Assessing the Skill of Medium-Range Ensemble Precipitation and Streamflow Forecasts from the Hydrologic Ensemble Forecast Service (HEFS) for the Upper Trinity River Basin in North Texas

SUNGHEE KIM, HOSSEIN SADEGHI,[a] REZA AHMAD LIMON, MANABENDRA SAHARIA,[b] AND DONG-JUN SEO

Department of Civil Engineering, The University of Texas at Arlington, Arlington, Texas

ANDREW PHILPOTT AND FRANK BELL

West Gulf River Forecast Center, NOAA/NWS, Fort Worth, Texas

JAMES BROWN

Hydrologic Solutions Limited, Southampton, United Kingdom

MINXUE HE

Hydrology Branch, California Department of Water Resources, Sacramento, California

(Manuscript received 7 February 2018, in final form 28 June 2018)

ABSTRACT

To issue early warnings for the public to act, for emergency managers to take preventive actions, and for water managers to operate their systems cost-effectively, it is necessary to maximize the time horizon over which streamflow forecasts are skillful. In this work, we assess the value of medium-range ensemble precipitation forecasts generated with the Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service (NWS) in increasing the lead time and skill of streamflow forecasts for five headwater basins in the upper Trinity River basin in north-central Texas. The HEFS uses ensemble mean precipitation forecasts from the Global Ensemble Forecast System (GEFS) of the National Centers for Environment Prediction (NCEP). For comparative evaluation, we verify ensemble streamflow forecasts generated with the HEFS forced by the GEFS forecast with those forced by the short-range quantitative precipitation forecasts (QPFs) from the NWS West Gulf River Forecast Center (WGRFC) based on guidance from the NCEP's Weather Prediction Center. We also assess the benefits of postprocessing the raw ensemble streamflow forecasts and evaluate the impact of selected parameters within the HEFS on forecast quality. The results show that the use of medium-range precipitation forecasts from the GEFS with the HEFS extends the time horizon for skillful forecasting of mean daily streamflow by 1–3 days for significant events when compared with using only the 72-h River Forecast Center (RFC) QPF with the HEFS. The HEFS forced by the GEFS also improves the skill of two-week-ahead biweekly streamflow forecast by about 20% over climatological forecast for the largest 1% of the observed biweekly flow.

1. Introduction

Accurate forecasting of river flow is not only important for flood prediction, but also for a range of applications associated with design, operation, and management of water resources infrastructure. To issue early warnings for the public to act, for emergency managers to take preventive actions, and for water managers to operate reservoirs and other systems effectively, it is necessary

[a] Current affiliation: Bannister Engineering, LLC, Mansfield, Texas.

[b] Current affiliation: National Center for Atmospheric Research, Boulder, Colorado.

Corresponding author: Reza Ahmad Limon, rezaahmad.limon@mavs.uta.edu

DOI: 10.1175/JHM-D-18-0027.1

to maximize the forecast lead time while properly accounting for the forecast uncertainties. In addition to short-range forecasting (~1–3 days), medium-range forecasting (~4–7 days) of streamflow is critical to meeting a variety of needs in operational hydrology and water resources management (Yuan et al. 2014). Whereas some users in the eastern United States may be interested in river forecasts with lead times of 3–7 days to manage and mitigate the potential impacts of flooding (Adams and Ostrowski 2010), those in the western United States may be interested in weekly or longer-period forecasts of inflow into water supply reservoirs (Georgakakos et al. 2006).

Skillful medium-range forecasting of precipitation and streamflow is particularly important in areas prone to extreme events such as floods and droughts. For example, in Texas, a severe drought which lasted for four and a half years since 2011 ended with extreme flooding from record-breaking rainfall in May 2015 resulting in at least 28 fatalities (Di Liberto 2015). In such situations, skillful precipitation and streamflow forecasting can, with sufficient warning, mitigate downstream flooding by allowing for preemptive releases of water from the reservoirs and enable more cost-effective management of water supply and treatment systems such as those operated by the Tarrant Regional Water District, the Trinity River Authority, and others in north-central Texas.

Currently, the National Weather Service (NWS) West Gulf River Forecast Center (WGRFC) in Fort Worth, Texas, uses short-range quantitative precipitation forecasts (QPFs) to produce operational river forecasts. The QPF is single-valued, or deterministic, and may comprise a forecast horizon of up to 168 h based on the National Centers for Environmental Prediction (NCEP) Weather Prediction Center's (WPC) guidance up to 72 h and the Global Forecast System (GFS) output thereafter. In practice, the forecast horizon is typically limited to 24 h or less with no precipitation assumed thereafter. Depending on the specific weather events, however, the entire 168-h forecast horizon may be used for contingency forecasts (WGRFC 2015). The above practice of limiting the period of nonzero QPF in single-valued streamflow forecasting stems from the limited predictive skill in single-valued QPF, particularly for convective events. In the southern plains of the United States, the use of single-valued QPF is likely to produce single-valued streamflow forecasts with unacceptably large errors beyond the first 24 h of lead time (Regonda et al. 2013). With such limited predictability, it is not possible, without risking credibility, to issue skillful single-valued streamflow forecasts consistently beyond the forecast horizon of the sum of the lead time of the skillful single-valued QPF and the hydrologic response time of the catchment.

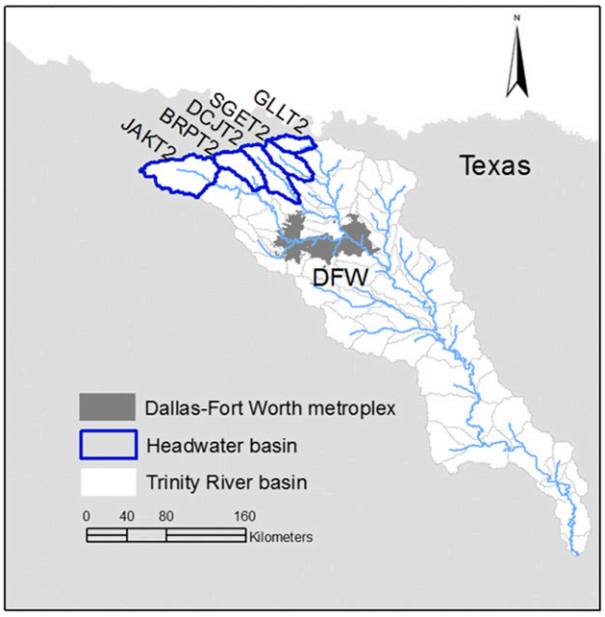

FIG. 1. Five headwater catchments in the upper Trinity River basin in north-central Texas: Jacksboro (JAKT2), Big Sandy Creek near Bridgeport (BRPT2), Denton Creek near Justin (DCJT2), Clear Creek near Sanger (SGET2), and Elm Fork of the Trinity River near Gainesville (GLLT2).

Precipitation forecasts of longer accumulations (3 days or longer), on the other hand, are significantly more skillful than those of shorter accumulations (daily or shorter). This is because for longer accumulations it is not necessary to predict accurately the granular temporal distributions of precipitation (Brown et al. 2014a). Even with larger skill in longer accumulations, however, precipitation forecasts are in general too uncertain for deterministic hydrologic forecasting, that is, as a single-valued input to hydrologic models. If precipitation forecasts are expressed as ensembles or in probabilistic terms, on the other hand, one may produce ensemble or probabilistic hydrologic forecasts that reflect skill present over the entire forecast horizon with which the users may make risk-based decisions (Hartman et al. 2007; Demargne et al. 2014; Seo et al. 2010).

The purpose of this work is to assess the skill of ensemble precipitation and streamflow forecasts produced with the NWS Hydrologic Ensemble Forecast Service (HEFS; Demargne et al. 2014) using precipitation forcing from the Global Ensemble Forecast System (GEFS) for improving the quality and lead time of streamflow forecasts in north-central Texas. The study area consists of five headwater basins located upstream of the DFW area in the upper Trinity River basin in north-central Texas (see Fig. 1). It is expected that, by utilizing the skill present in medium-range QPF at lead times beyond the current

maximum at WGRFC of 3 days, one may extend significantly the lead time of skillful hydrologic forecasts, in particular, of streamflow and soil moisture. It should be noted that these basins offer an extremely challenging test for the HEFS in that precipitation is dominated by convection and hence has very limited predictability, and that the basins are flashy with extreme variability including prolonged periods of little to no streamflow.

The HEFS includes the Meteorological Ensemble Forecast Processor (MEFP; Schaake et al. 2007; Wu et al. 2008) and the streamflow ensemble postprocessor (EnsPost; Seo et al. 2006), the two statistical models that quantify the meteorological input and hydrologic uncertainties, respectively. Both models generate ensembles via conditional stochastic simulation. The MEFP (NWS 2017a) inputs ensemble mean or single-valued forcings of precipitation and temperature and generates precipitation and temperature ensembles which are used to force the NWS hydrologic models and produce "raw" streamflow ensembles. The EnsPost (NWS 2017b) corrects for biases in the raw streamflow ensembles and models the total hydrologic uncertainty. Saharia (2013) applied the HEFS to five headwater basins in the upper Trinity River basin in north-central Texas (see Fig. 1). He found that the short-range ensemble QPFs generated with the MEFP forced by the WGRFC single-valued QPFs, referred to hereafter as the MEFP-RFC precipitation ensembles, were in general both reliable and skillful in keeping with similar studies in other areas (Wu et al. 2011; Brown et al. 2014b). He also found that, in comparison with using day 1 QPF only, using day 1–3 single-valued QPF significantly increased the skill in short-range ensemble streamflow forecast. His work also showed that the addition of day 2–3 QPF increased the probability of detection (PoD) of the 95th percentile flow by about 10% for day 3–4 streamflow prediction, extending the useful lead time by about a day. It was also found that, for high streamflow thresholds, the addition of day 2–3 QPF was more important than streamflow postprocessing, as high flows sensitively depend on the quality of the precipitation forcing.

