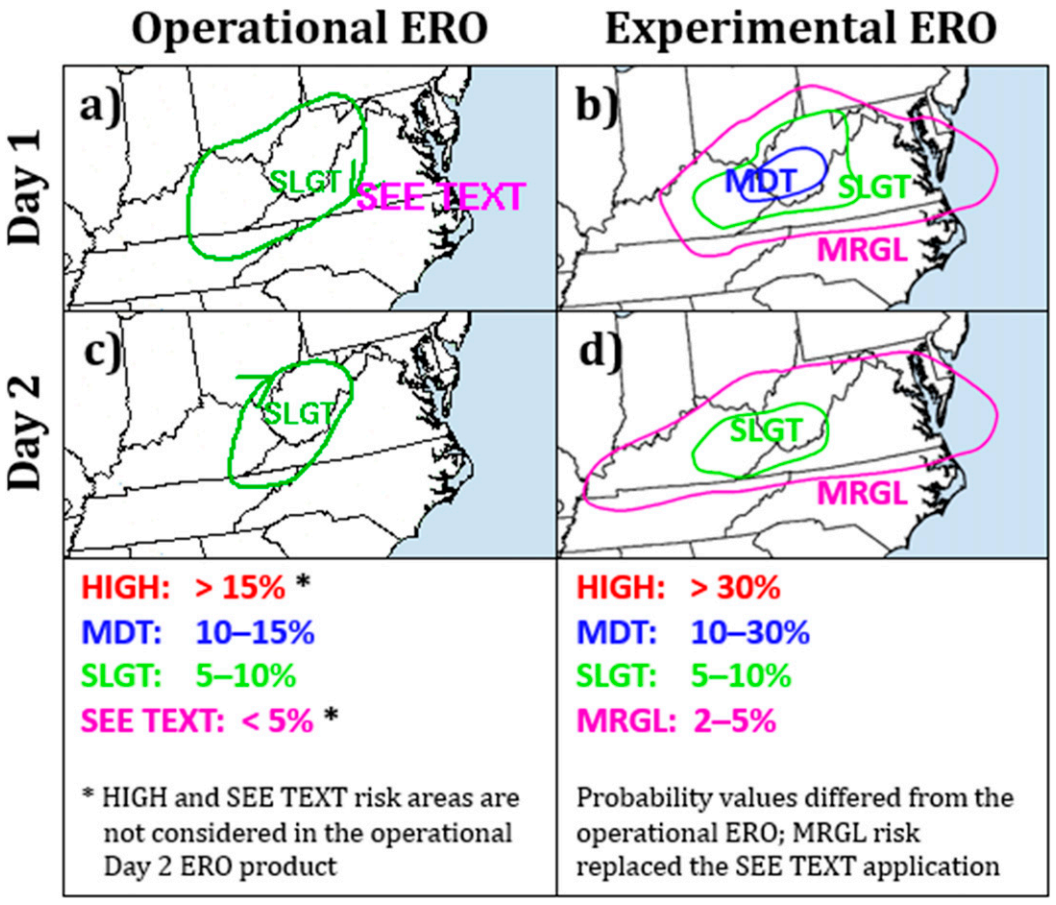

FIG. 18. Comparison of the Day 1 (a) operational ERO vs the (b) experimental ERO along with the Day 2 (c) operational ERO vs the (d) experimental ERO. The experimental Day 1 ERO was issued at 1415 UTC 23 Jun, and the operational Day 1 ERO was issued at 1459 UTC 23 Jun. Both Day 1 EROs were valid from 1500 UTC 23 Jun to 1200 UTC 24 Jun 2016. The experimental Day 2 ERO was issued at 2000 UTC 22 Jun, and the operational Day 1 ERO was issued at 1956 UTC 22 Jun. Both Day 2 EROs were valid from 1200 UTC 23 Jun to 1200 UTC 24 Jun 2016. Percentages for the defined risk categories differed between the operational and experimental versions of the ERO. The operational ERO does not consider high risk or "see text" areas on the Day 2 ERO. The experimental EROs contoured the marginal risk area that was equivalent to the "see text" risk area depicted in the operational ERO.

(Fig. 10a), while the HREFX also included a broad 50%–70% probability region over eastern Kentucky (Fig. 10b). The area of >30% probability of exceeding 3-h FFG was confined to West Virginia and the far northern extent of Kentucky in the WPC-SSEO, while the HREFX depicted a larger >30% probability area extending into western Kentucky and western Virginia. The HRRR-TLE depicted a 30%–50% probability of QPF exceeding the 100-yr 6-h ARI over central West Virginia,

TABLE 5. Subjective scoring of the various operational and experimental model QPFs for the 23 Jun 2016 event and the overall 2016 FFaIR experiment. The ensemble-based HREFX and SSEFX QPFs were evaluated using both the conventional ensemble mean (CM) and probability matched mean (PM) methodologies. Ratings were based on a score from 1 (very poor) to 10 (very good). The 23 Jun 2016 scoring had a sample size of eight for each model. The overall FFaIR scoring had a sample size ranging from 100 to 120 from 30 participants over the four weeks of the 2016 FFaIR experiment (there were some instances when a model was not available for a particular day).

Product status	Model	Model type	23 Jun 2016 scoring	Overall FFaIR scoring
Operational	NAM	Deterministic	6	3.84
	HRRRv2 (NCEP)	Deterministic	5	5.41
	WRF-NSSL	Deterministic	6	5.27
Experimental	ESRL HRRRv2	Deterministic	4	5.71
	NAMRR	Deterministic	7	4.77
	HREFX CM	Ensemble	5	4.48
	HREFX PM	Ensemble	8	4.99
	SSEFX CM	Ensemble	5	4.90
	SSEFX PM	Ensemble	5	5.19

TABLE 6. MRMS products available over the entire CONUS during 2016 HMT-Hydro experiment. Listed are the scale of each product as displayed during experimental operations and the spatiotemporal resolution. The spatial resolution of 0.01° × 0.01° is approximately 1 km × 1 km. Products that underwent specific evaluations during the 2016 HMT-Hydro experiment and for the 23 Jun 2016 event are denoted by an asterisk. See Zhang et al. (2016) for more information about the MRMS products. Table adapted from Martinaitis et al. (2017).

Product	Display scale	Spatiotemporal resolution
Seamless hybrid scan reflectivity	−30 to 100 dBZ	0.01° × 0.01°; 2 min
Radar quality index	0.0–1.0	0.01° × 0.01°; 2 min
Surface precipitation type	—	0.01° × 0.01°; 2 min
Surface precipitation rate	0–254 mm h^{-1}	0.01° × 0.01°; 2 min
Radar-only QPE (1-h accumulations)*	0–76 mm	0.01° × 0.01°; 2 min
Radar-only QPE (3-h accumulations)*	0–76 mm	0.01° × 0.01°; 1 h
Radar-only QPE (6-, 12-h accumulations)	0–152 mm	0.01° × 0.01°; 1 h
Radar-only QPE (24-h accumulations)	0–254 mm	0.01° × 0.01°; 1 h

which characterized the potential rarity of the forecasted precipitation (Fig. 11).

Hydrologic-related guidance depicted varied overland and channel impacts. The modeled 3-h surface runoff responses diverged significantly in magnitude for the 3-h period ending 0000 UTC 24 June. The NAMRR created a widespread area of >7.5 mm of surface runoff with localized areas exceeding 50 mm in central West Virginia (Fig. 12a); however, the maximum forecast surface runoff from the ESRL HRRRv2 was <10 mm across eastern portions of West Virginia (Fig. 12b) despite modeled soil saturation for the top 1.0-cm layer exceeding 95% (not shown). The NWM streamflow anomaly product highlighted "high flow" anomalies over West Virginia and northern Virginia based on earlier precipitation that moved across the region and forecast precipitation based on HRRRv2 QPFs (Fig. 13).

