Evaluation and Benchmarking of Operational Short-Range Ensemble Mean and Median Streamflow Forecasts for the Ohio River Basin

THOMAS E. ADAMS III

Department of Civil and Environmental Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, and NOAA/NWS Ohio River Forecast Center, Wilmington, Ohio

RANDEL DYMOND

Department of Civil and Environmental Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia

(Manuscript received 14 May 2018, in final form 20 August 2018)

ABSTRACT

This study presents findings from a real-time forecast experiment that compares legacy deterministic hydrologic stage forecasts to ensemble mean and median stage forecasts from the NOAA/NWS Meteorological Model-Based Ensemble Forecast System (MMEFS). The NOAA/NWS Ohio River Forecast Center (OHRFC) area of responsibility defines the experimental region. Real-time forecasts from subbasins at 54 forecast point locations, ranging in drainage area, geographic location within the Ohio River valley, and watershed response time serve as the basis for analyses. In the experiment, operational hydrologic forecasts, with a 24-h quantitative precipitation forecast (QPF) and forecast temperatures, are compared to MMEFS hydrologic ensemble mean and median forecasts, with model forcings from the NOAA/NWS National Centers for Environmental Prediction (NCEP) North American Ensemble Forecast System (NAEFS), over the period from 30 November 2010 through 24 May 2012. Experiments indicate that MMEFS ensemble mean and median forecasts exhibit lower errors beginning at about lead time 90 h when forecasts at all locations are aggregated. With fast response basins that peak at \( \approx 24 \) h, ensemble mean and median forecasts exhibit lower errors much earlier, beginning at about lead time 36 h, which suggests the viability of using MMEFS ensemble forecasts as an alternative to OHRFC legacy forecasts. Analyses show that ensemble median forecasts generally exhibit smaller errors than ensemble mean forecasts for all stage ranges. Verification results suggest that OHRFC MMEFS NAEFS ensemble forecasts are reasonable, but needed improvements are identified.

1. Introduction

The use of hydrologic ensembles to produce probabilistic flood and water resources forecasts, using ensemble prediction systems (EPSs), is rapidly gaining acceptance (National Research Council 2006a,b; Cloke and Pappenberger 2009; Demargne et al. 2014). However, full adoption of probabilistic forecasts by the public and decision-makers as a replacement to traditional, single-valued deterministic hydrologic forecasts is problematic, particularly with how risk-based forecasts are communicated to end users (Demeritt et al. 2010; Pappenberger et al. 2013; Michaels 2015) and because of “institutional conservatism” (Rayner et al. 2005). The National Academies of Sciences Engineering and Medicine (2018) report that with weather-related decision-making, end users of weather-related forecasts benefit from 1) their understanding of forecasts developed over time, 2) prior experience with severe weather, and 3) other factors, such as family relationships. Related to end user familiarity with hydrometeorological forecasts, National Research Council (2006a) and Joslyn and Savelli (2010) found that end users understand that hydrometeorological forecasts are uncertain, but they make internal adjustments to account for these uncertainties. Morss et al. (2008) also found that end users of weather forecasts understand forecasts are uncertain and that most preferred the inclusion of uncertainty information with the forecasts. But as Demeritt et al. (2010, 209–222) points out, resistance to the acceptance of EPS forecasts is “not simply cognitive or communicative”; there is also the need by decision-makers to “shift institutional
liability for decisions taken in the face of uncertainty. Murphy (1991) and Krzysztofowicz (1998) argue for the adoption of probabilistic hydrometeorological forecasting, pointing out that rational decision-making in such a system necessarily shifts decision-making from the forecaster to end users of forecasts. An intuitive understanding of this undoubtedly helps to shape the reluctance by end users to adopt probabilistic hydrometeorological forecasts. In other words, resistance to the adoption of forecasts derived from EPSs, in the form of a probabilistic forecast, by both individuals and many decision-makers is complex, even with prior understanding that single-valued deterministic forecasts are uncertain. There is the added issue pointed to by Stern and Easterling (1999) that addresses the need for climate forecasts to be relevant to make them useful. This need applies to weather and hydrologic forecasts as well, which points to the broad issue, not addressed in this paper, of how to best convey forecast uncertainty to end users in ways that are relevant to them. We might ask, however, if there is an interim step with the use of the hydrologic ensemble prediction system (HEPSs), that can be taken that addresses two issues related to flood forecasting and the eventual adoption of probabilistic hydrologic forecasts, namely,

1) Improving flood forecast accuracy over current deterministic hydrologic forecasting methods that rely on single-valued quantitative precipitation forecast (QPF)
2) Softening the landscape for end users for eventual adoption of forecasts derived from HEPSs in the form of probabilistic forecasts.