In this work, we extend the above study and assess the skill of medium-range precipitation forecasts in improving the quality and lead time of streamflow forecasts in north Texas. The precipitation forecasts are generated with forcing inputs from the GEFS using the MEFP, referred to hereafter as the MEFP-GEFS precipitation ensembles, which are compared with the MEFP-RFC precipitation ensembles. We then assess the skill of the MEFP-GEFS precipitation and streamflow ensembles, referred to collectively as the MEFP-GEFS ensembles, for multiday accumulation periods of up to 30 days, the impact of streamflow postprocessing with the EnsPost, and the impact of selected parameters within the MEFP and EnsPost on the quality of the MEFP-GEFS ensembles. Verification is carried out with the Ensemble Verification System (EVS; Brown et al. 2010; Brown 2015a) for a large sample of retrospective forecasts, or hindcasts, produced with the HEFS.

The new and significant contributions of this paper are as follows: 1) comparative verification of the MEFP-GEFS ensembles with the MEFP-RFC ensembles, 2) verification of the multiday MEFP-GEFS ensembles, 3) assessment of the impact of the EnsPost, and 4) assessment of the impact of the key parameters in the MEFP and EnsPost on the quality of the MEFP-GEFS ensembles for north-central Texas. This paper is organized as follows. Section 2 describes the methods, including the hydrologic models, study area, and data used; parameter estimation; and hindcasting and verification. Section 3 describes the results, including the impact of different parameter estimation options, precipitation results, and streamflow results. Section 4 provides the conclusions and future research recommendations.

2. Methods

In this section, we describe the hydrologic models, study area, and data used; parameter estimation; and hindcasting and verification.

a. Hydrologic models, study area, and data used

The HEFS can utilize any hydrologic models available within the NWS Community Hydrologic Prediction System (CHPS; Roe et al. 2010). In this work, we used the Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash 1995) and unit hydrograph (UH; Chow et al. 1988) for the five headwater basins in the study area (see Fig. 1). The above models are currently used operationally at the WGRFC, and hence the results presented in this paper represent what may be expected from the HEFS in the NWS operations today.

Recently, the NWS has implemented the National Water Model (NWM) over the continental United States (Graziano et al. 2017). It is expected, however, that the primary guidance for the NWS's flood watches and warnings will continue to come from the lumped models such as the SAC-SMA and UH run at the RFCs within the foreseeable future as explained below. Whereas the lumped models have been extensively calibrated over the years, the NWM is yet to undergo systematic calibration. Also, the precipitation forcings used at the RFCs are value-added by human forecasters (Nelson et al. 2016), whereas those used for the NWM are not. Last, the RFC forecasters perform extensive

TABLE 1. Characteristics of the study basins in the upper Trinity River basin.

Characteristics	JAKT2	BRPT2	DCJT2	SGET2	GLLT2
Latitude (outlet)	33.29	33.23	33.12	33.34	33.62
Longitude (outlet)	−98.08	−97.69	−97.29	−97.18	−97.15
Area (km^2)	1769.00	862.47	1036.00	764.05	450.66
Mean annual precipitation (mm)	931.1	980.8	1026.9	1076.9	1083.1
Average streamflow ($m^3 s^{-1}$ or cms)	2.48	1.46	2.97	2.92	2.67
Runoff ratio (%)	4.8	5.5	8.8	11.2	17.3
Mean elevation (m)	279	229	197	193	227
Time to peak (h)	24	24	12	12	12

manual data assimilation (DA) to keep the model states in line with reality (Seo et al. 2009), whereas the NWM's DA capability is currently limited to nudging, which operates more as a postprocessor than DA (D. Gochis et al. 2017, workshop presentation). One may anticipate that, as forecasters gain more experience with the NWM output and the forecast quality improves through improved forcings, calibration, and DA (Cosgrove et al. 2017), the RFCs will practice some form of multimodel ensemble forecasting for gauged locations (Georgakakos et al. 2004). In such a scenario, one may envision the HEFS evolving to support both the existing lumped model-based forecasting and the NWM.

The study area comprises the five headwater catchments in the upper Trinity River basin upstream of the Dallas–Fort Worth (DFW) metroplex (see Fig. 1). From the drier west to the wetter east, the catchments drain to the West Fork of the Trinity River near Jacksboro (JAKT2), Big Sandy Creek near Bridgeport (BRPT2), Denton Creek near Justin (DCJT2), and Clear Creek near Sanger (SGET2) and the Elm Fork of the Trinity River near Gainesville (GLLT2). Figure 1 and Table 1 show the locations and the physiographic and fluvial characteristics of the basins, respectively. Figures 2a and 2b show the mean daily precipitation and streamflow for the five catchments. Figure 2a and Table 1 show increasing mean daily precipitation and runoff ratio from west to east. In Fig. 2b, the substantially reduced streamflow in the fall wet season compared to that in the spring wet season is due to the fact that the very dry summer tends to deplete soil moisture [see section 3c(2)]. The DFW area is the largest inland population center and one of the fastest growing urban areas in the United States. This region is vulnerable to the impacts of urbanization and climate change on water sustainability due to the warmer climate conditions, rapid land conversion, high degree of impervious surface, and dependence on surface water. According to the Texas Water Development Board (TWDB 2015), more than 95% of the water used in the upper Trinity River basin is surface water. As such, skillful forecasting of precipitation and streamflow for these and other headwater

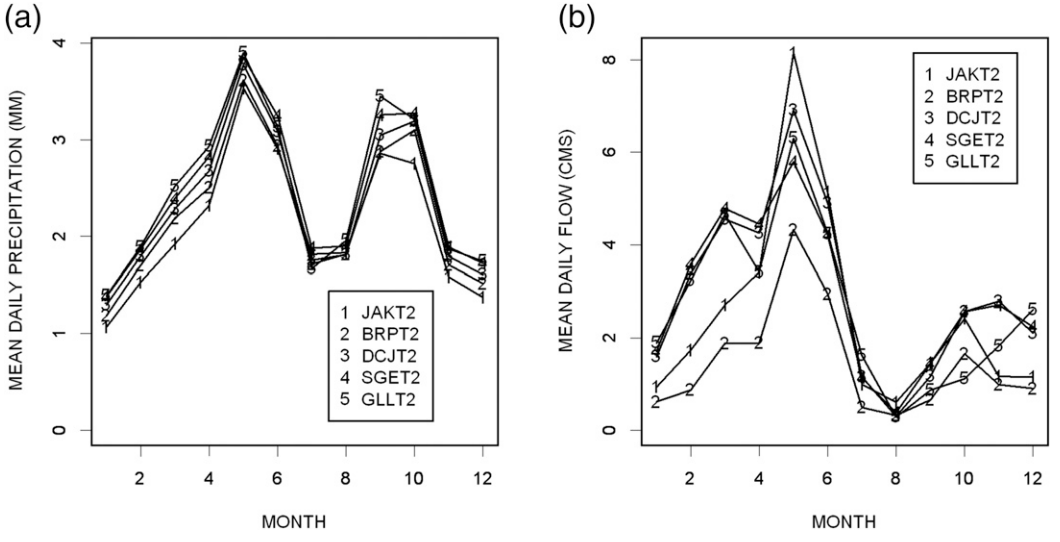

FIG. 2. (a) Mean daily precipitation (mm) for the five catchments. (b) Mean daily streamflow ($m^3 s^{-1}$ or cms) for the five catchments.

TABLE 2. Datasets used.

Name	Period of record	Description	Source
RFC QPF	Jan 2005 to Dec 2014	6-hourly single-valued precipitation forecast	WGRFC
MAP	Oct 1959 to Dec 2015	6-hourly observed mean areal precipitation	WGRFC
QME	Oct 1959 to Dec 2015	Observed mean daily streamflow	USGS via WGRFC
GEFS	Jan 1985 to Dec 2015	Ensemble mean precipitation forecast from GEFS	NWS
SQIN	Oct 1959 to Dec 2015	Simulated streamflow at 6-h interval	WGRFC

basins is particularly important for flood warning, water supply, reservoir operations, water quality management, and other applications.

Table 2 shows the data used to generate precipitation and streamflow hindcasts for the five headwater basins. The historical mean areal precipitation (MAP) time series, the historical RFC QPF, and the GEFS reforecast dataset are used to estimate the MEFP parameters and to generate ensemble precipitation hindcasts from the MEFP. The observed mean daily flow (QME) and the simulated mean daily flow derived from the simulated instantaneous flow (SQIN) at a 6-h interval are used to estimate the EnsPost parameters. The GEFS hindcasts comprise 6-hourly precipitation amounts, issued at 0000 UTC for a forecast horizon of 1–16 days (Hamill et al. 2013). Because the hydrologic forecasts are issued at 1200 UTC each day, the first 12 h of the GEFS forecast horizon is curtailed and the precipitation reforecasts are hence available only up to 15 days into the future for streamflow hindcasting.

b. *Parameter estimation*

The MEFP and EnsPost both employ statistical parameters whose values must be estimated from the historical data. The MEFP Parameter Estimator (MEFPPE) models the input uncertainty in forecast precipitation and produces the MEFP parameters. The EnsPost Parameter Estimator (EnsPostPE) models the hydrologic uncertainty in model-simulated streamflow and produces the EnsPost parameters. The quality of the ensembles produced by the MEFP and EnsPost depends very significantly on the quality of their parameters. It is therefore very important that the parameters are estimated carefully to maximize the skill in the ensemble precipitation and streamflow forecasts.