The combination of the various NWP precipitation guidance and the hydrologic-based forecasts prompted the issuance of a moderate risk area over West Virginia for the Day 1 ERO (Fig. 14a). A slight risk area extended from east-central Kentucky into western Virginia. This was an escalation from the slight risk that was forecast in the experimental Day 2 ERO valid for the 1200 UTC 23 June–1200 UTC 24 June period (Fig. 14b). A secondary slight risk area was noted in the Day 1 ERO over southern Kansas with a marginal risk contoured area extending from eastern Colorado to northwest Arkansas.

c. *1800–0000 UTC PFFF creation*

The Day 1 ERO moderate risk area over West Virginia became the focus of the experimental 1800–0000 UTC PFFF. Model interrogation initially focused on observed 2-km base radar reflectivity values (Fig. 15a) and forecast composite reflectivity radar simulations during the 6-h forecast period (e.g., Fig. 15b). Model exceedance probabilities then highlighted the significant rainfall threat over the 1800 UTC 23 June–0000 UTC 24 June period. The HRRR TLE probability of exceeding a 25.4 mm accumulation over a 6-h period had a pronounced signal (>85%) over West Virginia (Fig. 16a).

TABLE 7. FLASH products available over the entire CONUS during 2016 HMT-Hydro experiment. Products listed include two QPE comparison tools, various hydrologic model outputs from the featured Coupled Routing and Excess Storage (CREST) model along with the Sacramento Soil Moisture Accounting (SAC-SMA; Burnash et al. 1973) and hydrophobic (HP) water balance models. Listed are the scale of each product as displayed during experimental operations and the spatiotemporal resolution. The spatial resolution of 0.01° × 0.01° is approximately 1 km × 1 km. Products that underwent specific evaluations during the 2016 HMT-Hydro experiment and for the 23 Jun 2016 event are denoted by an asterisk. See Gourley et al. (2017) for more information about the FLASH products. Table adapted from Martinaitis et al. (2017).

Product	Display scale	Spatiotemporal resolution
QPE-to-FFG ratio (1, 3, 6 h; maximum value)*	0%–500%	0.01° × 0.01°; 2 min
QPE ARI (30 min, 1, 3, 6, 12, and 24 h; maximum value)*	0–200 yr	0.01° × 0.01°; 2 min
CREST streamflow (forcing: MRMS radar-only QPE)	0–100 000 $m^3 s^{-1}$	0.01° × 0.01°; 10 min
CREST streamflow (forcing: ESRL HRRRv2 QPF)	0–100 000 $m^3 s^{-1}$	0.01° × 0.01°; 10 min
CREST unit streamflow (forcing: MRMS radar-only QPE)*	0–20 $m^3 s^{-1} km^{-2}$	0.01° × 0.01°; 10 min
CREST unit streamflow (forcing: ESRL HRRRv2 QPF)*	0–20 $m^3 s^{-1} km^{-2}$	0.01° × 0.01°; 10 min
CREST soil moisture content (forcing: MRMS radar-only QPE)	0%–100%	0.01° × 0.01°; 10 min
SAC-SMA streamflow (forcing: MRMS QPE)	0–100 000 $m^3 s^{-1}$	0.01° × 0.01°; 10 min
SAC-SMA unit streamflow (forcing: MRMS QPE)	0–20 $m^3 s^{-1} km^{-2}$	0.01° × 0.01°; 10 min
HP streamflow (forcing: MRMS radar-only QPE)	0–100 000 $m^3 s^{-1}$	0.01° × 0.01°; 10 min
HP unit streamflow (forcing: MRMS radar-only QPE)	0–20 $m^3 s^{-1} km^{-2}$	0.01° × 0.01°; 10 min
SAC-SMA soil moisture content (forcing: MRMS radar-only QPE)	0%–100%	0.01° × 0.01°; 10 min
GFS prediction probability random forest model	0%–50%	0.25° × 0.25°; 6 h

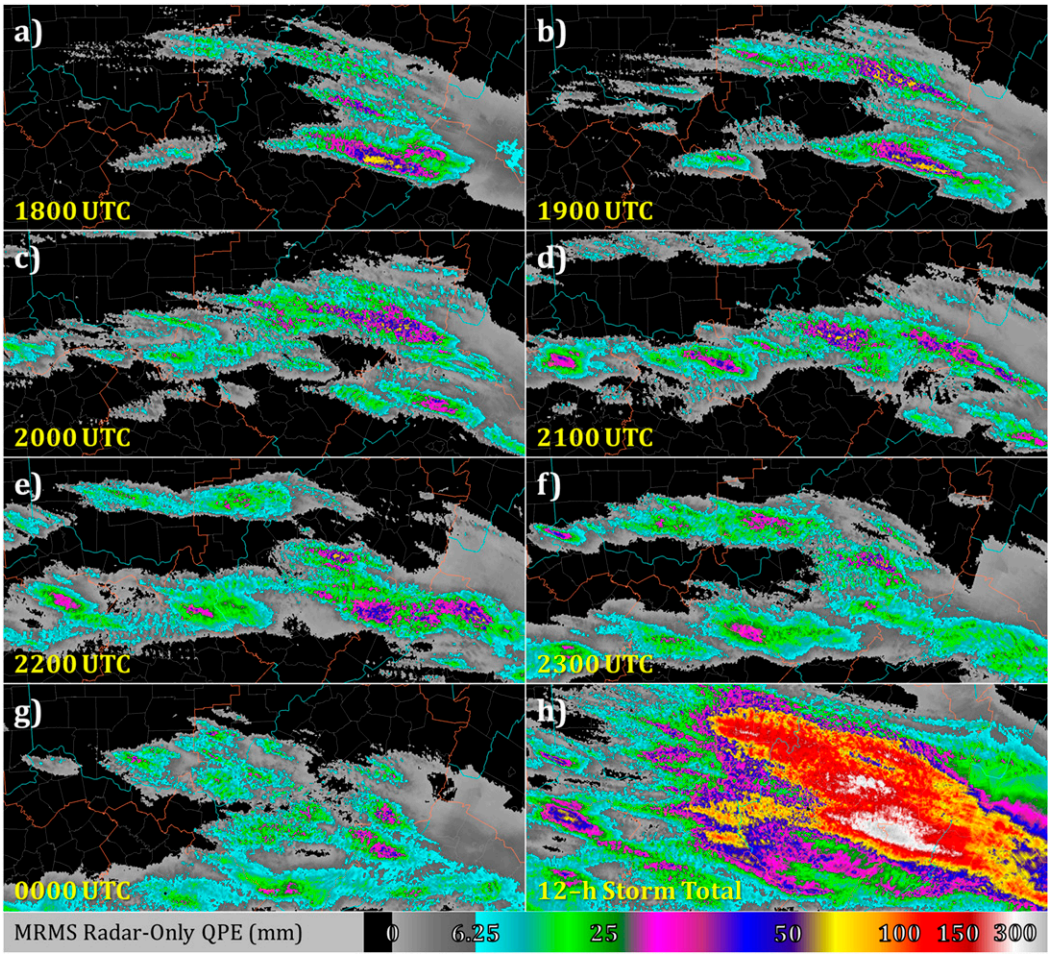

FIG. 19. Hourly MRMS radar-only QPE for (a)–(g) 1800 UTC 23 Jun 2016–0000 UTC 24 Jun 2016 with (h) the 12-h QPE accumulation of the MRMS radar-only QPE ending 0000 UTC 24 Jun 2016.

