Ensemble Streamflow Forecasting across the U.S. Mid-Atlantic Region with a Distributed Hydrological Model Forced by GEFS Reforecasts

RIDWAN SIDDIQUE AND ALFONSO MEJIA
Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania

(Manuscript received 17 October 2016, in final form 6 April 2017)

ABSTRACT

The quality of ensemble streamflow forecasts in the U.S. mid-Atlantic region (MAR) is investigated for short- to medium-range forecast lead times (6–168 h). To this end, a regional hydrological ensemble prediction system (RHEPS) is assembled and implemented. The RHEPS in this case comprises the ensemble meteorological forcing, a distributed hydrological model, and a statistical postprocessor. As the meteorological forcing, precipitation, and near-surface temperature outputs from the National Oceanic and Atmospheric Administration (NOAA)/National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast, version 2 (GEFSRv2), are used. The Hydrology Laboratory Research Distributed Hydrologic Model (HL-RDHM) is used as the distributed hydrological model, and a statistical autoregressive model with an exogenous variable is used as the postprocessor. To verify streamflow forecasts from the RHEPS, eight river basins in the MAR are selected, ranging in drainage area from 262 to 29,965 km² and covering some of the major rivers in the MAR. The verification results for the RHEPS show that, at the initial lead times (1–3 days), the hydrological uncertainties have more impact on forecast skill than the meteorological ones. The former become less pronounced, and the meteorological uncertainties dominate, across longer lead times (>3 days). Nonetheless, the ensemble streamflow forecasts remain skillful for lead times of up to 7 days. Additionally, postprocessing increases forecast skills across lead times and spatial scales, particularly for the high-flow conditions. Overall, the proposed RHEPS is able to improve streamflow forecasting in the MAR relative to the deterministic (unperturbed GEFSRv2 member) forecasting case.

1. Introduction

Managing water is a complex challenge faced with increasing difficulties due to climate change, rapid urbanization, competing demands for various water services, and socioeconomic (i.e., financial, governmental, and cultural) barriers and constraints (Famiglietti and Rodell 2013; Kelly 2014; Mekonnen and Hoekstra 2016; Vorösmarty et al. 2000). To improve decision-making in various areas of water policy and management (e.g., flood and drought preparedness, water supply, reservoir operations, hydropower generation, and navigation), streamflow forecasts are essential (Alfieri et al. 2014; Day 1985). Streamflow forecasts are often generated by a hydrological forecasting system forced by outputs from a numerical weather prediction (NWP) model, whereby the uncertainties in the meteorological outputs are propagated into the streamflow forecasts. To characterize and assess the uncertainty of hydrological forecasts, hydrological ensemble prediction systems (HEPS) are increasingly being implemented in both research and operational applications (Addor et al. 2011; Cloke and Pappenberger 2009; Demeritt et al. 2010; Fan et al. 2014; Khan et al. 2015; Olsson and Lindström 2008; Thielen et al. 2009). HEPS, although relatively recent, have demonstrated improved performance over deterministic forecasts in various water-related applications (Alemu et al. 2011; Anghileri et al. 2016; Bartholmes et al. 2009; Bennett et al. 2014; Boucher et al. 2011; Brown et al. 2014; Franz et al. 2008; Fundel and Zappa 2011; Georgakakos et al. 2014; Harshburger et al. 2012; Schellekens et al. 2011; Van Cooten et al. 2011; Verbunt et al. 2007; Wood et al. 2016).

HEPS consider current, plausible states of meteorological and hydrological variables to predict multiple realizations of future streamflows (Franz et al. 2008; Schaake et al. 2006, 2007). To account for input or forcing uncertainty, HEPS are forced with ensembles of meteorological outputs (e.g., precipitation and near-surface temperature) from NWP models to generate short-range (0–3 days) and medium-range (3–14 days) forecasts (Alfieri et al. 2014; Ramos et al. 2013;
Roulin and Vannitsem 2015; Yuan et al. 2014). For example, the current European Flood Awareness System uses operational weather forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) to produce medium-range flood forecasts (Thielen et al. 2009; Wetterhall et al. 2013). In the United States, the National Oceanic and Atmospheric Administration’s (NOAA) National Weather Service (NWS) is implementing ensemble weather forecasts operationally for hydrological forecasting (Demargne et al. 2014; NOAA 2014). With these developments in hydrological forecasting science, the need arises for scientific studies to verify and benchmark the performance of HEPS.