In this paper, we explore the use of ensemble mean and median hydrologic forecasts from an HEPS as alternatives to deterministic predictions that depend on single-valued QPF. The study region in this paper is the forecast area of responsibility of the National Oceanic and Atmospheric Administration (NOAA), National Weather Service (NWS), Ohio River Forecast Center (OHRFC), shown in Fig. 1, which is one of thirteen NOAA/NWS River Forecast Centers (RFCs). Single-valued, deterministic QPF is a commonly used model forcing in hydrologic forecasting (Georgakakos and Hudlow 1984; Sokol 2003; Adams and Pagano 2016; Li et al. 2017) and used by all NWS RFCs. Research has demonstrated that the use of deterministic QPF introduces considerable
error into hydrologic forecasting (Cuo et al. 2011; Diomede et al. 2014; T. E. Adams and R. Dymond 2018, unpublished manuscript; Adams and Dymond 2018, manuscript submitted to *J. Hydrometeor.*). We hypothesize that ensemble mean or median forecasts have smaller error than deterministic hydrologic forecasts that rely on single-valued QPF, suggested by Du et al. (2003) and Mylne et al. (2002) with numerical weather prediction (NWP) ensemble modeling systems.

a. Background

NWS RFCs are responsible for providing routine river stage/flow forecast guidance to NWS Weather Forecast Offices (WFOs) following procedures described in Adams (2016) and Adams and Dymond (2018, manuscript submitted to *J. Hydrometeor.*). The central responsibility of most RFCs is flood prediction, although for RFCs in the western United States, water supply forecasting, largely for reservoir inflows is, perhaps, of greater importance. RFCs utilize the NWS Community Hydrologic Prediction System (CHPS; Adams 2016), based on the Flood Early Warning System (FEWS; Deltares 2018). CHPS modeling is predominantly interactive, as described by Adams and Smith (1993), within the Linux-based NOAA/NWS Advanced Weather Interactive Processing System (AWIPS; National Research Council 1997, 2006a). The OHRFC employs several models within the CHPS operational environment, including the Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al. 1973; Burnash 1995), the SNOW-17 snow accumulation and ablation model (Anderson 1973), several lumped-parameter hydrologic routing models, and three reservoir simulation models. All OHRFC CHPS models were migrated from the legacy NWS River Forecast System (NWSRFS; U.S. Department of Commerce 1972) in 2011, with parallel NWSRFS and CHPS modeling through 2012. In addition to QPF, principal hydrologic model forcings are observed precipitation and observed and forecasted temperature. Observed precipitation forcings are obtained from a multisensor estimation process involving rain gauges, NWS Next Generation Weather Radar (NEXRAD) Doppler radar, and, at some RFCs, remotely sensed satellite estimates of precipitation (Zhang et al. 2011; Kitzmiller et al. 2013; He et al. 2018). Forecasted precipitation is derived from NWP models, usually with meteorological forecaster adjustments made at both the NWS Weather Prediction Center (WPC) and/or at local RFCs (Novak et al. 2014).

b. Research goals

The aim of this research is to determine the utility of using hydrologic ensemble mean or median forecasts of river stage from the NOAA/NWS Meteorological Model-Based Ensemble Forecast System (MMEFS), described in Adams and Ostrowski (2010), as an alternative to current, operational, single-valued deterministic hydrologic stage forecasts at the OHRFC and, possibly, elsewhere. Section 2 of this paper describes the real-time hydrologic forecasting experiment used in this study. Model simulations are restricted to watersheds in the OHRFC area of responsibility, shown in Fig. 1. The experiment consists of the concurrent generation of OHRFC operational river stage forecasts and MMEFS ensemble forecasts for the 30 November 2010–24 May 2012 period. Verification results of the ensemble median and mean forecasts relative to the OHRFC operational forecasts are presented in section 3. Experimental results and mean forecasts are discussed in relation to verification of the MMEFS ensemble forecasts in section 4. Section 5 summarizes the experimental results and presents conclusions.

2. Research approach

The approach of this study is to compare OHRFC operational forecasts to MMEFS ensemble mean and median forecasts that use NWP model precipitation and temperature output from the NOAA/NWS National Centers for Environmental Prediction (NCEP) North American Ensemble Forecast System (NAEFS; Candille 2009) as hydrologic model forcings. The MMEFS, which utilizes raw NAEFS forcings, shown in Fig. 2, does not include pre- or postprocessing of either ensemble forcings or hydrologic ensemble output. All hydrologic modeling occurs within the OHRFC CHPS-FEWS forecasting system within AWIPS. The study period was from 30 November 2010 through 24 May 2012. The NAEFS consists of 42 ensemble members. The research methodology includes the following:

1) Capturing OHRFC operational forecasts initialized at 1200 UTC (daily), with a 5-day forecast horizon
2) Capturing automated MMEFS NAEFS hydrologic ensemble forecasts based on OHRFC 1200 UTC saved model states (daily), with a 7-day forecast horizon
3) Deterministic verification of operational forecasts and MMEFS NAEFS ensemble mean and median forecasts (after 24 May 2012)
4) Verification of MMEFS NAEFS ensemble forecasts (after 24 May 2012).