The probability of exceeding a 50.8 mm 6-h accumulation produced a signal of 10%–15% over the same region (Fig. 16b). Participant feedback showed that the nonzero probability of exceeding a 50.8 mm 6-h accumulation in conjunction with other aforementioned 6-h precipitation QPF signals from the various experimental models, signals in the hydrologic model output, and real-time radar and satellite trends provided increased confidence in forecasting a significant hydrometeorological event. The experimental 1800 UTC 23 June to 0000 UTC 24 June 2016 PFFF placed a high risk probability over southern West Virginia (Fig. 17). The slight risk areal extent in the PFFF covered a similar area of the Day 1 ERO slight risk contour (Fig. 14a).

d. Evaluations and observations

The experimental Day 1 ERO received an average score of 8.25. Deductions in scoring were primarily based on the orientation of the moderate risk area compared to the orientation of the verification observations (Fig. 14a). The experimental Day 2 ERO valid for 23 June 2016 was given an average score of 8.00 based on the large slight risk area capturing the event (Fig. 14b). The issuance of a moderate risk in the experimental Day 1 ERO diverged from the operational Day 1 ERO that retained a slight risk for rainfall exceeding FFG (Figs. 18a,b); moreover, the slight risk area in the experimental Day 1 ERO was more confined to the reported flash flooding than the operational version. The experimental risk increase was likely contributed to the experimental model guidance that was not available to operational forecasters. The slight risk coverage area in the experimental Day 2 ERO was similar to the operational Day 2 ERO but with differing orientations (Figs. 18c,d).

Participants subjectively rated the 1800–0000 UTC PFFF highly with an average score of 9.63. The provided feedback noted the PFFF moderate and high risk contours were well placed and captured the majority of the observations but potentially had too high probabilities west of the greater concentration of reports (Fig. 17). There were no comparative operational products to the experimental 1800–0000 UTC PFFF.

Subjective evaluations of deterministic and ensemble models documented the various heavy rainfall signals over the West Virginia area and the influences of these strong

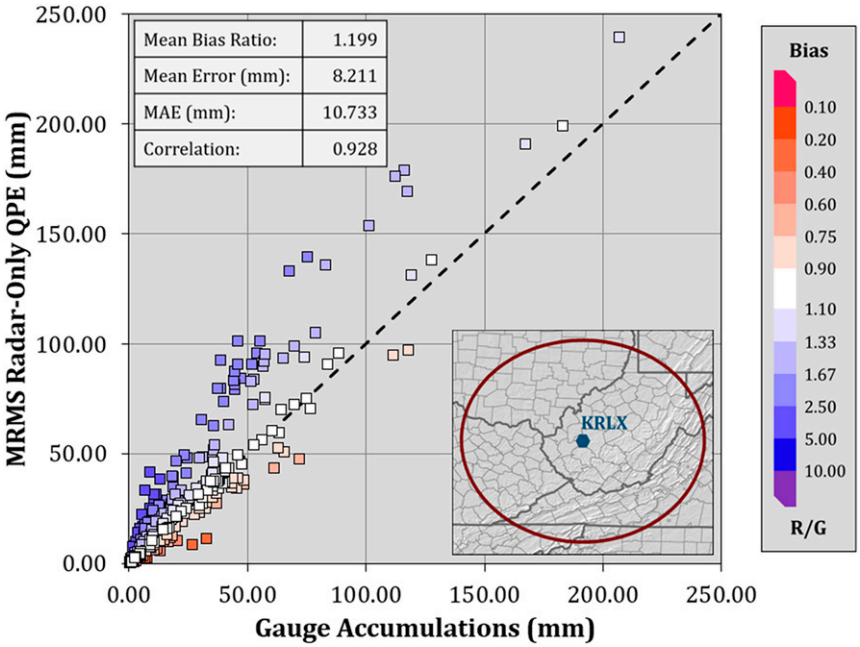

FIG. 20. Scatterplot of 24-h MRMS radar-only QPE (mm) vs gauge observations ending 0000 UTC 24 Jun 2016. The dashed line represents the one-to-one line between gauge and MRMS radar-only QPE values. Statistical evaluations of the mean bias ratio, mean error, mean absolute error (MAE), and correlation coefficient are shown in the upper-left corner. The mean bias ratio was calculated using the sum of the MRMS radar-only QPE divided by the sum of the gauge observations. Map inset showing the location of the Charleston, West Virginia, WSR-88D radar (KRLX) and a 230-km range ring to depict where the area analyzed is shown in the lower-right corner.

signals on participant confidence regarding the Day 1 ERO and associated PFFF issuance. All scored model evaluations emphasized the 6-h precipitation accumulation period from 1800 UTC 23 June to 0000 UTC 24 June 2016 and compared to the associated 6-h MRMS radar-only QPE (Fig. 8f). The experimental NAMRR was evaluated as the best deterministic model with a rating of 7, followed by the operational NAM and WRF-NSSL with a rating of 6 (Table 5). The NAMRR QPF pattern was similar to the MRMS radar-only QPE but farther west and with greater accumulations than the observed precipitation (Figs. 8e,f). Both the operational HRRRv2 and experimental ESRL HRRRv2 received the lowest subjective rating among the NWP models on 23 June, which contrasted to their receiving the best overall average scores throughout the 2016 FFaIR experiment (Table 5). Both HRRRv2 versions significantly underestimated the precipitation accumulations (Figs. 8b,d); moreover, the ESRL HRRRv2 struggled with precipitation coverage and orientation, notably having the local maximum accumulation located significantly farther south when compared to the other models and the observed QPE.

The ensemble PM technique was shown to be more representative versus the CM technique, and the ensemble HREFX PM QPF (Fig. 9b) was subjectively determined to be the best overall QPF with a rating of 8 (Table 5). The HREFX PM method had proper orientation and location of the precipitation despite a slight underestimation bias. The HREFX CM method had a similar precipitation orientation and location, yet the maximum QPF underestimated by 75% (Fig. 9a). Both SSEFX methodologies displaced the local QPF maximum to the west into northeast Kentucky with differing precipitation orientations and significant precipitation underestimation biases (Figs. 9c,d).

5. HMT-Hydro experiment

a. Warning decision-making products

Featured gridded products in the 2016 HMT-Hydro experiment were from the operational MRMS system (Zhang et al. 2016) and the experimental Flooded Locations and Simulated Hydrographs (FLASH; Gourley et al. 2017) system. The entire FLASH system product suite (Table 6) along with select MRMS reflectivity and QPE-related products (Table 7) were available CONUS-wide to the participants. The remaining MRMS product suite was available within a strategically positioned floating domain based on forecasted flash flooding. Four MRMS and FLASH products were the focus of experimental operations and follow-up assessments given their potential application to real-time flash flood detection based on past HMT-Hydro experiments (Martinaitis et al. 2017). Other products from the FLASH system were also available to assist in product issuance, yet those products were employed to a

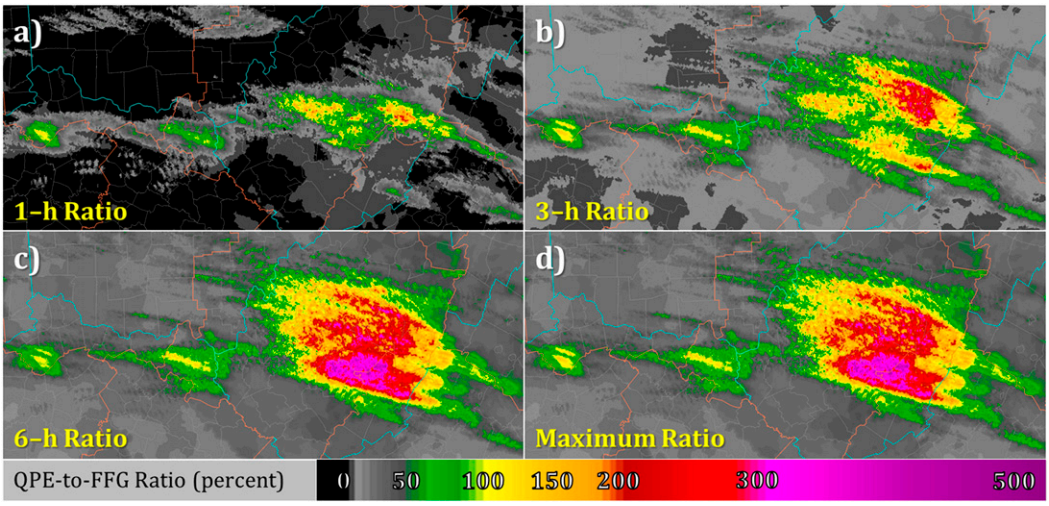

FIG. 21. QPE-to-FFG ratio (shown in percentage) at 2100 UTC 23 Jun 2016 for (a) 1-h QPE accumulation period, (b) 3-h QPE accumulation period, (c) 6-h QPE accumulation period, and (d) the maximum QPE-to-FFG ratio value for each grid cell based on the aforementioned QPE accumulation periods.

lesser degree within the decision-making process and not featured in this study (Table 7).