A key component of HEPS is the hydrological model(s) used to forecast streamflow or other hydrological outputs. Thus far, HEPS have been mostly evaluated using so-called lumped or semidistributed hydrological models that do not account, or only in a limited fashion, for the spatial variability of inputs (meteorological, topographical, pedological, land cover, etc.), parameters, and states (Carpenter and Georgakakos 2006). Indeed, there are many advantages to distributed hydrological models as demonstrated and extensively discussed elsewhere (Boyle et al. 2001; Carpenter and Georgakakos 2006; Michaud and Sorooshian 1994; Krajewski et al. 1991; Smith et al. 2012; Spies et al. 2015). In particular, they allow the spatially seamless prediction of different hydrological variables. The implementation of distributed hydrological models forced by ensemble meteorological forecasts, however, is computationally intensive; because of this and potentially other reasons (e.g., effort required to calibrate models), only a few studies have been carried out, and comprehensively evaluated, that apply a distributed hydrological model within an HEPS (Alfieri et al. 2014; Anghileri et al. 2016; Fan et al. 2014; Georgakakos et al. 2014; Jörg-Hess et al. 2015; Xuan et al. 2009; Yuan et al. 2014). For example, Alfieri et al. (2014) verified ensemble streamflow forecasts for several years from LISFLOOD, forced by meteorological ensembles from the ECMWF, while Yuan et al. (2014) used the National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast, version 2 (GEFSRv2; Hamill et al. 2013; Siddique et al. 2015), to force the Variable Infiltration Capacity model. Jörg-Hess et al. (2015) utilized the ECMWF meteorological ensembles to force the Precipitation Runoff Evapotranspiration Hydrotope distributed hydrological model in snow-dominated, mountainous river basins. All these studies demonstrate improved skill for medium-range ensemble streamflow forecasts generated by using a distributed hydrological model, although most of them use daily or even coarser temporal resolution. In contrast, we employ here hourly and 6-hourly resolutions for the hydrological model simulations and forecasts, respectively.

Streamflow forecasts generated from meteorological ensembles tend to exhibit systematic biases that make the determination of forecast probabilities from such streamflow data unreliable (Roulin 2007; Schaake et al. 2007). To correct the biases and improve the reliability of streamflow forecasts, statistical postprocessing techniques are used (Wood and Schaake 2008; Zalachori et al. 2012). Indeed, postprocessing is an integral component of an HEPS that must be considered when verifying the quality of the HEPS’s streamflow forecasts (Roulin and Vannitsem 2015). The general goal with postprocessing is to improve the performance (e.g., skill) of the HEPS by bias correcting the forecasts based on the statistical behavior of past forecasts from the same HEPS. A crucial prerequisite of postprocessing is thus the availability of long training datasets composed of past streamflow forecasts (Roulin and Vannitsem 2015). This can be challenging when dealing with operational systems that are constantly evolving, thereby making the use of weather reforecasts indispensable (Siddique et al. 2015). A number of postprocessing techniques have been proposed for streamflow forecasts (Hashino et al. 2007; Madadgar et al. 2014; Pagano et al. 2013; Van Steenbergen et al. 2012), which were recently categorized and discussed by van Andel et al. (2013). Additionally, weather preprocessing is often used to improve the performance of the meteorological forecasts prior to their implementation in the HEPS. The focus of this study is, however, on the benefits of jointly implementing distributed hydrological modeling and postprocessing to improve ensemble streamflow forecasts across spatial scales.