The GEFS already produces ensemble forecasts of precipitation and temperature along with many other variables (Hamill et al. 2013). Such "raw" ensemble forecasts are, however, generally biased in the mean, spread, and higher-order moments. In addition, the raw forecast probabilities cannot currently fully capture the regime-dependent forecast uncertainties (Wu et al. 2011). Accordingly, it is generally necessary to remove or reduce biases in the raw ensemble forecasts by statistical means. To bias-correct QPFs and to model the uncertainties associated with them statistically, considerable efforts have been made in recent years (Gneiting et al. 2007; Hamill et al. 2008, 2013; Scheuerer and Hamill 2015). To estimate reliably the parameters of statistical processors such as the MEFP, it is generally necessary to have historical forecasts and verifying observations over a long period.

In this work, we use the GEFSv10 (Zhou et al. 2017), which provides retrospective forecasts over a long period to support statistical postprocessing. Even with the large-sample hindcast dataset, the available sample size for extreme precipitation events for the specific season and location of interest may be too small for reliable estimation of the MEFP parameters. To increase the sample size, the MEFPPE pools all pairs of forecast and observed MAP within the user-specified time window. The window is centered on each Julian day so that the regression parameters may capture the seasonal variations. In this estimation process, there is a trade-off to consider between the sampling uncertainty of the MEFP parameters (larger window preferred) and their specificity in capturing the seasonal variations (smaller window preferred).

The MEFP can use multiple sources of forcing forecasts over different time horizons to produce bias-corrected forcing ensembles that are consistent from short to long ranges (see Table 3). To utilize in the above process all available skills present over the entire forecast horizon, the MEFP employs the so-called canonical events (CEs), which consist of base and modulation events (Collischonn et al. 2007; NWS 2017a; Roundy et al. 2015). Though named "events," the canonical events are predefined time windows of varying length over the forecast horizon. For each event, a regression model is constructed in the bivariate normal space (Brown 2015b). Once the model parameters are estimated for all events, they are ranked according to the strength of correlation.

There are two types of canonical events, base and modulation. The base events have an aggregation scale of 6 h through the first week of the forecast horizon and have larger time windows of 1 day, 2 days, etc., beyond the first week (see Fig. 3). There are two types of base events, fine and coarse. The fine base events consist of

TABLE 3. Forecasts used in MEFPPE (NWS 2017b).

Forecast horizon	Forecast	Source
Short range	Single-valued QPF (~3 days)	RFC, NCEP WPC
	Single-valued QPF (~5 days)	NCEP WPC
Medium range	Ensemble mean from GEFS (~15 days)	NCEP EMC
Long range	Time-lagged ensemble mean from CFSv2 (~9 months)	NCEP EMC
	Climatology (~1 year)	Historical observations

6-hourly time windows up to the first 120 h in the forecast horizon and 12-hourly or larger windows beyond the 120 h. The coarse base events consist of two 6-hourly time windows up to 12 h of forecast horizon, 12-hourly time windows between 12 and 120 h, and 24-hourly or larger windows beyond the 120 h. By design, these events, or time windows, do not overlap with one another. The modulation events are defined at time scales larger than (e.g., integer multiples of) the base events and may overlap with the base events (see Fig. 3). The purpose of the modulation events is to capture the joint distribution between the forecasts and observations at multiple temporal scales of aggregation. For example, with a 6-h scale alone, it is difficult to utilize skill that may be present at larger scales due to the high dimensionality of the multivariate probability distribution (Collischonn et al. 2007). One may therefore view the combined use of the base and modulation events in the MEFP as a form of multiscale nonlinear regression. A similar approach has also been used in postprocessing of raw streamflow ensembles using multiscale bias correction (S. K. Regonda and D.-J. Seo 2008, poster presentation).

In the conditional simulation process of the MEFP, the regression models are run for all canonical events in the ascending order of the strength of correlation. The base events typically have time windows that tend to increase with lead time whereas the modulation events have time windows that are aggregates of the base events. For operational use, the canonical events should be defined according to the weather and climate patterns of the forecast region. As such, hindcasting and verification studies are generally necessary to determine their optimal specification (NWS 2017a).

The purpose of the EnsPost is to correct for biases in streamflow simulation, that is, streamflow modeled with observed, as opposed to forecast, forcing, and to account for the total hydrologic uncertainty therein. The EnsPost uses an autoregressive-1 model with a single exogenous variable, or ARX(1,1), in the bivariate normal space (Seo et al. 2006). To account for seasonality, the EnsPostPE supports estimation of the EnsPost parameters at different time scales, such as monthly, seasonal (spring, summer, fall, and winter), semiannual (wet and dry), or annual. The EnsPost parameters are estimated with historical pairs of simulated and observed streamflow on a user-defined time scale. As the seasonal scale increases, the sample size increases but potentially at the expense of not being able to capture the seasonal variations in streamflow. As with the sampling window in the MEFPPE, there is a trade-off to consider between the sampling uncertainty of the EnsPost parameters and their specificity in capturing the seasonal variations. In this work, we assess the impact of the choices of the sampling window and the canonical events in the MEFP and the time scale of seasonal stratification in the EnsPost on the quality of the MEFP-GEFS ensembles.

c. *Hindcasting and verification*

To assess the comparative skill of medium-range ensemble precipitation and streamflow forecasts, we designed and carried out a set of hindcasting experiments using the HEFS as depicted in Fig. 4. The MEFP-RFC and MEFP-GEFS ensembles were generated every day for the 10- and 31-yr periods of 2005–14 and 1985–2015, respectively. In this process, each hindcast is reinitialized every day in the hindcast horizon with the soil moisture states valid for that day as obtained from the SAC-SMA forced by MAP and climatological mean areal potential evapotranspiration following a warmup period of 1960–84. Whereas the GEFS reforecast is available from 1985, the RFC QPF has been archived only since 2005. The above 10-yr period hence represents the largest common period of record between the two forcing QPF datasets. For reference climatological forecasts, we generated the so-called resampled climatological ensembles of precipitation and streamflow by using climatological ensemble mean as the forcing input for the MEFP and using the resulting climatological precipitation ensembles to generate streamflow ensembles (Brown et al. 2014a). The ensemble forecasts comprise 55 ensemble members corresponding to the number of historical years for observed precipitation for the Schaake Shuffle (Clark et al. 2004) used in the MEFP.

To assess the impact of selected MEFP and EnsPost parameters on forecast quality, we examined the skill of the ensemble hindcasts generated by the MEFP and EnsPost using the six different sets of the MEFPPE- and

Lead time (hour)	CE1*	CE2						CE3*	CE4		
	Base	Base	Modulation					Base	Base	Modulation	
6	1	1	1	4	7	10	21	1	1	1	4
12	2	2						2	2		
18	3	3						3	3		
24		4						4	4		
30	4	5	2					5	5	2	
36		6						6	6		
42	5	7						7	7		
48		8						8	8		
54	6	9	3					9	9	3	
60		10						10	10		
66	7	11						11	11		
72		12						12	12		
78	8	13	5	9				13	13	5	9
84		14						14	14		
90	9	15						15	15		
96		16						16	16		
102	10	17	6					17	17	6	
108		18						18	18		
114	11	19						19	19		
120		20						20	20		
126	12	21	8		14			21	21	7	
132											
138											
144											
150	13	22		12				22	22		10
156											
162											
168											
174	14	23	11			20		23	23	9	
180											
186											
192											
198	15	24						24	24		
204											
210											
216											
222		25	13	16				25	25	11	13
228											
234											
240											
246	16	26			19			26	26		
252											
258											
264											
270		27	15	18				27	27	12	15
276											
282											
288											
294	17	28						28	28		
300											
306											
312											
318		29	17					29	29	14	
324											
330											
336											
342		30						30	30		
348											
354											
360											

* Include no modulation events.

FIG. 3. Definition of canonical events used.

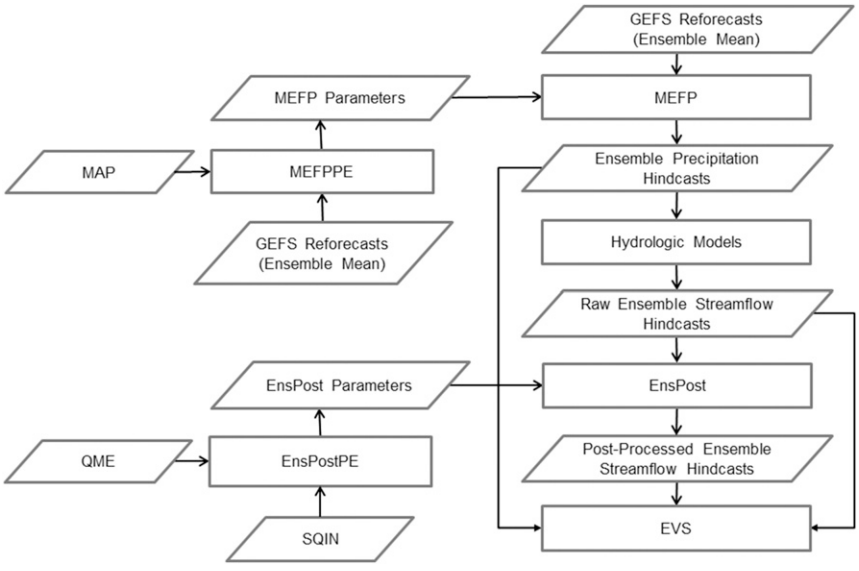

FIG. 4. Ensemble hindcasting and verification process using the HEFS.

EnsPostPE-estimated parameters (see Table 4). The parameters examined are the sampling window and canonical events for the MEFPPE and the time scale of seasonal stratification for the EnsPostPE. We then carried out hindcasting experiments using the six cases and verified the resulting ensemble precipitation and streamflow hindcasts. For comparative verification of the MEFP-GEFS ensembles versus the MEFP-RFC, the period of record available is only 10 years. To reduce sampling uncertainty, we pooled the hindcasts for the five study basins. Such pooling is not a significant issue for precipitation because the basins share very similar MAP climatology (see Table 1, Fig. 2a). For streamflow, however, variations in catchment size, physiography, anthropogenic effects, and the quality of modeling may produce significantly different results for different catchments. If the model simulation is particularly poor for some basin due to, for example, large timing errors, it is likely to skew the pooled results. As such, we examined the above attributes and the catchment-specific verification results to assess their comparability for pooling. Because the verification is carried out for different percentile-based thresholds of the observed flow, it is particularly important to examine how the thresholds for the pooled results may compare with those for the individual basins.