The MRMS radar-only QPE was generated from a seamless mosaic of quality-controlled single-radar reflectivities combined with NWP variables to apply different reflectivity–rate relationships at each 0.01° × 0.01° grid cell (Zhang et al. 2016). Emphasis was placed on the MRMS radar-only QPE given its real-time availability (latency < 90 s) and its ingest into the FLASH system. Localized hourly QPE accumulations exceeded 63.5 mm, and rainfall estimates for the 12-h period ending 0000 UTC 24 June 2016 were 75–240 mm (Fig. 19). The 24-h MRMS radar-only QPE accumulations within 230 km of the Charleston, West Virginia, WSR-88D (KRLX) overestimated by approximately 20% (Fig. 20), which would likely inflate values within the FLASH product suite.

The QPE-to-FFG ratio product was derived from the WPC-mosaicked FFG grids and compared to 1-, 3-, and 6-h MRMS radar-only QPE accumulations. This was similar to operational comparisons of QPE to FFG values, but the version within the FLASH system was based solely on MRMS radar-only QPEs and depicted at the MRMS spatiotemporal resolution. A maximum QPE-to-FFG ratio value was also generated between those three accumulation periods. Ratio values exceeding 1.00 (i.e., QPE > 100% of FFG) implied that the amount of MRMS radar-only QPE accumulated over a given accumulation period would exceed guidance for bank-full conditions. QPE values surpassed 100% of FFG at the 1-h time scale, while QPE-to-FFG ratio values were >2.00 over areas that experienced more significant rain rates. QPE-to-FFG ratios were substantially greater for the longer temporal

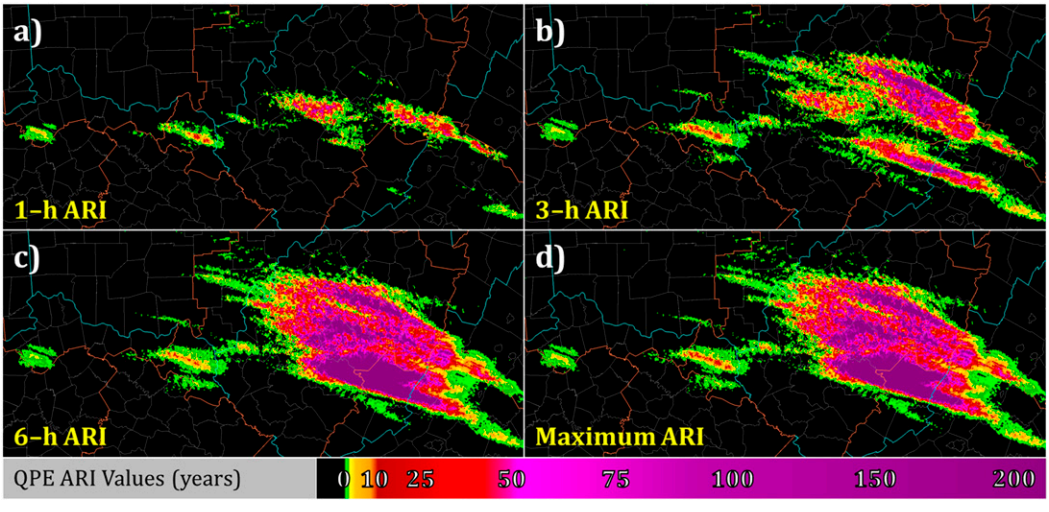

FIG. 22. As in Fig. 21, but for the QPE average recurrence intervals (ARIs) shown in years.

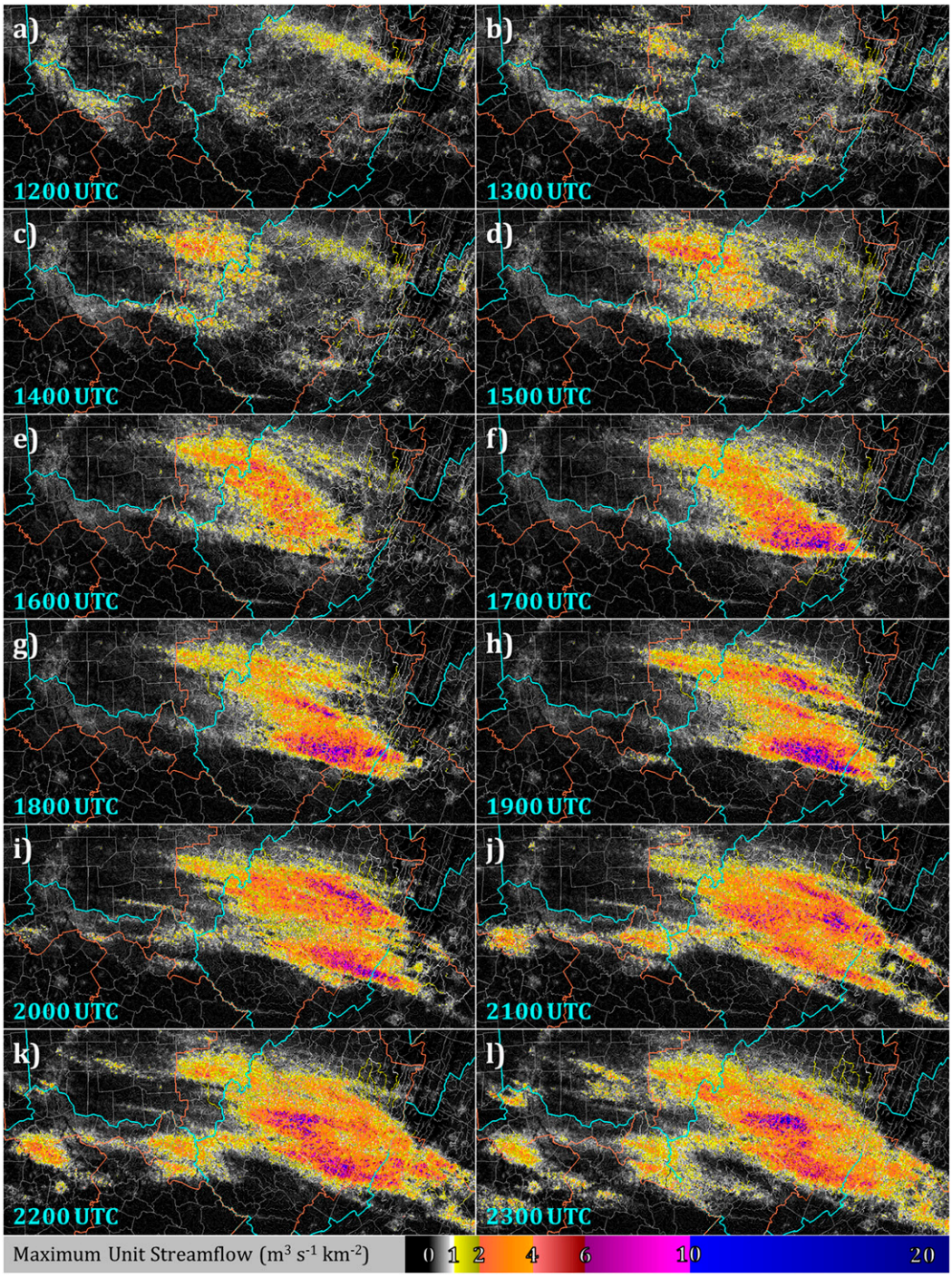

FIG. 23. CREST maximum unit streamflow ($m^3\ s^{-1}\ km^{-2}$) for every hour from (a) 1200 to (l) 2300 UTC 23 Jun 2016 across the region from northern Kentucky to western Virginia.

accumulations, notably with widespread regions of QPE exceeding FFG by 300%–400% for 6-h accumulations (Fig. 21). The exceedance of FFG across the various accumulation periods demonstrated that the flash flooding observed during this event was both rate-driven and driven by long-term accumulations.