The real-time hydrologic forecasts were made using the legacy NWSRFS, relying on a geographically broad distribution of forecast point locations, with varying basin sizes and hydrologic response times for the OHRFC. All model forcing inputs and internal and output time steps
are 6-hourly. A total of 54 basins, shown in Fig. 3, were selected for the study. Calibrations of SAC-SMA, SNOW-17, channel routing, and reservoir simulation models for operational use for all OHRFC subbasins were completed long before the experiments started, following guidelines presented by Anderson (2002) and Smith et al. (2013). Operational and MMEFS simulations utilize a 6-h time step for model forcings, internally, and output. Forecasts are evaluated on the basis of comparisons between U.S. Geological Survey (USGS) observed stages and model-estimated river stage values, which were transformed from simulated flow values using USGS station rating curves. Deterministic verification followed methods proposed by Welles et al. (2007) and Demargne et al. (2009).

The 54 study basins (Fig. 3) are categorized as fast (Table 1), medium (Table 2), and slow (Table 3) responding. These include 26 fast, 20 medium, and 8 slow responding forecast point locations. The terms slow, medium, and fast refer to hydrograph time-to-peak-response times from the center of mass of the observed precipitation to the hydrograph peak. Response times less than 24 h are classified as fast, response times between 24 and 60 h are considered medium, and response times greater than 60 h are considered slow; see Office of Hydrologic Development (2000).

**a. Operational legacy forecasts**

The study relied on operational forecasts, using the OHRFC operational modeling system outlined in section 1a, covering the period 30 November 2010–24 May 2012. All operational forecasts used 24-h duration (four 6-h periods per 24 h) QPF. The experimental period spans 541 days at 54 locations, with 28 forecast periods.
each (four 6-h periods per day for five days), resulting in 817,992 forecast verification pairs for analysis. It should be pointed out that the operational forecasts used in this study include modeling of all 696 subbasins in the OHRFC area, an approximately 450,000 km² region, shown in Fig. 3. Operational forecast horizons are five days.

b. MMEFS ensemble forecasts

The automated MMEFS NAEFS hydrologic model ensemble simulations exactly parallel OHRFC operational forecasts. All simulations begin with 1200 UTC initializations, but they are not run until about 1800 UTC when NAEFS data are available from NCEP. Simulations utilize the full operational suite of models and follow identical operational workflows as the legacy deterministic OHRFC model forecast runs. Forecast horizons are seven days. An example MMEFS NAEFS forecast is shown in Fig. 4.

c. Ensemble forecast verification

Ensemble verification results are presented to demonstrate that MMEFS NAEFS ensemble median and mean forecasts are derived from a reasonably robust ensemble forecast system.

Demargne et al. (2010) discuss the need for verification of hydrologic ensemble forecasts, identifying the need to improve research and operations, specifically aimed at 1) monitoring changes in forecast quality over time, 2) analyzing sources of forecast error, and 3) evaluating forecast skill improvements resulting from the introduction of new science and technology. Consequently, we present some ensemble verification results to serve as a baseline evaluation of MMEFS forecast quality to identify areas of needed system improvement, which should further reduce MMEFS ensemble median and mean prediction errors.

Many statistical measures have been used to evaluate probabilistic forecasts. Johnson and Bowler (2009) suggest that, for example, resolution, the property that there should be large variability of observed frequencies associated with different forecast probabilities around the climatological value, is desirable. Also, ensemble forecasts should be reliable; that is, forecast probabilities should give an estimate of the expected frequencies of the event occurring. Welles et al. (2007) and Demargne et al. (2009) explain the necessity of hydrologic forecast verification. Hydrologic ensemble forecast verification methods, as recommended and discussed by Brown et al. (2010) and
Demargne et al. (2010) are used to assess the MMEFS NAEFS ensemble forecasts.

The statistical measures to evaluate MMEFS ensemble forecasts are discussed below. Ensemble verification results are presented only to indicate the overall reasonableness of the MMEFS NAEFS-based ensemble forecasts, not as a comprehensive evaluation of MMEFS ensemble forecasts.