Figure 5 shows the empirical cumulative distribution functions (CDFs) of observed mean daily flow for the five catchments. For clarity, only the tails above the probability levels less than or equal to 90% are shown. Note that the CDFs are very similar among the four basins of DCJT2, GLLT2, JAKT2, and SGET2 but the CDF for BRPT2 is significantly different. The pooled CDF (in black) shows that the percentile thresholds based on pooling are representative of the four basins but not of BRPT2. Because the streamflow at BRPT2 is smaller than the flows at other catchments for the same level of exceedance probability (see also Fig. 2b), the pooled verification results reflect the BRPT2 ensembles at higher thresholds than its own. The consequence is that the pooled results underrepresent the skill in the BRPT2 ensembles, and that the marginal gains are likely to be slightly underestimated in the quality of streamflow ensembles due to the MEFP-GEFS versus the MEFP-RFC and due to the EnsPost versus without the EnsPost.

The resulting large-sample ensemble precipitation and raw and postprocessed streamflow hindcasts were verified using the EVS (Brown et al. 2010). The EVS includes a comprehensive set of metrics for verification

TABLE 4. List of cases examined.

	MEFPPE		EnsPostPE
Case No.	CE No.[a]	Sampling window (days)	Seasonal stratification
1	1	61[b]	Monthly
2	1	61[b]	Semiannual[c]
3	1	91	Monthly
4	2	91	Monthly
5	3	91	Monthly
6	4	91	Monthly

[a] See Fig. 3 for definition of CEs.
[b] Default value recommended by NWS.
[c] Wet season: Mar, Apr, May, Jun, Sep, and Oct; dry season: Jan, Feb, Jul, Aug, Nov, and Dec.

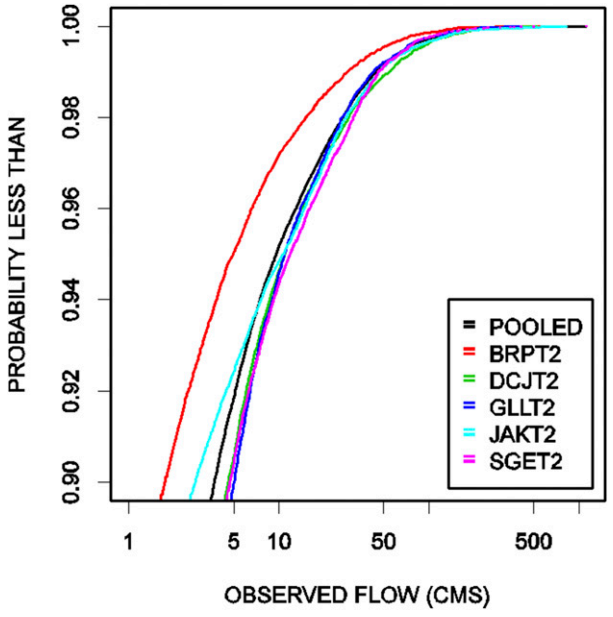

FIG. 5. Empirical CDFs (upper tail only) of observed mean daily flow for the five catchments.

of both single-valued and ensemble forecasts. To verify ensemble forecasts, we used the reliability diagram, the mean continuous ranked probability score (CRPS), and the relative operating characteristic (ROC), among others, to examine reliability, overall skill and discrimination (Jolliffe and Stephenson 2003; Wilks 2006), respectively. The area under the curve (AUC) represents the area under the ROC curve as calculated via direct integration of the empirical ROC curves (Green and Swets 1966). The ROC measures the ability of a forecasting system to correctly predict the occurrence of an event, expressed as the PoD, while avoiding too many incorrect forecasts when it does not occur, expressed as the probability of false detection (PoFD; Mason and Graham 2002; Wilks 2006). Hence, the ROC measures forecast's ability to discriminate an event as defined by the user from a nonevent (Demargne et al. 2010). For a particular exceedance probability threshold d, the empirical PoD and PoFD are given by

$$\text{PoD} = \frac{\sum_{i=0}^{n} I_{X_i}\left[F_{X_i}(q) > d | Y_i > q\right]}{\sum_{i=0}^{n} I_{Y_i}(Y_i > q)} \quad \text{and} \qquad (1)$$

$$\text{PoFD} = \frac{\sum_{i=0}^{n} I_{X_i}\left[F_{X_i}(q) > d | Y_i \le q\right]}{\sum_{i=0}^{n} I_{Y_i}(Y_i \le q)}. \qquad (2)$$

where n denotes the number of pairs of the probabilistic forecast Y_i and the verifying observation X_i; $I(\,)$ denotes the indicator function of the variable subscripted which maps to unity if the outcome parenthesized is realized and to zero if not; q denotes the conditioning threshold for the variable being verified; and $F_{X_i}(q)$, $F_{Y_i}(q)$ denote the ith observed and forecast probabilities that the variable being verified exceeds q, respectively. The ROC plots the PoD versus the PoFD for all possible values of the threshold d in [0, 1]. The ROC for a perfect forecast connects (0, 0), (0, 1), and (1, 1) on the PoFD–PoD plane, and that for a skill-less forecast connects (0, 0) and (1, 1). The AUC for a perfect forecast is hence 1 and that for a skill-less forecast is 0.5.

Because the ROC is made of the PoD and PoFD, it is possible to compare directly the PoD among multiple ensemble forecasts at a user-defined level of the PoFD as well as between the ensemble and single-valued forecasts at the level of the PoFD of the single-valued forecast. In this work, we translate the increase or decrease in the ROC score (RS), which is related to the AUC via RS = (AUC − 0.5)/0.5, to an increase or decrease in the PoD given the user-acceptable level of the PoFD. Such translation allows for a straightforward comparison between single-valued and ensemble forecasts and therefore provides an effective means to communicating with water managers on the use of ensemble forecasts. Some decision-makers, including forecasters and emergency managers, may prefer a lower PoFD at the expense of a lower PoD whereas others may prefer a higher PoD even if it may increase the PoFD.

The CRPS represents the integral squared difference between the CDF of the predicted variable $F_Y(q)$, and that of the verifying observed variable $F_X(q)$ (i.e., a step function):

$$\text{CRPS} = \int [F_Y(q) - F_X(q)]^2 \, dq. \qquad (3)$$

The mean CRPS reflects the overall quality of the probabilistic forecast (the smaller the mean CRPS is, the better) and, similarly to the Brier score (Wilks 2006), is decomposed into reliability, resolution, and uncertainty (Hersbach 2000). The mean continuous ranked probability skill score (CRPSS) measures this skill relative to climatology (1 means perfect, 0 means skill-less):

$$\overline{\text{CRPSS}} = \frac{\overline{\text{CRPS}}_{\text{clim}} - \overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{clim}}}. \qquad (4)$$

3. Results

In this section, we present the results in three parts: 1) the impact of different parameter options, 2) the

precipitation results, and 3) the streamflow results. The results focus on comparative verification of the MEFP-GEFS ensembles relative to the MEFP-RFC ensembles, and verification of the MEFP-GEFS ensembles over a range of temporal scales of aggregation. Resampled climatology was used beyond day 3 for the MEFP-RFC ensembles and beyond Day 15 for the MEFP-GEFS ensembles, respectively. For comparative verification of the MEFP-GEFS ensembles versus the MEFP-RFC, we pooled the hindcasts for the five study basins to reduce sampling uncertainty. For verification of the MEFP-GEFS ensembles, the sample size is much larger and hence we present selected catchment-specific results as well. Because of space limitations, it is not possible to present results for different thresholds of precipitation and streamflow. For the main results, we focus on the 99th percentiles of the verifying observed precipitation or streamflow which represent the largest thresholds before sampling uncertainty makes interpretation difficult. The above thresholds are of the largest impact and hence interest for water management in the study area and offer a rather challenging test for the HEFS given the limited hydrometeorological and hydrologic predictability in the region.

a. Impact of different parameter estimation options

To arrive at the MEFP and EnsPost parameters used in the hindcasting experiments, we assessed the impact of different parameter estimation options (see Table 4) in the MEFPPE and EnsPostPE. Here we present the precipitation and streamflow results together so that one may easily assess the impact of any changes in the skill of precipitation ensembles on that of streamflow ensembles. The assessment is based on the 31-yr MEFP-GEFS hindcasts.

Comparison between the 61- and 91-day sampling windows in MEFPPE, case 1 versus case 3, indicates that the differences are negligible in precipitation or streamflow hindcasts with or without the EnsPost. The above lack of sensitivity suggests that the 31-yr period of the GEFS record is sufficiently long for calibration of the MEFP with a sampling window of 61 days, which is the HEFS default.

The verification results for the different combinations of the canonical events show that the combination of coarse base events and no modulation events (case 3) improves the mean CRPSS by about 5% for day 1 and 10% for days 2–5 over the combination of fine base events and no modulation events (case 5). No gain was observed for days 6–8 because the temporal aggregation scheme in the canonical event definitions is the same over this part of the forecast horizon (see Fig. 3). From day 9, however, the gain reappears due to the larger temporal aggregation in the canonical event definitions used in the coarse event set (48 vs 24 h). The above findings indicate that a coarse base event layer produces marginally more skillful daily precipitation hindcasts. The above gain, however, is too small to translate into improved skill in raw or postprocessed streamflow hindcasts.