Rainfall accumulations compared to gridded static ARIs from NOAA Atlas 14 point precipitation frequency estimates highlighted the potential rarity of the estimated rainfall. ARI comparisons to the MRMS radar-only QPE in the FLASH system were generated for accumulation periods from 30 min to 24h along with a maximum value across

TABLE 8. Subjective spatial coverage scoring of the primary MRMS and FLASH products for the Kanawha County, West Virginia, and Boyd County, Kentucky, flash flood events along with the average overall score during the 2016 HMT-Hydro experiment. Ratings were based on a score from 0 (very poor) to 100 (very good). The 23 Jun 2016 scoring had a sample size of five for each product. The overall HMT-Hydro scoring had a sample size of 59 from a total of 16 participants across the three weeks of the 2016 HMT-Hydro experiment.

Product (spatial coverage)	Kanawha County, WV	Boyd County, KY	Overall HMT-Hydro scoring
MRMS radar-only QPE	80	86	75.93
QPE-to-FFG ratio	94	78	74.57
QPE ARI	82	68	72.37
CREST maximum unit streamflow	94	80	75.25

all analyzed time periods; however, the use of ARIs in the HMT-Hydro experiment focused on the 1-, 3-, and 6-h QPE accumulations. The ARI product indicated that the greater QPE accumulations had at least a 2.0% exceedance probability (50 yr) for 1-h totals and 1.0% exceedance probability (100 yr) for 3-h totals (Fig. 22). The 6-h QPE accumulations had widespread areas of 0.5% exceedance probabilities (200 yr), the maximum value calculated in the FLASH system.

The experimental FLASH system output designed to be most applicable to flash flood detection and forecasting was the Coupled Routing and Excess Storage (CREST; Wang et al. 2011) hydrologic model maximum unit streamflow (Gourley et al. 2017). The CREST model was forced by the MRMS radar-only QPE and relied on mass balance and kinematic wave routing, among other parameterizations, to generate forecast surface hydrologic conditions for each grid cell out to 12 h. Unit streamflow is defined as a normalization of the discharge by the upstream drainage area, which allowed participants to focus on areas experiencing anomalous flows in the overland cell and in small streams as part of the flash flood warning decision-making process (Martinaitis et al. 2017). Previous HMT-Hydro experiments observed that unit streamflow values of 1.0–2.0 $m^3 s^{-1} km^{-2}$ were likely to lead to the issuance of a FFW (Martinaitis et al. 2017). A real-time version of the CREST maximum unit streamflow product forced with ESRL HRRRv2 QPFs was also presented in the HMT-Hydro experiment.

The evolution of the CREST maximum unit streamflow product from 1200 to 2300 UTC 23 June highlighted the initial hydrologic model response from precipitation that passed through the region between 0700 and 1200 UTC and the influences from the following waves of precipitation that traversed the area through 2300 UTC (Fig. 23). A broad region of maximum unit streamflow values of 1.0–3.0 $m^3 s^{-1} km^{-2}$ were observed throughout the region with localized areas of 6.0–10.0 $m^3 s^{-1} km^{-2}$.

b. MRMS and FLASH product subjective evaluations

Two detailed subjective product evaluations for the 23 June 2016 event were conducted. The catastrophic event in northern Kanawha County, West Virginia, was chosen for evaluation due to the historical significance of the event; moreover, there was sufficient FLASH system recovery from the aforementioned technical issues to allow for a product evaluation. The isolated minor flash flood event in Boyd County within far northeast Kentucky was chosen to assess the FLASH system performance during a more isolated and minor flash flood hazard.

All evaluated products for both events were rated favorably regarding the capturing of the flash flood spatial coverage and rated above the overall average for each respective product except for the QPE ARI product for the northeast Kentucky event (Table 8). Participants noted that the coverage area of greater ARI values was too broad and displaced to the south of the verified flash flooding. The magnitude of the MRMS radar-only QPE and the QPE-to-FFG ratio product were also perceived to match the event well (Table 9). The CREST maximum unit streamflow values were rated lower for both events but scored above the overall testbed average. Product degradation from the earlier technical issues potentially biased the participant scoring of the magnitude of values for the Kanawha County event despite the system recovery. The MRMS radar-only QPE overestimation bias could have also compensated for some product magnitude degradation.

The addition of ESRL HRRRv2 QPFs into the CREST hydrologic model were evaluated in real-time operations and the postevent assessment. The spatiotemporal and magnitude variances in precipitation in the ESRL HRRRv2 versus observations hindered the general use of short-term QPFs to extend FFW lead time. Follow-up assessments were characterized by mixed agreement on the utility of ESRL HRRRv2 QPFs to improve warning lead time. Subjective participant consensus for potential additional warning lead time

TABLE 9. As in Table 8, but for the product magnitude.

Product (magnitude)	Kanawha County, WV	Boyd County, KY	Overall HMT-Hydro scoring
MRMS radar-only QPE	90	78	72.71
QPE-to-FFG ratio	90	78	61.36
QPE ARI	74	52	58.64
CREST maximum unit streamflow	72	70	66.78

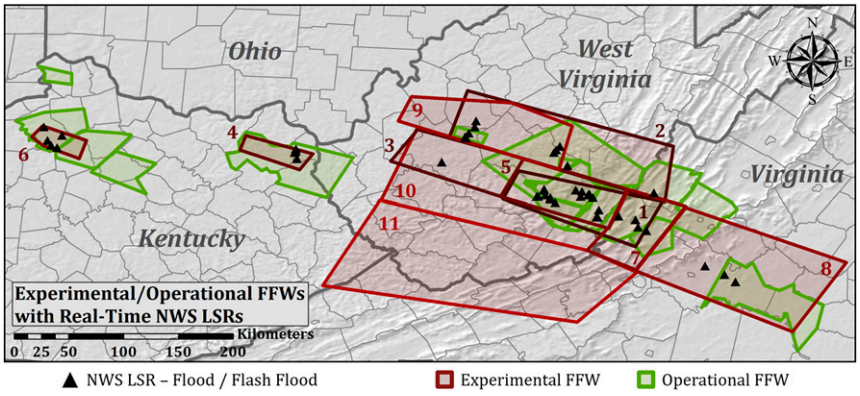

FIG. 24. Experimental FFWs (red polygons) issued by participants during the HMT-Hydro experiment and operational FFWs (green polygons) issued by NWS forecasters with NWS flash flood and flood LSRs that occurred in real time. The number listed by each FFW corresponds to the FFW information provided in Table 10.

improvements from the ingest of ESRL HRRRv2 QPFs ranged from no perceived additional benefits to up to an approximately additional 30 min.

c. FFW issuance and analysis

An initial experimental FFW was generated by the participants at 1838 UTC to define the ongoing flash flooding in order to delineate where future FFW polygons could be issued (Fig. 24; Table 10); therefore, postevent objective analysis did not include the first experimental FFW. A total of 10 additional experimental FFWs were issued across the area from northern Kentucky to central Virginia (Fig. 24). Five experimental FFWs verified during the event (Table 10). There was one missed flash flood event, which occurred between the expiration of experimental FFW 1 and the issuance of experimental FFW 5 for the expired area. The probability of detection for the experimental FFWs was 0.83, which was an increase of 0.12 over the operational FFWs that were relevant to the flash flooding observed during the experimental real-time warning operations; however, the false alarm ratio increased from 0.29 for operational FFWs to 0.44 for the experimental FFWs (Table 11). The critical success index values were similar between the two warning datasets, yet the average lead time for verified experimental FFWs improved by 32.8 min over the operational FFWs to 84.2 min. The 63.8% increase in lead time was remarkable given that participants were not permitted to view operational FFWs or forecast products, lacked local knowledge of areas more susceptible to flash flooding, and the earlier technical challenges that impacted the FLASH products. It was also noted that the experimental FLASH product suite was not available to operational forecasters.