1) CONTINUOUS RANKED PROBABILITY SKILL SCORE

The ranked probability score (RPS), shown in Eq. (1), is a measure of how well forecasts that are expressed as probability distributions are in matching observed outcomes,

\[
\text{RPS} = \frac{1}{r-1} \sum_{i=1}^{r} \left( \sum_{j=1}^{i} p_j - \sum_{j=1}^{r} e_j \right)^2, \tag{1}
\]

where \( r \) is the number of outcomes, \( p_j \) is the forecasted probability of outcome \( j \), and \( e_j \) is the actual probability of outcome \( j \). As a note, the special case where \( r = 2 \) gives the Brier score (Wilks 2006). The RPS applies to probability forecasts for discrete categories and the continuous ranked probability score (CRPS) extends the measure to continuous forecasts. Bradley and Schwartz (2011) show that the continuous ranked probability skill score [CRPSS; Eq. (4)] is a summary measure representing the weighted-average skill score, using climatology as the reference forecast, CRPS_{clim} in Eq. (4), over the continuous range of outcomes \( y \). The CRPSS, which can be derived from the mean-square error (MSE), is given by

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{r} (Y_i - \hat{Y}_i)^2, \tag{2}
\]

\[
\text{CRPS} = \int_{-\infty}^{\infty} \text{MSE}(y) \, dy, \quad \text{and} \tag{3}
\]

\[
\text{CRPSS} = 1 - \frac{\text{CRPS}}{\text{CRPS}_{\text{clim}}}, \tag{4}
\]

where \( Y_i \) and \( \hat{Y}_i \) are the observed and predicted values, respectively, of a verification pair. The overbar (CRPS) refers to averaging of CRPS values across the sample of events. CRPSS values can range from \(-\infty\) to 1, with perfect skill equal to 1 and negative values when the forecast has worse CRPS than the reference forecast.

2) RELIABILITY DIAGRAM

Reliability diagrams (Hartmann et al. 2002) represent, graphically, the observed frequency of an event plotted
against the forecast probability of an event. This expresses how often (as a relative frequency) a forecast probability actually occurred. The hit rate is calculated from the sets of forecasts for each probability separately. Consequently, the hit rate for each probability bin \( n \) is given by:

\[
\text{hit rate}_n = \frac{O_n}{O_n + N_n}, \quad \text{and}
\]

\[
F_n = \frac{O_n + N_n}{T},
\]

where \( F \) is the forecast frequency, \( O \) is the number of observed instances, \( N \) is the number of nonobserved instances, and \( T \) is the total number of forecasts.

3) **RANK HISTOGRAM**

Rank histograms are useful for evaluating ensemble forecasts because they can efficaciously assess the reliability and errors in the mean and spread of ensemble forecasts (Hamill 2001). Rank histograms are created by tallying the rank of observations relative to values from an ensemble sorted from lowest to highest rank, which produces, ideally, a uniform distribution across the ranks.

4) **RELATIVE OPERATING CHARACTERISTIC**

The relative operating characteristic (ROC; Kharin and Zwiers 2003) provides information on the hit rates and false alarm rates that can be expected from the use of different probability thresholds. The ROC is a summary score used to describe the ability of forecasts to discriminate between events and nonevents, from Table 4.

For Table 3, as in Table 1, but for slow response basins.

<table>
<thead>
<tr>
<th>ID</th>
<th>USGS ID</th>
<th>Name</th>
<th>Area (km²)</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAR12</td>
<td>03381500</td>
<td>Little Wabash River at Carmi, IL</td>
<td>8034</td>
<td>Slow</td>
</tr>
<tr>
<td>EVVI3</td>
<td>03322000</td>
<td>Ohio River at Evansville, IN</td>
<td>277699</td>
<td>Slow</td>
</tr>
<tr>
<td>GOLI2</td>
<td>03384500</td>
<td>Ohio River at dam 51 at Golconda, IL</td>
<td>372 699</td>
<td>Slow</td>
</tr>
<tr>
<td>HUFI3</td>
<td>03341500</td>
<td>Wabash River at Terre Haute, IN</td>
<td>31 766</td>
<td>Slow</td>
</tr>
<tr>
<td>NHR13</td>
<td>03378500</td>
<td>Wabash River at New Harmony, IN</td>
<td>75 716</td>
<td>Slow</td>
</tr>
<tr>
<td>NWBI3</td>
<td>03360500</td>
<td>White River at Newberry, IN</td>
<td>12 142</td>
<td>Slow</td>
</tr>
<tr>
<td>PTRI3</td>
<td>03373980</td>
<td>White River above Petersburg, IN</td>
<td>28 808</td>
<td>Slow</td>
</tr>
<tr>
<td>WTVO1</td>
<td>04195500</td>
<td>Maumee River at Waterville, OH</td>
<td>16 395</td>
<td>Slow</td>
</tr>
</tbody>
</table>
3. Study results