In particular, our primary objective with this study is to investigate the ability of a regional HEPS (RHEPS) to improve short- to medium-range streamflow forecasts in the U.S. mid-Atlantic region (MAR). The objective is also to quantify the relative importance of different sources of uncertainty (meteorological and hydrological) in streamflow forecasts. This research represents one of only a handful of studies to actually generate, and rigorously verify, ensemble streamflow forecasts over relatively large basin areas by forcing a high-resolution (both spatially, $2 \times 2$ km$^2$ grid cells, and temporally, 6-hourly forecasts over a 10-yr period out 7 days) distributed hydrological model with weather ensembles. Although similar implementations have been carried out using forecasts from the ECMWF (Jörg-Hess et al. 2015), there are very few published studies that use GEFSRv2 data to force a distributed hydrological model. To meet the proposed research objectives, we assemble and implement the RHEPS, which here comprises NOAA’s Hydrology Laboratory Research Distributed Hydrologic Model.
(HL-RDHM), forced by ensemble precipitation, and near-surface temperature outputs from the GEFSRv2 (Hamill et al. 2013; Siddique et al. 2015). The RHEPS is used in this study to produce and verify ensemble streamflow forecasts for lead times from 6 to 168 h across eight river basins of varying spatial scales in the MAR. Additionally, a computationally efficient, operationally relevant statistical postprocessor is applied to correct the biases in the streamflow forecasts. The study area, details about the selected case study basins, and the datasets used are discussed in section 2. In section 3, we describe the methods used, including the distributed hydrological model, statistical postprocessor, and verification strategy. The main results are summarized in section 4. Section 5 discusses the study findings and some limitations. Finally, in section 6, the key conclusions are outlined.

2. Study area and data

a. Study area

The MAR is selected as the study area (Fig. 1). Streamflow forecasting is crucially relevant in the MAR because of its high population density, large cities, and high frequency (relative to other parts of the United States) of extreme precipitation events (Hitchens et al. 2013; Jones et al. 1997; Siddique et al. 2015). Moreover, the MAR is highly dependent on streamflow since a major share of its total water withdrawals (~90%) are from riverine (streamflow) sources, as opposed to groundwater sources (Maupin et al. 2014). In the MAR, eight river basins are selected (Fig. 1), ranging in drainage area from ~262 to 29,965 km² and covering the major rivers in the MAR, including the Delaware, James, Potomac, and Susquehanna Rivers. Table 1 summarizes the key characteristics of the selected river basins.

For each major river in the MAR, one large basin and a smaller, nested subbasin are selected in order to account for the effect of spatial scale when implementing the RHEPS and verifying the quality of its streamflow forecasts. For example, the large basin for the Delaware River has a drainage area of 17,574 km² while its nested subbasin is only 860 km² (Table 1). All of the selected basins are gauged by the U.S. Geological Survey (USGS) and represent forecast points used by the Middle Atlantic River Forecast Center (MARFC) to produce daily flow forecasts and communicate them to the public. The USGS gauge ID associated with each basin is included in Table 1.

b. Data

1) FORECASTS

As part of the RHEPS, we use ensemble meteorological forecasts (precipitation and near-surface temperature) from the GEFSRv2 to force HL-RDHM. GEFSRv2 uses the NCEP Global Ensemble Forecast System (GEFS) model (version 9.0.1) to produce retrospective forecasts or reforecasts across the globe for 16 days of lead time. The system consists of 11 ensemble members, one of which is an unperturbed (control) member, and the rest are perturbed members generated with perturbed initial conditions. From days +1 to +8, it uses the T254L42 model resolution, which runs at a spatial grid of 0.5° or ~55 km. From days +9 to +16, the
Table 1. Main characteristics of the eight study basins.