The two isolated flash flooding events in Kentucky demonstrated the potential for reduced warning false alarm area with

TABLE 10. List of experimental FFWs issued during the HMT-Hydro experiment. Included are the start and end times of each experimental FFWs (in UTC), the assigned probabilities of minor and major flash flooding, the verification of each experimental FFW, and the lead time based on LSRs received in real time during the event. The asterisk for FFW 1 denotes that this FFW was issued at the onset of the operational period to denote the ongoing threat area at the start of the experimental operations and was not included in the postevent warning evaluations.

FFW No.	Start time (UTC)	End time (UTC)	Probabilities		Verified	Lead time (min)
			Minor	Major		
1*	1838	2038	—	—	—	—
2	1959	2259	80	20	Yes—Major	1
3	2036	2236	60	10	No	—
4	2045	2345	70	10	Yes—Minor	73
5	2100	2300	40	0	Missed event	—
6	2108	0008	70	10	Yes—Minor	37
7	2204	0004	50	0	No	—
8	2216	0016	50	10	Yes—Minor	104
9	2234	0134	90	30	Yes—Major	206
10	2316	0116	100	80	No	—
11	2342	0142	70	30	No	—

TABLE 11. FFW metrics for the experimental warnings and operational warnings related to the flash flooding that occurred from 1650 UTC 23 Jun to 0100 UTC 24 Jun 2016. Listed are the number of hit events (i.e., verified FFWs), the number of missed flash flood events, the number of unverified FFWs, the probability of detection (POD), the false alarm ratio (FAR), the critical success index (CSI), and the average FFW lead time of the hit events (min).

	Hit events	Missed events	Unverified FFWs	POD	FAR	CSI	Lead time (min)
Operational	10	4	4	0.71	0.29	0.56	51.4
Experimental	5	1	4	0.83	0.44	0.50	84.2

the incorporation of the FLASH product suite. Experimental FFW 4 that covered the flash flooding in Boyd County encompassed a region that was 1905 km^2 less than the collocated operational FFW issued 7 min later (Table 12). The flash flooding threat covered by experimental FFW 6 was operationally warned by two FFWs from neighboring NWS forecast offices, which were issued 10 and 35 min prior to the experimental FFW, respectively. The warned area for experimental FFW 6 was 1744 km^2 less than the two combined operational FFWs. Both warned areas were reduced by >70% in the HMT-Hydro experimental real-time warning operations.

The surveyed experimental FFW decision-making processes for all warnings highlighted the familiarity of current operational products and contrasting results between the two QPE comparison products. Participants noted that the MRMS radar-only QPE and the QPE-to-FFG ratio product, the two concepts used operationally, greatly influenced 90% and 80% of their warning decisions, respectively (Fig. 25a); however, the QPE ARI product influence on the warning decision-making process contrasted with the QPE-to-FFG ratio product. Participants stated that QPE ARIs had no great influence on the decision-making process; moreover, 60% of the warnings issued were not influenced at all by QPE ARIs (Fig. 25a). Experimental FFWs where QPE ARIs were deemed to have no influence on the warning issuance had ARI values < 6 yr. The QPE ARI product had some influence in the warning decision-making process when ARI values increased to 10–20 yr. Participants noted that the QPE ARI values were supportive of the overall situational awareness of significant rainfall accumulations that could generate flash flooding, yet participants had little confidence in constructing a range of ARI values that coincided with potential flash flooding due to widely varying values while having a perceived high bias in the ARI values.

The CREST maximum unit streamflow product greatly influenced 60% of the analyzed warning decisions during this event. Similar distributions of surveyed influences were noted between verified and unverified experimental FFWs (Figs. 25b,c). Values observed by participants for nine of the experimental FFWs issued were in the range of 1.0–3.0 $m^3 s^{-1} km^{-2}$. These observations matched previous findings that identified values for delineating regions of potential flash flooding (Martinaitis et al. 2017). Experimental FFW 2 that covered the Kanawha County, West Virginia, flash flooding was the single warning decision that the CREST maximum unit streamflow product was declared to have no influence on the decision-making process and had values < 1.0 $m^3 s^{-1} km^{-2}$ in the real-time operations period at the time of warning issuance. Some residual product degradation from the earlier network communication difficulties still existed, and the surveyed response might not be representative. The participant who issued experimental FFW 2 commented that the observed values at the time of warning issuance did not play a role in the decision but described how the rising CREST unit streamflow values did increase situational awareness of that particular event.

Postevent evaluations discussed the prospective utility of a rate-of-change field to highlight localized areas of increasing flash flood potential, including the catastrophic flash flood event in northern Kanawha County. Participants noted that situational awareness and multiple warning decisions were influenced by the rapid increase in values of all FLASH products, most notably the CREST maximum unit streamflow product. Example output depicting 10-min differences in the CREST maximum unit streamflow product for the historic Kanawha County event showed an initial decline in overland flow coupled with increased anomalous unit streamflow in larger waterways from prior precipitation (Figs. 26a–d). The additional rainfall over already saturated soils resulted in prolonged unit streamflow increases of 0.1–0.5 $m^3 s^{-1} km^{-2}$ $(10 \text{ min})^{-1}$ from 1950 to 2140 UTC (Figs. 26e–p). Localized areas were modeled having 10-min increases in unit streamflow exceeding 1.0 $m^3 s^{-1} km^{-2}$.

6. Summary and future efforts

The flash flooding of 23 June 2016 presented an opportunity for research scientists and operational forecasters participating

TABLE 12. Analysis of the warned area for the two isolated flash flood events in Kentucky. Listed for the two events over Boyd County (FFW 4) and Owen County (FFW 6) are the FFW issuance times (UTC), the experimental and operational FFW area (km^2), the area difference between the experimental and operational FFWs (km^2), and the percent reduction of the operationally warned area.

Event	Experimental FFW issuance (UTC)	Operational FFW issuance (UTC)	Experimental FFW area (km^2)	Operational FFW area (km^2)	Area difference (km^2)	Percent reduction of warned area
FFW 4	2045	2052	751.5	2656.1	1904.6	71.7%
FFW 6	2108	2033, 2058	599.2	2343.0	1743.8	74.4%

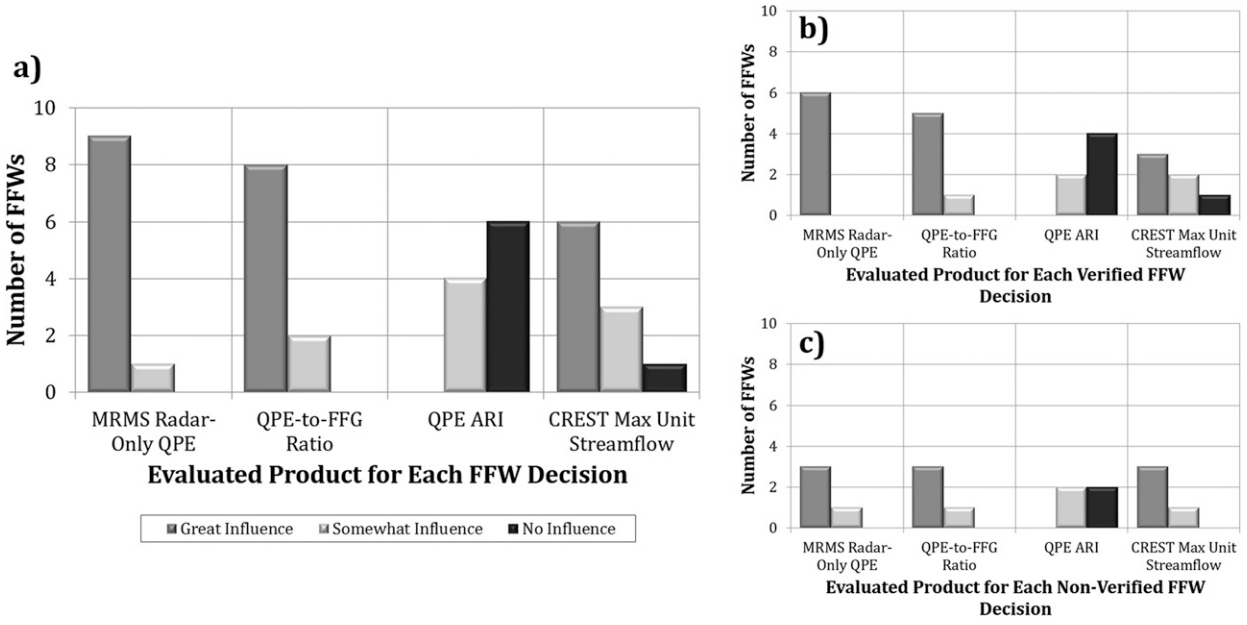

FIG. 25. Influence of the primary MRMS and FLASH products on warning issuance during the HMT-Hydro experiment for (a) all FFWs (i.e., FFW 2–11 from Table 10), (b) only verified FFWs, and (c) only unverified FFWs. Participants rated the influence of each product as "great influence" (gray), "somewhat influence" (light gray), and "no influence" (black).