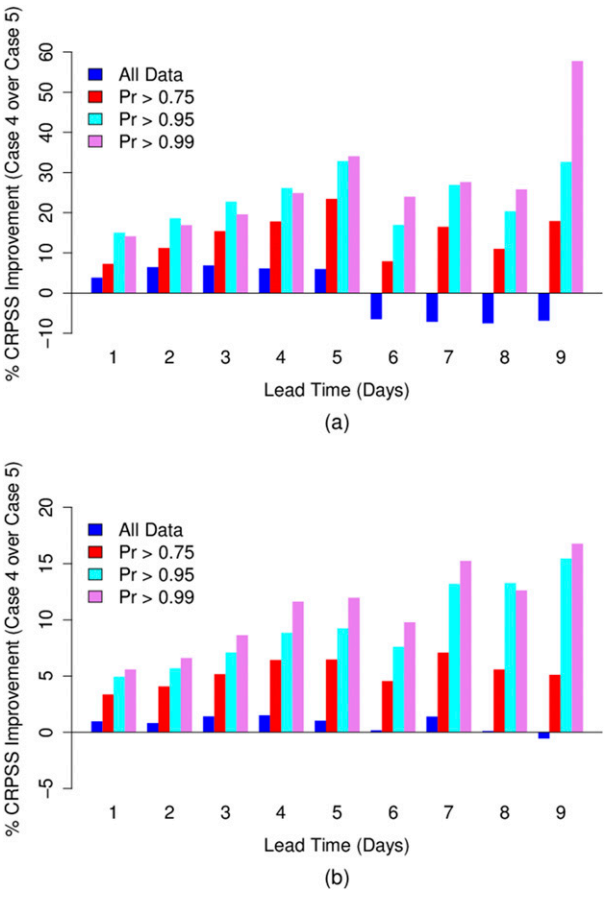

FIG. 6. (a) Percent increase in CRPSS of MEFP daily precipitation ensemble hindcasts due to the addition of five layers of modulation events (see Fig. 3 and Table 4) vs using the fine base events only. (b) As in (a), but for mean daily raw streamflow hindcasts.

Figure 6a shows the percent increase in mean CRPSS of the MEFP-GEFS ensemble hindcasts for daily precipitation due to adding the five layers of the modulation events shown in Fig. 3 versus using the fine base events only. The figure shows that the skill improvement in precipitation hindcasts ranges from 8% to 23% for the 75th percentile threshold and from 14% to 34% for thresholds of the 95th percentile or higher up to day 8 of forecast lead time. Figure 6b shows the percent increase in mean CRPSS of raw streamflow ensembles forced by the precipitation ensembles associated with Fig. 6a in reference to

those associated with using the fine base events only. Similar verification was carried out for the postprocessed streamflow forecasts with similar results. Figure 6b indicates that the use of the modulation events improves the skill in the raw and postprocessed streamflow hindcasts well beyond day 5 and that the improvement is up to 15% and 10% for the raw and postprocessed streamflow forecasts up to 8 days of lead time, respectively. The above findings indicate that the benefits of including modulation events are greater for larger precipitation and streamflow thresholds. We also evaluated adding only two layers of modulation events instead of five, that is, case 6 versus case 4. Compared to the skill improvement with five layers of modulation events, skill in precipitation improved only up to day 3 and decreases afterward with two layers only. The gain in skill in the raw and postprocessed streamflow hindcasts was also smaller, with only two layers of modulation events than with five. The findings indicate that adding a larger number of layers of modulation events generally improves skill in the MEFP ensemble precipitation forecasts as well as in the raw and postprocessed streamflow forecasts.

Comparisons between the monthly and semiannual scales of seasonal stratification, that is, case 1 versus case 2, indicate that, in the mean CRPSS sense, the monthly EnsPost parameters produce more skillful streamflow forecasts than the semiannual, and that the monthly parameters improve skill by up to 10%. That monthly stratification performs better than the semiannual may not be seen as surprising given the dependent nature of this validation. The previous hindcasting and verification experiments (Wu et al. 2010; Brown et al. 2014a,b) suggest, however, that the difference between dependent and independent validation is not very significant for the HEFS ensembles. Given the above, it is seen that monthly stratification is preferred for the EnsPost if the period of record is 55 years or longer. The above finding, however, is not expected to hold in the presence of nonstationarity for which additional research is needed. Based on the above, we used case 4 in Table 4, which employs five layers of modulations events, for the results presented below.

b. Precipitation results

This subsection presents the verification results for ensemble precipitation forecasts. The comparative verification of the MEFP-GEFS ensembles versus the MEFP-RFC is for the 10-yr period of 2005–14. The verification of multiday MEFP-GEFS ensembles is for the 31-yr period of 1985–2015. All precipitation results are based on pooling over all five basins.

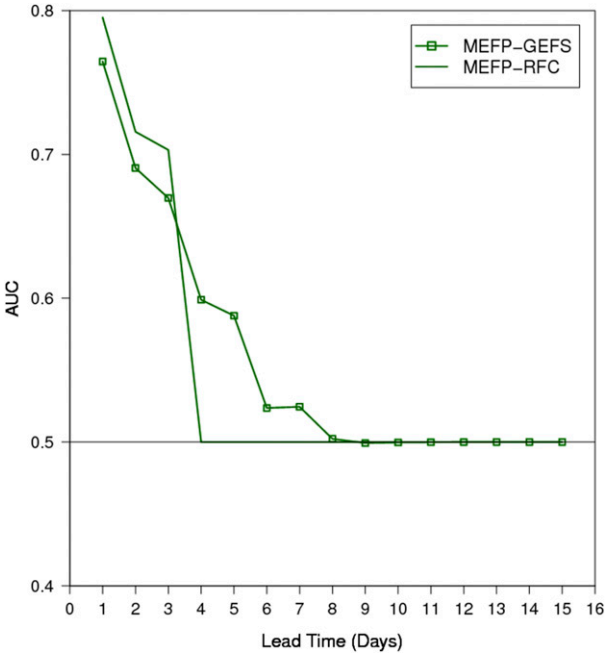

FIG. 7. AUCs for MEFP-RFC and MEFP-GEFS daily precipitation hindcasts with the 99th percentile (38.4 mm) of observed precipitation as the threshold.

1) MEFP-GEFS VERSUS MEFP-RFC ENSEMBLE FORECAST OF DAILY PRECIPITATION

Figure 7 shows the AUC for the MEFP-GEFS and MEFP-RFC precipitation ensemble forecasts at the 99th percentile of daily precipitation of 38.4 mm. In all ROC-related results in this paper, an event is defined as precipitation or streamflow exceeding the indicated threshold. Figure 7 shows that, for days 4 and 5, the MEFP-GEFS forecast has substantial discriminatory skill, which cannot be utilized effectively in the current single-valued forecast process. At the 97.5th percentile of 22.6 mm, the marginal gain in AUC by the MEFP-GEFS forecast over the MEFP-RFC is larger and extends to Day 7. At the 90th and 95th percentiles of 5.2 and 13.2 mm, respectively, the AUC is generally larger than that at the 97.5th percentile for both the MEFP-RFC and MEFP-GEFS forecasts. At the 75th percentile of 0.2 mm, however, the opposite is observed. The above observations indicate that in the study area both the MEFP-RFC and MEFP-GEFS forecasts of daily precipitation are most skillful in discriminating light (<5.2 mm) from significant (>13.2 mm) amounts.

2) MEFP-GEFS ENSEMBLE FORECAST OF MULTIDAY PRECIPITATION

Reservoir management in the study area requires inflow predictions over a wide range of temporal scales of

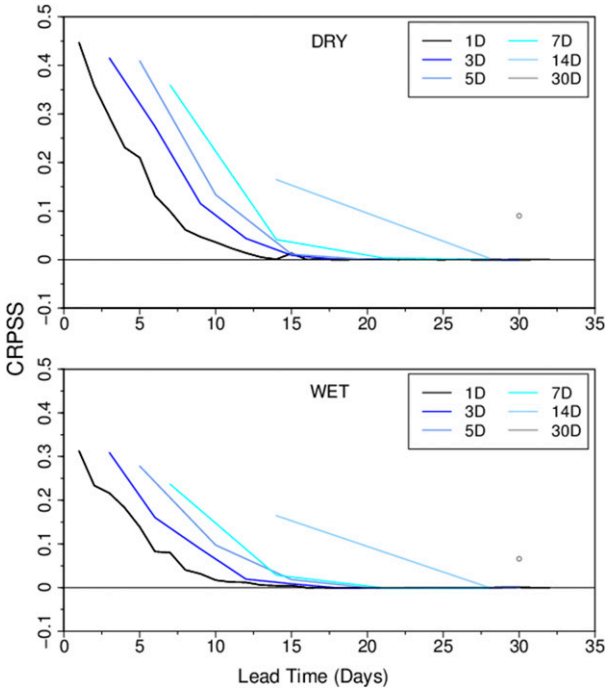

FIG. 8. Mean CRPSSs of the MEFP-GEFS precipitation ensembles for aggregation periods of 1, 3, 5, 7, 14, and 30 days for the (top) wet (March, April, May, June, September, October) and (bottom) dry (January, February, July, August, November, December) seasons conditional on the verifying observation exceeding the 99th percentile.