Verification results from the experiment, comparing the MMEFS NAEFS ensemble mean and median forecasts to OHRFC operational deterministic forecasts, are summarized in Figs. 5 and 6. Mean error (ME), mean absolute error (MAE), and root-mean-square error (RMSE), based on predicted and observed stage pairs, are shown by forecast lead time in hours. Figure 5 compares results from fast response basins to the results for all basins. Figure 6 shows MAE for medium response and slow response basins. Several observations can be made from Fig. 5, namely,

1) In most instances, there is little difference between ensemble mean and median values by lead time. With the exception of RMSE, where ensemble mean values are smaller than ensemble median values, ensemble median values are always smaller in magnitude than ensemble mean values, which suggest that ensemble median forecasts should be preferred over ensemble mean forecasts, since less error is incurred.

2) With the aggregation of all 54 basins, with respect to ME, little difference exists between the ensemble median forecast (and mean) and the OHRFC operational forecast (OHRFC 24-h QPF) through lead time 54 h. Beginning with lead time 60 h, the OHRFC operational forecast becomes increasingly more negatively biased with increased lead times, whereas the MMEFS ensemble median forecasts remain unbiased through lead time 168 h.

3) With fast response basins, the ensemble median forecast always shows ME values equal to or smaller preferred over ensemble mean forecasts, since less error is incurred.

**Table 4. Contingency table.**

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observed</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 4.** Example MMEFS NAEFS ensemble forecast showing 42 individual ensemble model members (various colors), the ensemble median (black line identified with triangles), and the 75%–25% probability of exceedance confidence band is shown as the shaded region. The minor and moderate flood levels are indicated for reference.
in magnitude than the OHRFC operational forecasts, which get increasingly more negative with longer lead times after lead time 72 h compared to ensemble median forecasts that very slowly become more negative with longer lead times.

4) OHRFC operational forecasts have smaller MAE values compared to MMEFS ensemble median and mean forecasts until lead time 96 h, with all basins aggregated; however, for fast response basins, MMEFS ensemble median and mean forecasts have MAE values equal to or smaller than OHRFC operational forecasts beginning at about lead time 36 h.

5) With respect to RMSE, with all basins aggregated, OHRFC operational forecasts exhibit smaller error compared to MMEFS ensemble median and mean forecasts until lead time 90 h, after which MMEFS ensemble median and mean forecast RMSE values are smaller than OHRFC operational forecast RMSE values; however, for fast response basins, MMEFS ensemble median and mean forecast RMSE values are approximately equal to or less than OHRFC operational forecast RMSE values beginning with lead time 72 h.

FIG. 5. ME, MAE, and RMSE (m) by lead time for all and fast response basins identified in Fig. 3 and in Tables 1, 2, and 3. Results are shown for the operational forecast (OHRFC 24-h QPF) and MMEFS NAEFS ensemble mean and median forecasts, 30 Nov 2010–24 May 2012.
Figure 6 shows that for medium and slow response basins, OHRFC operational forecasts tend to exhibit smaller forecast error compared to MMEFS ensemble median and mean forecasts until longer lead times are reached, ≥102 h. An explanation for this finding is discussed in section 4.

a. Forecast verification

Verification of the operational legacy forecasts use the R language and environment for statistical computing (R Core Team 2017) and contributed verification package (NCAR 2015). MMEFS NAEFS ensemble mean and median forecast verification statistics were obtained from ensemble analyses utilizing the NOAA/NWS ensemble verification service (EVS; Brown et al. 2010; Demargne et al. 2010). Operational forecast data are stored in the OHRFC PostgreSQL verification database, and MMEFS simulations are written to NWSRFS Ensemble Streamflow Prediction (ESP; Day 1985) format files. Verification measures used are ME, MAE, and RMSE, given in Eqs. (9), (10), and (11):

\[
\text{ME} = \frac{1}{n} \sum_{k=1}^{n} (y_k - o_k),
\]

\[
\text{MAE} = \frac{1}{n} \sum_{k=1}^{n} |y_k - o_k|,
\]

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (y_k - o_k)^2},
\]

where the quantities \(y_k\) and \(o_k\) are the predicted and observed \(k\)th stage values, respectively, for \(n\) total paired values. Units of measure for stage are meters unless reported otherwise. Values for ME, MAE, and RMSE = 0 imply perfect agreement, that is, no error.