in two real-time testbed experiments to evaluate experimental products for an historic, catastrophic hydrometeorological event. The FFaIR experiment focused on short-term forecasting through various experimental NWP and hydrologic model guidance. The HMT-Hydro experiment emphasized the prediction and warning of flash flooding in the region. The collaborative efforts between the FFaIR and HMT-Hydro experiments in real time permitted for the simulation of operations and workflow from a national center to a local office environment while assessing the performance of multiple modeling platforms and hydrometeorological tools.

Findings from the FFaIR experiment showed the experimental Day 1 ERO and PFFF for the 23 June 2016 event were highly rated by participants and correlated well with observed flash flood reports and operational FFWs; moreover, the experimental products issued by the FFaIR experiment outlined a greater forecasted risk area. The strengths of the evaluated forecast model guidance varied, yet some models and modeling methodologies captured the orientation of the event with some variances in location and magnitude. Forecasts issued by the FFaIR experiment were critical to the experimental real-time warning operations of the HMT-Hydro experiment. Warning issuance within the defined enhanced threat area began after the FFaIR experiment daily briefing and continued throughout the evolution of the flash flood event. Experimental FFW metrics from the 23 June 2016 event showed an increase in the probability of detection and a 63.8% increase in warning lead time to 84.2 min. Isolated flash flooding in Kentucky revealed the potential ability for significantly reducing warned areas.

Deterministic and ensemble high-resolution convective-allowing models utilized in the FFaIR experiment were shown to be essential to the forecasting process for both the 6–24-h short-term forecast and for longer-term 2- and 3-day forecast periods (Erickson et al. 2019). Probabilistic exceedance information provided by the experimental models were identified as being very useful to probabilistic flash flood forecasting, while work will continue on the skill and application of point-based probabilities versus neighborhood techniques. Model output that supported QPF versus FFG comparisons were desirable among participants, yet QPF comparisons to extreme precipitation information (e.g., ARIs) were shown to be beneficial from a more situational awareness perspective. This was also seen in the HMT-Hydro experiment when comparing QPE to ARIs had a lesser influence on the warning decision-making process but provided situational awareness to define potential threat areas.

Potential benefits of the FLASH product suite for warning issuance were demonstrated during the 23 June 2016 event; moreover, there was increased confidence in the warning decision-making when applying the FLASH hydrologic modeling products when coupled with traditional operational tools and techniques used to predict flash flooding. Future work will investigate methodologies that relate FLASH product values to societal effects and tiered flooding threats. Prototype graphical representations of the temporal changes in CREST maximum unit streamflow values and other hydrometeorological guidance could highlight immediate, significant flash flooding threats embedded within a widespread, ongoing hydrologic event. Detecting significant increases in CREST maximum unit streamflow values within a flood-impacted region could contextualize a potential escalation in flash flood severity and the subsequent use of enhanced wording within a FFW to illicit a greater public response.

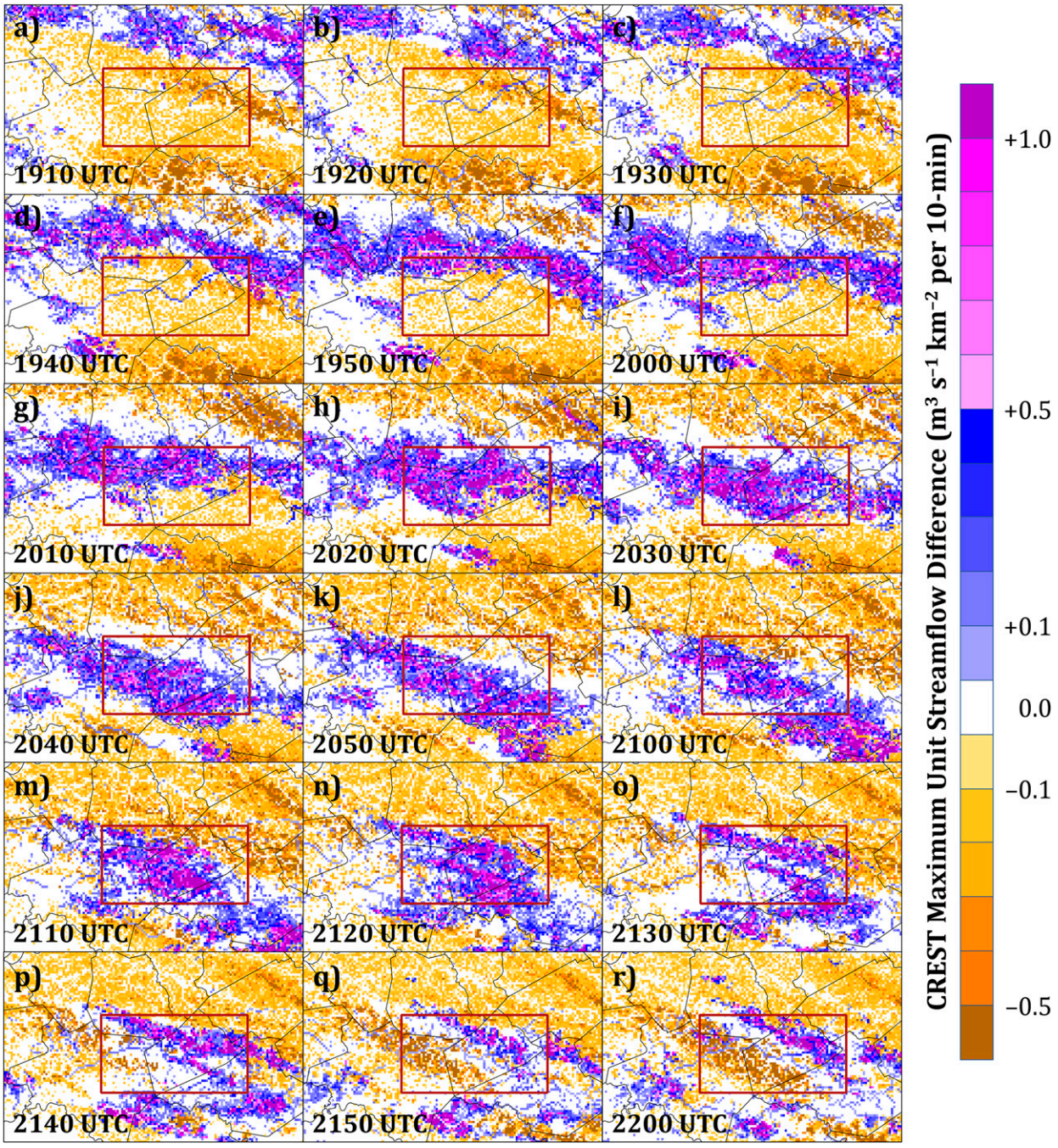

FIG. 26. The 10-min difference of CREST maximum unit streamflow values ($m^3 s^{-1} km^{-2}$) from 1910 to 2200 UTC over west-central West Virginia. The red box denotes the area of greatest impact over northern Kanawha County and Clay County.