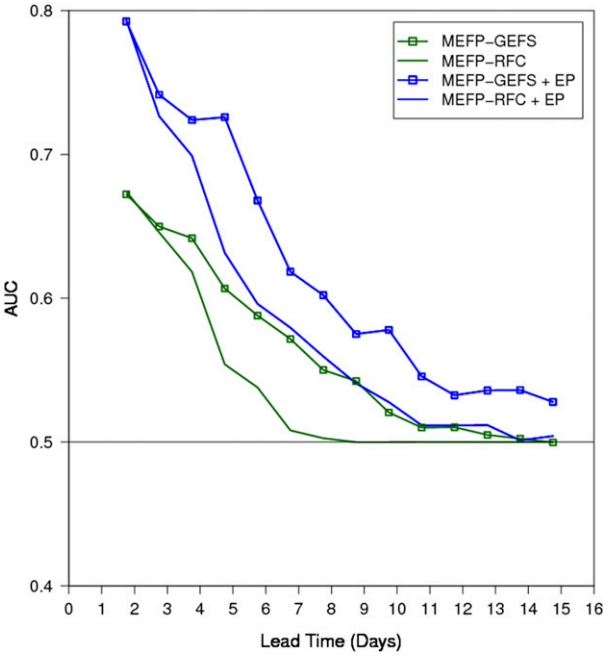

FIG. 9. AUCs of MEFP-RFC-forced and MEFP-GEFS-forced raw and postprocessed ensemble streamflow hindcasts for a threshold of the 99th percentile (31.6 cms) of observed mean daily flow.

aggregation. Because most of the predictive skill for precipitation is within the first two weeks or so of lead time in the study area, we focus here on verification of the MEFP-GEFS ensembles at aggregation periods of 1, 3, 5, 7, 14, and 30 days. Figure 8 shows the mean CRPSS of the MEFP-GEFS precipitation ensembles for all aggregation periods for the wet (March, April, May, June, September, October) and dry (January, February, July, August, November, December) seasons conditional on the verifying observation exceeding the 99th percentile. For reference forecast, resampled climatology is used. The figure indicates that the MEFP-GEFS ensemble forecasts for significant accumulations of 1-, 3-, 5-, and 7-day precipitation have mean CRPSS greater than 0.2 for lead times of up to about 3, 5, 7, and 8 days for the wet season and up to about 5, 7.5, 8.5 and 11 days for the dry season, respectively. A mean CRPSS of 0.2 corresponds to a 20% reduction in mean CRPS over climatological ensemble forecast (i.e., the reference forecast) and hence represents a significant skill. The above results indicate that there exists very significant skill in the MEFP-GEFS precipitation ensemble forecasts of up to about 14-day accumulations.

c. *Streamflow results*

Here we present the verification results for ensemble streamflow forecasts in two subsections that correspond to those for ensemble precipitation forecasts presented above.

1) MEFP-GEFS VERSUS MEFP-RFC ENSEMBLE FORECAST OF MEAN DAILY FLOW

Figure 9 shows the AUCs of raw and postprocessed ensemble streamflow hindcasts conditional on the verifying observed flow exceeding the 99th percentile threshold of 31.6 cms. At an AUC of 0.66, the MEFP-GEFS streamflow ensembles extend the forecast lead time only with postprocessing by about 1.5 days at this threshold. For the conditioning threshold of the 97.5th percentile of 14.2 cms, the increase in lead time is over a day without postprocessing and over 2.5 days with postprocessing at the same level of AUC. An AUC of 0.66 corresponds to the discriminatory skill of Day-2 MEFP-RFC streamflow ensemble forecast at the 99th percentile. Recall that the WGRFC routinely uses 3-day-ahead QPF in their operations. As such, the above level of AUC represents a skill level that may safely be considered useful for operational forecasting and provides a stringent reference for the assessment of the quality of the MEFP-GEFS streamflow ensembles and EnsPost. It was

observed that the AUCs peak when conditioned on the 75th percentile threshold of 0.3 cms and tend to decrease as the threshold increases or decreases. In general, the lower the conditioning threshold is, the larger the benefit from the EnsPost is, a reflection of the fact that low flow conditions tend to persist strongly. Examination of the basin-specific results indicates that the improvement in skill due to the EnsPost is relatively small for JAKT2 and SGET2 whether forced by the RFC QPF or the GEFS ensemble mean. The largest contributing factor to the reduced performance for these catchments is the more pronounced no-flow conditions in the dry season, which is not modeled with the current version of the EnsPost.

For flood forecasting, the PoD is a very important measure of forecast quality as it directly relates to the quality of warnings. Figure 10a shows the PoD at a PoFD of 5% for the MEFP-RFC and MEFP-GEFS streamflow ensembles at the 99th percentile threshold. Figure 10b shows the corresponding increase or decrease in the PoD at the same PoFD due to using the MEFP-GEFS ensembles relative to using the MEFP-RFC ensembles at the 99th percentile threshold. The benefit of the EnsPost is readily seen in Fig. 10a. Figure 10b shows that the MEFP-GEFS ensembles increase the PoD by close to 10% or more at the 99th percentile threshold for day 5–8 forecasts, a very significant improvement given the relatively modest PoD levels seen in Fig. 10a. While the evaluation above was carried out using the RFC QPF and GEFS ensemble mean separately to discern the value of each QPF source, in practice one would use both QPF sources in the MEFP to generate precipitation ensemble forecasts that are more skillful than using only a single source (see Tables 2 and 3).

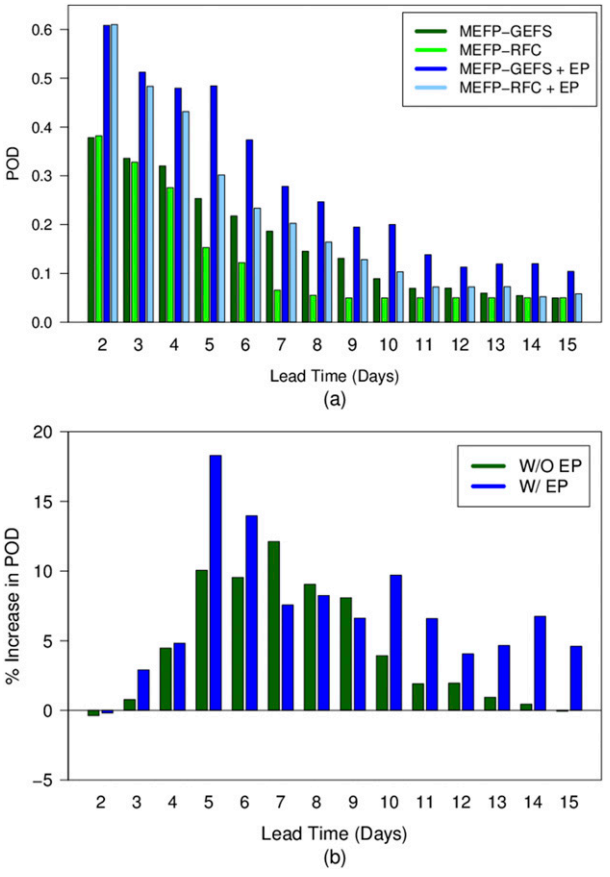

FIG. 10. (a) PoD at a PoFD of 5% of the MEFP-RFC and MEFP-GEFS streamflow ensembles for a threshold of the 99th percentile of observed mean daily flow. (b) Increase or decrease in the PoD at 5% PoFD by the MEFP-GEFS ensembles over the MEFP-RFC for a threshold of the 99th percentile of observed mean daily flow.

2) MEFP-GEFS ENSEMBLE FORECAST OF MULTIDAY FLOW

Figure 11 shows the mean CRPSS of the MEFP-GEFS streamflow ensemble forecasts with and without the EnsPost for aggregation periods of 1, 3, 5, 7, 14, and 30 days for the wet and dry seasons conditional on the verifying observation exceeding the 99th percentile. Figure 11 is based on pooling all five basins together. The positive impact of the EnsPost is readily seen. Figure 12 shows the 90% (between 5% and 95%) Monte Carlo intervals for mean CRPSS for SGET2, which indicates that the improvement due to the EnsPost is statistically significant. The catchment-specific results without the EnsPost are similar among all five catchments, but those with the EnsPost show significant differences with Fig. 11 for DCJT2 and JAKT2, for which the EnsPost provides larger and smaller improvement than the pooled results, respectively. The reduced positive impact of the EnsPost for JAKT2 is due to the significantly longer periods of no flow compared to the other catchments. As with the precipitation results, we also use mean CRPSS of 0.2 as a reference skill level for streamflow ensembles. Note that, because climatological streamflow ensemble forecasts are generally very skillful for short lead times owing to the memory of the hydrologic initial conditions, the above-referenced CRPSS represents a significantly larger absolute skill than that for precipitation ensemble forecasts. Figure 11 shows that the mean CRPSS of accumulations of 1- and 7-day accumulated streamflow ensemble forecasts approach or exceed 0.2 for short lead times even without the EnsPost, and that, with the EnsPost, the mean CRPSS of all streamflow ensemble forecasts approach or exceed 0.2 except those of 30-day accumulation. The above results indicate that the HEFS ensemble streamflow forecasts of up to 14-day accumulation have very significant skill which cannot be effectively utilized in the current single-valued forecast process.