b. Ensemble verification

1) CRPSS

Figure 7 summarizes the forecast skill of all 54 basins, with aggregation across all forecast stage ranges (Fig. 7a) and for stage ranges ≥ 0.90 probability of nonexceedance (Fig. 7b). The results show that MMEFS NAEFS forecasts are skillful relative to unconditional sample climatology \(f_X\) utilizing all observations over the period 30 November 2010–24 May 2012. Results also show that forecasts for stage ranges ≥ 0.90 probability of nonexceedance are more skillful than forecasts that are aggregated across all stage ranges, which is encouraging since RFCs emphasize flood forecasting. Figure 7 shows differentiation between fast, medium, and slow response basins, illustrating that MMEFS ensemble forecasts are most skillful for slow response basins and that forecast skill is most variable for medium response basins. Forecast skill, as expected, declines with increased lead times.
An interesting point relates to some of the lowest CRPSS values in Fig. 7a, which correspond to the forecast point location at Pittsburgh, Pennsylvania (PTTP1). Referring to Table 5, CRPSS values aggregated across all forecast stage ranges demonstrate little to no skill. However, for stage ranges $\geq 0.90$ probability of non-exceedance, MMEFS ensemble forecasts show reasonable skill. The reason for this difference relates to the
complex physical setting described by Adams et al. (2018) involving downstream control at the Dashields, Pennsylvania, lock and dam that regulates the pool level at the PTTP1 stream gauge at low flows in a manner that is not well captured by the OHRFC modeling system. With higher flows, particularly flood flows, this control does not exist.

2) RELIABILITY

Figure 8 shows the reliability diagram for stage ranges $0.50$ and $0.90$ probability of nonexceedance, for lead times of 24, 48, 96, 120, and 168 h, aggregated across all 54 basins. Generally, the ensemble forecasts show reasonable reliability, but that forecast overconfidence exists between forecast probability ranges 0.50–0.75 for both stage ranges. Results for lead time of 24 h for the $\geq 0.90$ stage range is, most likely, representative of small-sample-size problems, which is a general concern since the study period was short, only 541 days, 30 November 2010–24 May 2012. But for $\geq 0.50$ stage range, 24-h lead-time reliability is better, possibly because of either a larger sample size or reduced error or both. The $\geq 0.50$ stage range forecasts do exhibit a small degree of underconfidence at smaller forecast probability levels compared to the $\geq 0.90$ stage range that may likely be related to an overstatement of the importance of near-term QPF uncertainty relative to model state influences.

3) ROC

Figure 9 shows the ROC diagram, aggregated across all 54 basins, for stage ranges $\geq 0.90$ probability of nonexceedance for forecast lead times of 24, 96, 120, and 168 h, indicating that MMEFS NAEFS ensemble forecasts discriminate between events and nonevents very well. This result is representative of similar analyses done for individual basins and for all forecast stage ranges.

4) RANK HISTOGRAM

Rank histogram analyses are summarized with aggregation across all 54 basins in Fig. 10, showing severe underspread of MMEFS ensemble forecasts, particularly at shorter lead times, illustrated for lead time of 24 h.

<table>
<thead>
<tr>
<th>Lead time (h)</th>
<th>CRPSS (all)</th>
<th>CRPSS ($p = 0.90$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.2172</td>
<td>0.5758</td>
</tr>
<tr>
<td>24</td>
<td>0.2294</td>
<td>0.6261</td>
</tr>
<tr>
<td>48</td>
<td>-0.0265</td>
<td>0.4014</td>
</tr>
<tr>
<td>72</td>
<td>-0.0751</td>
<td>0.3930</td>
</tr>
<tr>
<td>96</td>
<td>-0.1014</td>
<td>0.3686</td>
</tr>
<tr>
<td>120</td>
<td>-0.0834</td>
<td>0.3810</td>
</tr>
<tr>
<td>144</td>
<td>-0.0800</td>
<td>0.3527</td>
</tr>
<tr>
<td>168</td>
<td>-0.1182</td>
<td>0.3351</td>
</tr>
</tbody>
</table>

TABLE 5. CRPSS for Pittsburgh (PTTP1) for all stage ranges and for stages with probability of exceedance $p = 0.90$ by lead time.

FIG. 8. Reliability diagrams for all 54 basins for lead times of 24, 48, 96, 120, and 168 h. Shown for stage ranges (left) $\geq 0.50$ and (right) $\geq 0.90$ probability of nonexceedance.
An explanation for this result is shown in Fig. 4, which was specifically selected to illustrate the underspread problem. In this example, the 42 MMEFS NAEFS ensembles appear as a single-valued forecast from the beginning of the forecast at 1800 UTC 4 April 2018 through 1200 UTC 7 April 2018. This occurred because there were 1) no model forcings to substantially perturb the hydrologic models over the period from 1800 UTC 4 April through 1200 UTC 7 April and 2) because inherent model error is not included in the MMEFS.