<table>
<thead>
<tr>
<th>Location of outlet</th>
<th>USGS ID</th>
<th>Area (km²)</th>
<th>Lat</th>
<th>Lon</th>
<th>Min daily flowa (m³ s⁻¹)</th>
<th>Max daily flowa (m³ s⁻¹)</th>
<th>Mean daily flowa (m³ s⁻¹)</th>
<th>Climatological flow (Pr = 0.5)b (m³ s⁻¹)</th>
<th>Climatological flow (Pr = 0.9)b (m³ s⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delaware</td>
<td>WAL-N</td>
<td>0423900</td>
<td>39.09</td>
<td>-75.06</td>
<td>12.84 (12.92)</td>
<td>71.39 (109.92)</td>
<td>27.29 (21.60)</td>
<td>36.41 (32.00)</td>
<td>45.82 (37.40)</td>
</tr>
<tr>
<td></td>
<td>WAL-NO</td>
<td>0423800</td>
<td>39.06</td>
<td>-75.03</td>
<td>21.08 (21.08)</td>
<td>30.87 (61.00)</td>
<td>15.58 (15.58)</td>
<td>18.93 (24.20)</td>
<td>24.01 (24.01)</td>
</tr>
<tr>
<td></td>
<td>TREN4</td>
<td>04146300</td>
<td>40.20</td>
<td>-74.87</td>
<td>14.08 (12.92)</td>
<td>41.39 (41.39)</td>
<td>16.08 (16.08)</td>
<td>28.01 (28.01)</td>
<td>37.00 (37.00)</td>
</tr>
<tr>
<td></td>
<td>SHBN8</td>
<td>0415800</td>
<td>39.90</td>
<td>-75.00</td>
<td>15.08 (15.08)</td>
<td>75.08 (75.08)</td>
<td>25.08 (25.08)</td>
<td>40.00 (40.00)</td>
<td>70.00 (70.00)</td>
</tr>
<tr>
<td></td>
<td>WYNN</td>
<td>04145800</td>
<td>40.20</td>
<td>-74.90</td>
<td>16.08 (16.08)</td>
<td>76.08 (76.08)</td>
<td>26.08 (26.08)</td>
<td>41.00 (41.00)</td>
<td>71.00 (71.00)</td>
</tr>
<tr>
<td></td>
<td>WYNX</td>
<td>04145900</td>
<td>40.20</td>
<td>-74.90</td>
<td>17.08 (17.08)</td>
<td>77.08 (77.08)</td>
<td>27.08 (27.08)</td>
<td>42.00 (42.00)</td>
<td>72.00 (72.00)</td>
</tr>
<tr>
<td></td>
<td>YYXN</td>
<td>04146000</td>
<td>40.20</td>
<td>-74.90</td>
<td>18.08 (18.08)</td>
<td>78.08 (78.08)</td>
<td>28.08 (28.08)</td>
<td>43.00 (43.00)</td>
<td>73.00 (73.00)</td>
</tr>
</tbody>
</table>

Notes:
- a The number in parentheses is the historical (based on entire available record, as opposed to the period 2004–11 used in this study) daily min, max, or mean recorded flow.
- b Pr = 0.5 indicates flows with exceedance probability of 0.5 and Pr = 0.9 indicates flows with exceedance probability of 0.10.

2) Observations

We use multisensor precipitation estimates (MPEs) as the observed precipitation data to calibrate the hydrological model, perform the model simulation runs, and initialize the forecasting system. MPEs are produced hourly through the optimal combination of multiple radars and hourly rain gauge data at 4 × 4 km² grid resolution (Rafieeinasab et al. 2015a; Zhang et al. 2011). The MPE product used here was obtained from the MARFC and is similar to the NCEP stage IV MPEs (Moore et al. 2015; Prat and Nelson 2015). At NOAA’s River Forecast Centers, MPEs are routinely monitored and quality controlled for different hydrological modeling applications, including streamflow forecasting (Lin and Mitchell 2005). Gridded MPE products are now widely used in verification studies (Habib et al. 2012; Sharma et al. 2017; Siddique et al. 2015), hydrological modeling (Kitzmiller et al. 2011; Rafieeinasab et al. 2015b), and data assimilation (Lee et al. 2011; Lin and Mitchell 2005; Rafieeinasab et al. 2014). HL-RDHM requires gridded temperature observations to obtain monthly potential evaporation and, as input to the SNOW-17 model, to determine snow accumulation and melt. The gridded temperature data were obtained from the MARFC, which generated the data by combining multiple observation networks (METAR, USGS stations, and NWS Cooperative Observer Program). All the gridded data used in this study were resampled using bilinear interpolation onto the regularly spaced grid (4 × 4 km² cell size) required by HL-RDHM. For the verification of the streamflow simulation and forecasts, daily discharge data from the relevant USGS gauges (Table 1) were used. In total, 10 years (2004–13) of streamflow observations were used.