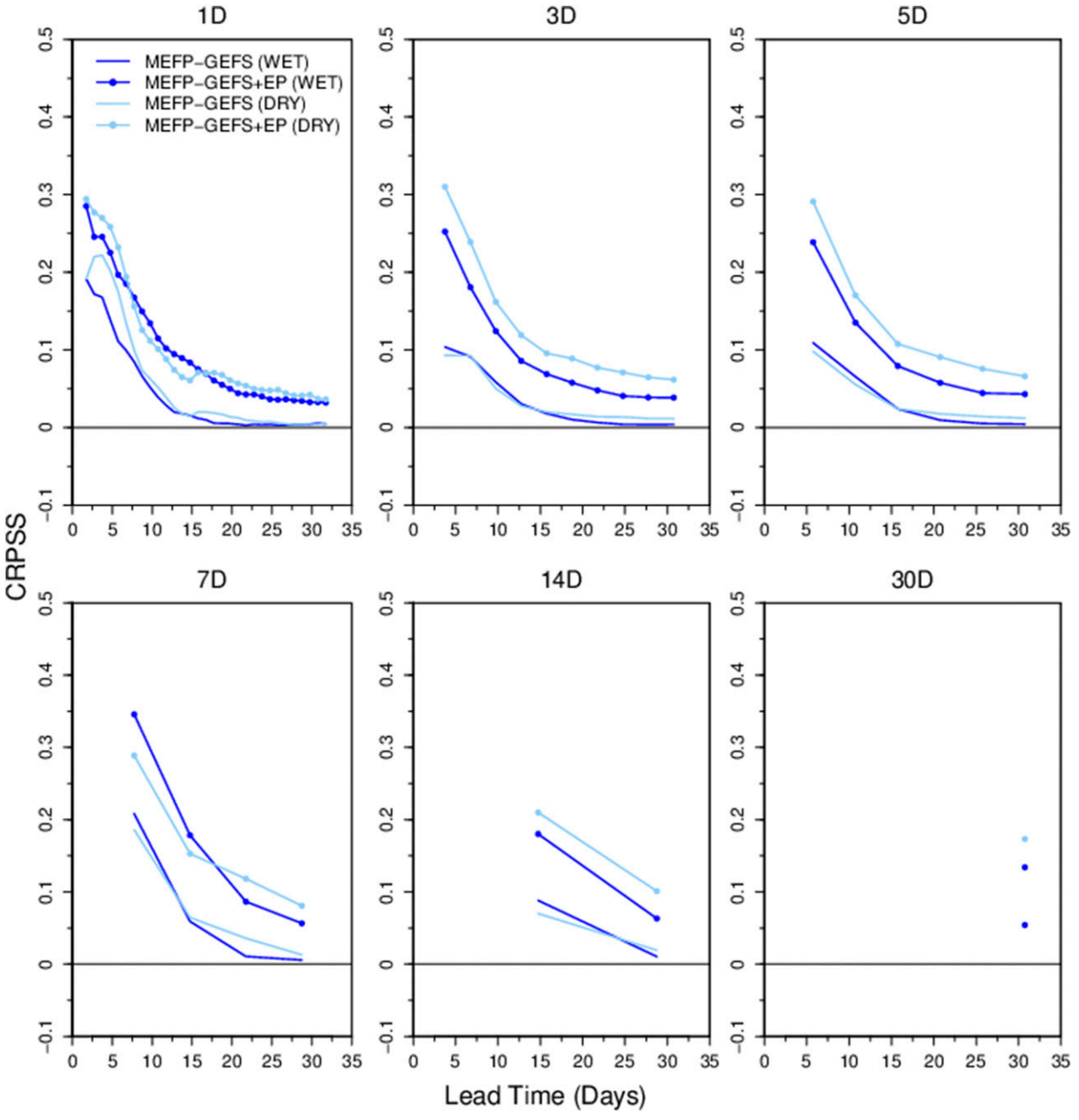

FIG. 11. Mean CRPSSs of the MEFP-GEFS streamflow ensemble forecasts with and without the EnsPost for all aggregation periods for the wet and dry seasons conditional on the verifying observation exceeding the 99th percentile.

Finally, Fig. 13 shows the monthly variation of the average mean CRPSS for day 1–15 forecasts versus the average monthly soil water depth simulated by SAC-SMA with observed precipitation forcing for SGET2 for all ranges of verifying observed flow. The figure indicates that, without the EnsPost, the MEFP-GEFS streamflow ensemble forecast provides larger improvement over climatological forecast in the fall wet months than in the spring wet months, due presumably to more skillful precipitation forecast in the cool season (Brown et al. 2014b), but offers little improvement in the dry summer months where very low soil moisture conditions persist. The EnsPost significantly improves skill not only in the wet spring and fall months but also in the winter dry season owing to the relatively wet soil moisture conditions. In the hydrologically very dry summer months of August and September, however, the EnsPost provides little improvement because climatology-forced

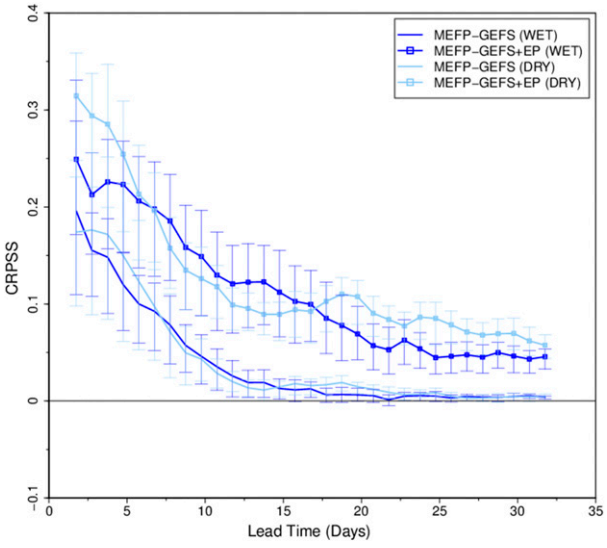

FIG. 12. The 90% (between 5% and 95%) Monte Carlo intervals for mean CRPSS of daily streamflow hindcasts for SGET2 conditional on the verifying observation exceeding the 99th percentile.

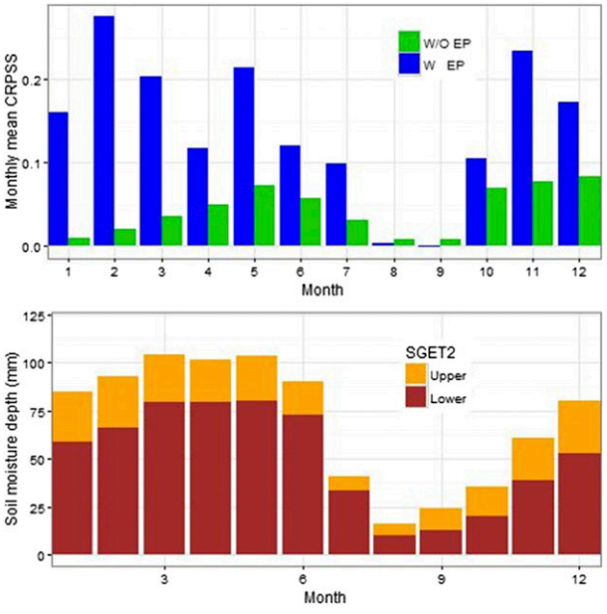

FIG. 13. Monthly variations of the average mean CRPSS for day 1–15 forecasts with and without the EnsPost vs the average monthly soil water depth simulated by SAC-SMA for SGET2 for all ranges of verifying observed flow.

streamflow ensembles are able to capture baseflow or no-flow conditions just as well.

4. Conclusions and future research recommendations

For emergency and water management, it is necessary to maximize forecast lead time while properly accounting for forecast uncertainties. In this work, we assess the skill of medium-range ensemble precipitation and streamflow forecasts generated with the HEFS developed by the NWS in extending the lead time and skill of operational streamflow forecasts.

The main conclusions of this work are as follows. The use of medium-range precipitation forecasts from the GEFS with the HEFS extends the time horizon for skillful forecasting of mean daily streamflow by 1–3 days for significant events when compared with using only the 72-h RFC QPF with the HEFS. For forecasting of multiday flow, the time horizon is extended significantly further. The GEFS-forced ensemble hindcasts of biweekly streamflow generated with the HEFS have mean CRPSS (reference forecast is resampled climatology) of about 0.2 for two-week-ahead prediction of observed flow of 99th percentile or larger. Without the EnsPost, however, the skill is considerably lower. The examination of the sensitivity of ensemble quality to the choice of the canonical events in the MEFP suggests that the use of the modulation events, which are associated with larger time scales than the base events, significantly improves the predictive skill in ensemble precipitation and streamflow forecasts. The results indicate that by employing modulation events in the MEFP the HEFS is able to capture at least partly the multiscale forecast skill in the GEFS and translate it into skill in streamflow forecasting. The overall findings strongly suggest that the operationalization of the HEFS in the region and elsewhere is expected to provide skillful medium-range ensemble precipitation and streamflow forecasts for high-impact events, particularly at multiday scales for a wide range of applications.

The main recommendations for future research are as follows. Most basins in the study area have significant periods of little or no flow during the dry season. To account for streamflow intermittency, improvement in the EnsPost is necessary. Implicit in the current statistical modeling of the HEFS is an assumption of stationarity, that is, the statistical relationships do not change materially over time and hence the past is a guide to the present and future (Brown et al. 2014a; NWS 2017a,b). Purely statistical techniques for modeling hydrologic and input uncertainties may have limited potency in the study area due to possible nonstationarities in the hydrologic and hydrometeorological processes arising from urbanization and climate change (Nazari et al. 2016; Norouzi 2016; Norouzi et al. 2018, manuscript submitted to *Stochastic Environ. Res. Risk Assess.*). To model predictive hydrologic uncertainty under nonstationarity and to allow parsimonious stochastic

modeling, more physically based approaches such as DA (Liu et al. 2012; Seo et al. 2014) are necessary. Parsimony in stochastic modeling is also necessary given that data-intensive modeling of probability distributions may not be viable under nonstationarity in many parts of the country.

Acknowledgments. This work is supported by the Sectoral Applications Research Program of the NOAA Climate Program Office Grant NA15OAR4310109. This support is gratefully acknowledged. We thank Mr. Mark Fresch and Drs. Hank Herr, Limin Wu, and Haksu Lee of the NWS Office of Water Prediction for various help during the course of this research; Messrs. Tom Donaldson (retired) and Robert Corby (retired) of the NWS WGRFC for various support in the early stages of this work; Mr. Kris Lander for helpful review of the manuscript; Mr. Brett Whitin of the NWS California–Nevada RFC (CNRFC) for sharing their HEFS parameters and data; and Dr. Edwin Welles of Deltares USA for facilitating the use of the Flood Early Warning System of Deltares for the NWS CHPS.