4. Discussion

Two main topics of discussion are as follows: 1) MMEFS ensemble median and mean forecasts compared to OHRFC operational forecasts and 2) needed improvements to the MMEFS. The discussion on the first point focuses on the differences found between MMEFS ensemble median and mean forecasts for fast response basins versus what is seen with medium and slow response basins compared to OHRFC operational forecasts. We suggest there are two principal explanations. First, there are differences in the quality of model calibrations, and second, there are independent operational modeling complications that account for larger initial errors for medium and slow response basins that are not present with fast response basins.

a. MMEFS ensemble median and mean forecasts

Results presented in Fig. 5 clearly show, in section 3, with all 54 basins aggregated, that MMEFS ensemble median and mean forecasts add considerable value at longer lead times above OHRFC operational forecasts at lead times $\geq 90$ h. An important result is that the MMEFS ensemble median and mean forecasts for fast response basins have smaller forecast errors than basins with longer basin response times beginning with lead times $\geq 36$ h for MAE. MMEFS ensemble median and mean forecasts for fast response basins display smaller error immediately with ME, at lead time of 6 h. This finding is important because it demonstrates the viability of using MMEFS ensemble median and mean forecasts as an alternative to the continued use of OHRFC operational forecasts relying on 24- or, currently, 48-h QPF.

There are likely several factors that contribute to lower errors found with shorter lead times, $\leq 90$ h, for OHRFC operational forecasts compared to MMEFS ensemble median and mean forecasts. These include the following:

1) The fact that OHRFC operational forecasts benefit from real-time interactive forecaster adjustments in the near term to account for hydrologic model state inaccuracies, which are infeasible with ensemble processing workflows.
2) The incorporation of observed flows in streamflow routings for nonheadwater basins with OHRFC operational forecasts, which is not tenable with hydrologic ensemble forecasting.

3) The QPF used in OHRFC operational forecasts is closer in time to the beginning of forecast initiation than is currently possible with MMEFS ensemble QPF inputs with the delay in computer processing. That is, meteorological conditions are likely to change enough in the near term that are not reflected in the MMEFS ensemble QPF inputs but maybe better captured in OHRFC operational QPF.

Several factors influence the disparity in MMEFS ensemble median and mean forecast verification results for medium and slow response basins compared to fast response basins. These include the following:

1) Model calibrations are good for fast response, headwater basins, but calibrations for medium and slow response basins are more problematic. Calibration for nonheadwater basins is more difficult because of, in part, overwhelming upstream flow influences that cannot be accurately separated from total observed streamflow at local nonheadwater gauges. The separation of flows routed from upstream basins from the observed total streamflow is necessary to estimate the local observed streamflow. With the local observed streamflow, model parameter adjustments to the local basin can be made appropriately to reflect the hydrologic response of the local watershed being calibrated. The consequence is that hydrologic model parameters used for downstream basins usually can only be estimated, not calibrated. The outcome is that local downstream
basin simulations can be considerably more in error than fast response headwater basins that have been calibrated.

2) Overall modeling for downstream medium and slow response basins is more complex, involving the use of flow routing and reservoir simulation models, which adds to modeling uncertainty and error.

3) For some medium and all slow response basins, complex channel flow dynamics, including hydrodynamic backwater effects, adversely affects stage–discharge relationships and, on the main stem of the Ohio River, lock and dam controls influence stage–discharge relationships (Adams et al. 2018). The implication here is that, while flow simulations may be good, stage–discharge relationships are nonunique because of hysteresis. This nonunique stage–discharge relationship precludes unique conversion from flow to stage. Consequently, considerable error is incurred, which adversely influences verification scores because of the introduction of erroneous stage values. Manual forecaster adjustments in the OHRFC operational workflow are made, in part, to minimize stage–discharge ratings hysteresis effects. Such adjustments are not possible in MMEFS NAEFS simulations. The need for dynamic flow routing modeling served as the basis for the development and operational implementation of the Ohio River Community Hydrologic Engineering Center River Analysis System (HEC-RAS) Model (USACE) deterministic reservoir release schedules, including flow releases from locks and dams, which are incorporated into the operational forecasts. MMEFS modeling relies on NWSRFS-based (now CHPS) reservoir model simulations, which, under many scenarios, can be quite erroneous compared to actual USACE reservoir releases, which involves human decision-making that is difficult to capture in a reservoir simulation model. Figure 3 shows that all slow and all but two medium response basin locations are downstream of significant reservoirs with maximum storage capacities \( \geq 250,000 \) acre feet (ac ft; \( 308,370,000 \) m\(^3\)). All of the fast response basins are upstream of these reservoirs. Consequently, OHRFC operational forecast errors are reduced in the near term relative to MMEFS NAEFS ensemble median and mean forecasts that rely solely on model simulations of reservoir outflows.