Both experiments throughout this event demonstrated the importance of cross-testbed collaborations, continuous product development, and end-user evaluation of next-generation forecast and prediction products for flash flooding across various scales of space and time. The collective use of all experimental models and applications improved confidence with the forecasting of the event and the warning decision-making process. Both experiments also demonstrated the importance of having the fusion of subject-matter experts with forecasters and other end-users in a testbed experiment environment. The discussions between the research and operational groups over the years helped advance the applicational development and operational training of new tools and techniques. This was validated during the two testbed experiment operations of the 23 June 2016 historic flash flood event.

Research scientists with both testbed experiments recognize the importance of probabilistic information and how incorporating predictive uncertainty can shape decision-making processes. New objectives within both testbed experiments look to identify ways that short-term high-resolution ensemble modeling, like the Warn-on-Forecast system (Stensrud et al. 2009, 2013), can benefit national and local hydrometeorological operations with respect to short-term forecasting and increased warning lead time. Long-term planning focuses on the evolution of hydrometeorological products to be encompassed within the probabilities-based and social, behavioral, and economic science-informed Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) paradigm and how the FACETs concepts could perform during a historic hydrometeorological event.

Acknowledgments. The authors thank the participants that contributed in the FFaIR and HMT-Hydro experiments, for this work would not be made possible without their involvement and insight. The authors also thank the anonymous reviewers for their comments, insight, and feedback, as well as Heather Grams (NSSL) for her insight and review of the manuscript. The HMT-Hydro experiment was held within the NOAA Hazardous Weather Testbed, and the authors thank those associated with the facility for the ability to utilize their space, time, and resources. The authors also thank Alexander Zwink (OU/CIMMS–WDTD) for visualization of and access to AWIPS versions of the MRMS and FLASH data. Work related to the FFaIR experiment was supported by NOAA's Hydrometeorology Testbed Program and the U.S. Weather Research Program (USWRP). Funding for the HMT-Hydro experiment was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. Funding for research associated with the HMT-Hydro experiment was also provided by the Disaster Relief Appropriations Act of 2013 (P.L. 113-2), which provided support to the Cooperative Institute for Mesoscale Meteorological Studies at the University of Oklahoma under Grant NA14OAR4830100 and the HMT Program under Grant NA15OAR4590158.

REFERENCES

Alexander, C., S. S. Weygandt, S. G. Benjamin, T. G. Smirnova, J. M. Brown, P. Hofmann, and E. P. James, 2011: The High Resolution Rapid Refresh (HRRR): Recent and future enhancements, time-lagged ensembling, and 2010 forecast evaluation activities. *24th Conf. on Weather and Forecasting/20th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 12B.2, https://ams.confex.com/ams/91Annual/webprogram/Paper183065.html.

Argyle, E. M., J. J. Gourley, Z. L. Flamig, T. Hansen, and K. Manross, 2017: Toward a user-centered design of a weather forecasting decision-support tool. *Bull. Amer. Meteor. Soc.*, **98**, 373–382, https://doi.org/10.1175/BAMS-D-16-0031.1.

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.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire, 1973: A general streamflow simulation system—Conceptual modeling for digital computers. Joint Federal and State River Forecast Center Tech. Rep., U.S. National Weather Service and California Department of Water Resources, 204 pp.

Carley, J. R., and Coauthors, 2015: Ongoing development of the hourly-updated version of the NAM forecast system. *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 2A.1, https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273567.html.

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.

Ebert, E. E., 2001: Analysis of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

——, J. E. Janowiak, and C. Kidd, 2007: Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bull. Amer. Meteor. Soc.*, **88**, 47–64, https://doi.org/10.1175/BAMS-88-1-47.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.*, **108**, 8851, https://doi.org/10.1029/2002JD003296.

Erickson, M. J., J. S. Kastman, B. Albright, S. Perfater, J. A. Nelson, R. S. Schumacher, and G. R. Herman, 2019: Verification results from the 2017 HMT-WPC flash flood and intense rainfall experiment. *J. Appl. Meteor. Climatol.*, **58**, 2591–2604, https://doi.org/10.1175/JAMC-D-19-0097.1.

Fritsch, J. M., and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85**, 955–966, https://doi.org/10.1175/BAMS-85-7-955.

Gochis, D. J., W. Yu, and D. N. Yates, 2015: The WRF-Hydro model technical description and user's guide, version 1.0. NCAR Tech. Doc., 120 pp., https://www.ral.ucar.edu/projects/wrf_hydro.

Gourley, J. J., 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.

Janjić, Z., 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.*, **82**, 271–285, https://doi.org/10.1007/s00703-001-0587-6.

——, J. Gerrity, and S. Nickovic, 2001: An alternative approach to nonhydrostatic modeling. *Mon. Wea. Rev.*, **129**, 1164–1178, https://doi.org/10.1175/1520-0493(2001)129<1164:AAATNM>2.0.CO;2.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC storm-scale ensemble of opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *26th Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., P9.137, https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html.

——, A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the optimal configuration of the High Resolution Ensemble Forecast system. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.

Lewis, W. R., W. J. Steenburgh, T. I. Alcott, and J. J. Rutz, 2017: GEFS precipitation forecasts and the implications of statistical downscaling over the western United States. *Wea. Forecasting*, **32**, 1007–1028, https://doi.org/10.1175/WAF-D-16-0179.1.

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, http://ams.confex.com/ams/pdfpapers/83847.pdf.

Lincoln, W. S., R. F. Thomason, M. Stackhouse, and D. S. Schlotzhauer, 2017: Utilizing crowd-sourced rainfall and flood impact information to improve the analysis of the North Central Gulf Coast flood event of April 2014. *J. Oper. Meteor.*, **5**, 26–41, https://doi.org/10.15191/nwajom.2017.0503.

Maddox, R. A., C. F. Chappell, and L. R. Hoxit, 1979: Synoptic and meso-α-scale aspects of flash flood events. *Bull. Amer. Meteor. Soc.*, **60**, 115–123, https://doi.org/10.1175/1520-0477-60.2.115.

Martinaitis, S. M., and Coauthors, 2017: The HMT Multi-Radar Multi-Sensor Hydro experiment. *Bull. Amer. Meteor. Soc.*, **98**, 347–359, https://doi.org/10.1175/BAMS-D-15-00283.1.

Nietfeld, D., 2013: Fostering R2O and O2R with a unique AWIPS2 testbed in NWS-WFO Omaha. *Third Conf. on Transition of Research to Operations*, Austin, TX, Amer. Meteor. Soc., 6.1, https://ams.confex.com/ams/93Annual/webprogram/Paper222766.html.

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

Rogers, E., and Coauthors, 2009: The NCEP North American Mesoscale modeling system: Recent changes and future plans. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 2A.4, https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154114.htm.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A proposed next-generation paradigm for high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note NCAR/TN-468+STR, 88 pp., https://doi.org/10.5065/D6DZ069T.

——, and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Snook, N., F. Kong, K. A. Brewster, M. Xue, K. W. Thomas, T. A. Supinie, S. Perfater, and B. Albright, 2019: Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016–17 NOAA hydrometeorology testbed flash flood and intense rainfall experiments. *Wea. Forecasting*, **34**, 781–804, https://doi.org/10.1175/WAF-D-18-0155.1.

Stensrud, D., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

——, and Coauthors, 2013: Progress and challenges with warn-on-forecast. *Atmos. Res.*, **123**, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme quantitative precipitation forecast performance at the Weather Prediction Center from 2001 to 2011. *Wea. Forecasting*, **29**, 894–911, https://doi.org/10.1175/WAF-D-13-00061.1.

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.

Viterbo, F., and Coauthors, 2020: A multiscale, hydrometeorological forecast evaluation of national water model forecasts of the May 2018 Ellicott City, Maryland, flood. *J. Hydrometeor.*, **21**, 475–499, https://doi.org/10.1175/JHM-D-19-0125.1.

Wang, J., and Coauthors, 2011: The Coupled Routing and Excess Storage (CREST) distributed hydrological model. *Hydrol. Sci. J.*, **56**, 84–98, https://doi.org/10.1080/02626667.2010.543087.

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.