REFERENCES

Adams, T., and J. Ostrowski, 2010: Short lead-time hydrologic ensemble forecasts from numerical weather prediction model ensembles. *World Environmental and Water Resources Congress 2010,* Providence, RI, American Society of Civil Engineers, 2294–2304, https://doi.org/10.1061/41114(371)237.

Brown, J. D., 2015a: Ensemble Verification System (EVS) user's manual. Hydrologic Solutions Limited, 130 pp., https://vlab.ncep.noaa.gov/documents/207461/1893026/EVS_MANUAL.pdf .

——, 2015b: An evaluation of the minimum requirements for meteorological reforecasts from the Global Ensemble Forecast System (GEFS) of the U.S. National Weather Service (NWS) in support of the calibration and validation of the NWS Hydrologic Ensemble Forecast Service (HEFS). Tech. Rep. prepared by Hydrologic Solutions Limited for the Office of Hydrologic Development, 120 pp., http://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/publications_presentations/HSL_LYNT_DG133W-13-CQ-0042_SubK_2013_1003_Task_3_Deliverable_04_report_FINAL.pdf.

——, J. Demargne, D.-J. Seo, and Y. Liu, 2010: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Modell. Software*, **25**, 854–872, https://doi.org/10.1016/j.envsoft.2010.01.009.

——, L. Wu, M. He, S. Regonda, H. Lee, and D.-J. Seo, 2014a: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. *J. Hydrol.*, **519**, 2869–2889, https://doi.org/10.1016/j.jhydrol.2014.05.028.

——, M. He, S. Regonda, L. Wu, H. Lee, and D.-J. Seo, 2014b: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *J. Hydrol.*, **519**, 2847–2868, https://doi.org/10.1016/j.jhydrol.2014.05.030.

Burnash, R. J. C., 1995: The NWS River Forecast System–Catchment Modeling. *Computer Models of Watershed Hydrology*, V. P. Singh, Ed., Water Resources Publications, 311–366.

Chow, V. T., D. R. Maidment, and L. W. Mays, 1988: *Applied Hydrology.* McGraw-Hill, 572 pp.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.*, **5**, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Collischonn, W., C. E. M. Tucci, R. T. Clarke, S. C. Chou, L. G. Guilhon, M. Cataldi, and D. Allasia, 2007: Medium-range reservoir inflow predictions based on quantitative precipitation forecasts. *J. Hydrol.*, **344**, 112–122, https://doi.org/10.1016/j.jhydrol.2007.06.025.

Cosgrove, B., D. Gochis, T. Graziano, and E. Clark, 2017: From continental scale to neighborhood scale: Operational hydrologic modeling with the National Water Model. Accessed 2 May 2017, http://www.awrancrs.org/images/Presentations/Feb23-2017/Cosgrove.AWRA.NWM.Overview.2017.pdf.

Demargne, J., J. D. Brown, Y. Liu, D.-J. Seo, L. Wu, Z. Toth, and Y. Zhu, 2010: Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmos. Sci. Lett.*, **11**, 114–122, https://doi.org/10.1002/asl.261.

——, and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1.

Di Liberto, T., 2015: Flood disaster in Texas and Oklahoma. Climate.gov, accessed 4 April 2016, https://www.climate.gov/news-features/event-tracker/flood-disaster-texas-and-oklahoma.

Georgakakos, K. P., D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts, 2004: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *J. Hydrol.*, **298**, 222–241, https://doi.org/10.1016/j.jhydrol.2004.03.037.

——, and Coauthors, 2006: Integrated forecast and reservoir management (INFORM) for Northern California: System development and initial demonstration. HRC Tech. Rep. 5, 446 pp., http://www.hrc-lab.org/projects/projectpdfs/INFORM_REPORTS/FINAL_PHASE_I/HRC%20Technical%20Report%20No.%205.pdf.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Graziano, T., E. Clark, B. Cosgrove, and D. Gochis, 2017: Transforming National Oceanic and Atmospheric Administration (NOAA) water resources prediction. *31st Conf. on Hydrology*, Seattle, WA, Amer. Meteor. Soc., 2A.2, https://ams.confex.com/ams/97Annual/webprogram/Paper314016.html.

Green, D. M., and J. A. Swets, 1966: *Signal Detection Theory and Psychophysics.* Wiley, 455 pp.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, https://doi.org/10.1175/2007MWR2411.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

Hartman, R., D.-J. Seo, B. Lawrence, S. Shumate, J. Ostrowski, J. Halquist, C. Dietz, and M. Mullusky, 2007: The Experimental Ensemble Forecast System (XEFS) Design and Gap

Analysis. Rep. of the XEFS Design and Gap Analysis Team, 50 pp., http://www.nws.noaa.gov/oh/rfcdev/docs/XEFS_design_gap_analysis_report_final.pdf.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification. A Practicioner's Guide in Atmospheric Science*. Wiley, 240 pp.

Liu, Y., and Coauthors, 2012: Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities. *Hydrol. Earth Syst. Sci.*, **16**, 3863–3887, https://doi.org/10.5194/hess-16-3863-2012.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, https://doi.org/10.1256/003590002320603584.

Nazari, B., D.-J. Seo, and R. Muttiah, 2016: Assessing the impact of variations in hydrologic, hydraulic and hydrometeorological controls on inundation in urban areas. *J. Water Manage. Model.*, https://doi.org/10.14796/JWMM.C408.

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

Norouzi, A., 2016: Improving hydrologic prediction for large urban areas through stochastic analysis of scale-dependent runoff response, advanced sensing and high-resolution modeling. Ph.D. dissertation, Dept. of Civil Engineering, The University of Texas at Arlington, 280 pp., https://rc.library.uta.edu/uta-ir/handle/10106/27137.

NWS, 2017a: Meteorological Ensemble Forecast Processor (MEFP) user's manual. Office of Hydrologic Development, National Weather Service, 168 pp., https://vlab.ncep.noaa.gov/documents/207461/1893026/MEFP_Users_Manual.pdf.

——, 2017b: Ensemble Postprocessor (EnsPost) user's manual. Office of Hydrologic Development, National Weather Service, 69 pp., https://vlab.ncep.noaa.gov/documents/207461/1893026/EnsPost_Users_Manual.pdf.

Regonda, S. K., D.-J. Seo, B. Lawrence, J. D. Brown, and J. Demargne, 2013: Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts–A Hydrologic Model Output Statistics (HMOS) approach. *J. Hydrol.*, **497**, 80–96, https://doi.org/10.1016/j.jhydrol.2013.05.028.

Roe, J., and Coauthors, 2010: NOAA's Community Hydrologic Prediction System. *Second Joint Federal Interagency Conf.*, Las Vegas, NV, Advisory Committee on Water Information, 12 pp., https://training.weather.gov/nwstc/CHPS/roe.pdf.

Roundy, J. K., X. Yuan, J. Schaake, and E. F. Wood, 2015: A framework for diagnosing seasonal prediction through canonical event analysis. *Mon. Wea. Rev.*, **143**, 2404–2418, https://doi.org/10.1175/MWR-D-14-00190.1.

Saharia, M., 2013: Ensemble streamflow forecasting for the upper Trinity River basin in Texas. M.S. thesis, Dept. of Civil Engineering, The University of Texas at Arlington, 76 pp., http://hdl.handle.net/10106/23926.

Schaake, J., and Coauthors, 2007: Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth Syst. Sci. Discuss.*, **4**, 655–717, https://doi.org/10.5194/hessd-4-655-2007.

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1.

Seo, D.-J., H. D. Herr, and J. C. Schaake, 2006: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci. Discuss.*, **3**, 1987–2035, https://doi.org/10.5194/hessd-3-1987-2006.

——, L. Cajina, R. Corby, and T. Howieson, 2009: Automatic state updating for operational streamflow forecasting via variational data assimilation. *J. Hydrol.*, **367**, 255–275, https://doi.org/10.1016/j.jhydrol.2009.01.019.

——, J. Demargne, L. Wu, Y. Liu, J. D. Brown, S. Regonda, and H. Lee, 2010: Hydrologic ensemble prediction for risk-based water resources management and hazard mitigation. *Second Joint Federal Interagency Conf.*, Las Vegas, NV, Advisory Committee on Water Information, 27 pp., http://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/publications_presentations/Seo_et_al_JFIC_June2010.pdf.

——, Y. Liu, H. Moradkhani, and A. Weerts, 2014: Ensemble prediction and data assimilation for operational hydrology. *J. Hydrol.*, **519**, 2661–2662, https://doi.org/10.1016/j.jhydrol.2014.11.035.

TWDB, 2015: Water Data for Texas. Texas Water Development Board, http://waterdatafortexas.org.

WGRFC, 2015: WGRFC QPF page. West Gulf River Forecast Center, https://www.weather.gov/wgrfc/wgrfcqpfpage.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 648 pp.

Wu, L., D.-J. Seo, J. Demargne, and J. D. Brown, 2008: Generation of ensemble precipitation forecasts from single-value QPF via mixed-type meta-Gaussian model. *2008 Fall Meeting*, San Francisco, CA, Amer. Geophys. Union, Abstract H14D-07.

——, J. Schaake, J. D. Brown, J. Demargne, and R. Hartman, 2010: Generation of medium-range precipitation ensemble forecasts from the GFS ensemble mean at the basin scale. *2010 Fall Meeting*, San Francisco, CA, Amer. Geophys. Union, Abstract H23A-1165.

——, D.-J. Seo, J. Demargne, J. D. Brown, S. Cong, and J. Schaake, 2011: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *J. Hydrol.*, **399**, 281–298, https://doi.org/10.1016/j.jhydrol.2011.01.013.

Yuan, X., E. F. Wood, and M. Liang, 2014: Integrating weather and climate prediction: Toward seamless hydrologic forecasting. *Geophys. Res. Lett.*, **41**, 5891–5896, https://doi.org/10.1002/2014GL061076.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the New NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.