4) Reservoir releases significantly alter downstream flows. OHRFC operational forecasts benefit from the inclusion of U.S. Army Corps of Engineers (USACE) deterministic reservoir release schedules, which, under many scenarios, can be quite erroneous compared to actual USACE reservoir releases, which involves human decision-making that is difficult to capture in a reservoir simulation model. Figure 3 shows that all slow and all but two medium response basin locations are downstream of significant reservoirs with maximum storage capacities \( \geq 250,000 \) acre feet (ac ft; \( 308,370,000 \) m\(^3\)). All of the fast response basins are upstream of these reservoirs. Consequently, OHRFC operational forecast errors are reduced in the near term relative to MMEFS NAEFS ensemble median and mean forecasts that rely solely on model simulations of reservoir outflows.

b. MMEFS improvements

MMEFS ensemble forecasts do not currently make use of ensemble model-error correction methods suggested by, for example, Bogner and Kalas (2008) and Li et al. (2016) or postprocessing bias correction of raw ensemble forecasts proposed by, for example, Hashino et al. (2007), Brown and Seo (2010), and Yuan and Wood (2012) to reduce uncertainties arising from model inputs and outputs, initial and boundary conditions, and the structure and parameter estimates of models. Wentao et al. (2017) recently reviewed statistical post-processing methods for hydrometeorological ensemble forecasting, citing the need for further work on many fronts. These concerns include the need to address stationarity assumptions; handle extreme events, including the timing of flood peaks in the case streamflow modeling; further investigate methods proposed to make adjustments at ungauged locations; and continue research into methods that attempt to address total uncertainty, including model structure, parameter estimation, and model initial and boundary conditions. The need for the use of such techniques is illustrated in section 4b where MMEFS ensemble verification results are discussed related to ensemble underspread and apparent biases. Although not strictly an ensemble-modeling-related issue, MMEFS ensemble forecasts would benefit from improved hydrologic model calibrations for many basins. Model simulation error would be reduced further and MMEFS verification would also improve by incorporating the Ohio River Community HEC-RAS Model directly into MMEFS ensemble simulations for slow response rivers where complex hydrodynamics are inadequately handled by simple streamflow routing models. MMEFS ensemble verification metrics are clearly better for basins known to have good calibrations than for basins with suboptimal model calibrations.

5. Summary and conclusions

Experimental results from this study demonstrate that NAEFS-based MMEFS ensemble median forecasts have smaller forecast error than ensemble mean forecasts based on ME and MAE verification measures. Even though RMSE values suggest slightly lower forecast error with ensemble mean compared to ensemble median forecasts, analyses, overall, suggest that ensemble median forecasts should be preferred over ensemble mean forecasts. More importantly, when forecasts spanning all ranges in stage and basin response times are aggregated, MMEFS ensemble mean and median forecasts show lower forecast error than legacy OHRFC operational forecasts, based on 24-h deterministic QPF,
at long forecast lead times, that is, beginning at approximately \( \approx 90 \) h. This result shows the viability of using ensemble mean/median forecasts for extended forecasts beyond the 4-day forecast horizon. When analyses are restricted to fast response basins only, MMEFS ensemble mean and median forecasts have smaller forecast error than legacy OHRFC operational forecasts beginning at about lead time of 36 h. This finding has potentially significant implications on forecast operations at the OHRFC and other hydrologic forecast centers. Specifically, lower MMEFS ensemble mean/median forecast errors compared to the current operational OHRFC deterministic forecasts suggests the feasibility of changing operational workflows from manually intensive, interactive forecasting procedures to a more automated operational environment using ensemble forecasting methodologies. Of course, the generation of full probabilistic forecasts from ensemble methods is greatly preferred over the use of the ensemble mean or median because the former conveys forecast uncertainty to end users, where the latter does not (National Research Council 2006a). In addition, a wide range of water resources applications requires the use of flow hydrographs; the proposed ensemble median/mean stage forecasts should not be translated to flow hydrographs because mass balance is not preserved. Consequently, for most other water resources applications, full hydrologic ensemble forecasts are needed.

MMEFS ensemble verification results covering the 30 November 2010–24 May 2012 study period demonstrate forecast skill and reasonable forecast discrimination, reliability, and sharpness. However, verification results also identify needed areas of improvement, such as the need to account for model error, utilizing a priori hindcast experiments to quantify model error, and the adoption of a postforecast, ensemble bias correction methodology.

Acknowledgments. The authors are grateful to James Brown, Hydrologic Solutions, Limited, Winchester, United Kingdom, for providing helpful guidance in this study and by providing updated EVS software to the authors. We also thank the anonymous reviewers for their comments, which helped to improve the manuscript.

REFERENCES