3. Methods

In this study, the RHEPS is composed of the following four main components: (i) meteorological forecasts resolution is changed to T190L42, which runs at a 0.67° resolution or ~73 km. This variable resolution approach is also implemented by the ECMWF (Buizza et al. 2007). For GEFSRv2, the model is initiated each day at 0000 UTC to produce reforecasts for the next 16 days of lead time. From days +1 to +3, 3-hourly forecast accumulations are available; after that, forecasts are saved every 6 h, providing 6-hourly accumulations of forecasts from days +4 to +16. In total, more than 30 years of reforecast data are archived (from 1984 to present) for a large number of selected meteorological variables. Further details about the GEFSRv2 are discussed by Hamill et al. (2013).
(precipitation and near-surface temperature ensembles), (ii) distributed hydrological model, (iii) statistical postprocessor, and (iv) verification strategy. This section describes the latter three components since the meteorological ensembles were described in the previous section.

a. Distributed hydrological model

NOAA’s HL-RDHM is used as the distributed hydrological model (Koren et al. 2004). Recent applications of HL-RDHM (Lee et al. 2015; Rafieeinasab et al. 2015b; Spies et al. 2015; Thorstensen et al. 2016; Wood et al. 2016) as well as further details about the model (Burnash and Singh 1995; Burnash et al. 1973; Koren et al. 2004; Sorooshian and Gupta 1983) are discussed elsewhere. Within HL-RDHM, we implement the heat transfer version of the Sacramento Soil Moisture Accounting model (SAC-HT) to represent rainfall–runoff generation, and the SNOW-17 model to represent snow accumulation and melt (Koren et al. 2006).

SAC-HT is a physics-based, conceptual model where the basin system is divided into regularly spaced, square grid cells to account for spatial heterogeneity and variability. Each grid cell, in turn, is composed of storage components that store and transmit water. The cells are ultimately connected to each other through the stream network system, that is, each cell acts as a hillslope capable of generating surface and subsurface runoff that discharges directly into the streams. Specifically, HL-RDHM assumes at each grid cell a two-layer subsurface soil moisture storage called the upper and lower zone, as well as a surface interception and groundwater storage. In each of these layers, soil moisture can be stored in two ways, tension or free water, and surface runoff occurs with upper-zone saturation excess. Through the SNOW-17 submodel, each cell can also accumulate snow and generate hillslope snowmelt based on the near-surface temperature. The hillslope runoff, generated at each grid cell by SAC-HT and SNOW-17, is routed to the stream network using a nonlinear kinematic wave algorithm (Koren et al. 2004; Smith et al. 2012). Similarly, flows in the stream network are routed downstream using a nonlinear kinematic wave that accounts for parameterized stream cross-sectional shapes (Koren et al. 2004; Smith et al. 2012). Here we run HL-RDHM in a fully distributed manner at a spatial resolution of $2 \times 2$ km$^2$. Note that HL-RDHM requires the forcing to be input at the $4 \times 4$ km$^2$ resolution, but the model itself can actually be run at different resolutions. The $2 \times 2$ km$^2$ resolution mainly allows for a more realistic representation of the stream network.

To calibrate HL-RDHM, the model is first run using a priori parameter estimates previously derived from available datasets (Anderson et al. 2006; Koren et al. 2000; Reed et al. 2004). To calibrate the a priori parameters, we select 10 out of the 17 SAC-HT parameters based upon prior experience and preliminary parameter sensitivity tests. To calibrate the selected HL-RDHM parameters, the a priori parameter fields are adjusted manually first; once the manual changes do not yield noticeable improvements in the model performance, the parameter values are tuned up using an automatic technique, namely, stepwise line search (SLS; Kuzmin et al. 2008; Kuzmin 2009). SLS is used since this method is readily available within HL-RDHM and has been shown to provide reliable parameter estimates (Kuzmin et al. 2008; Kuzmin 2009). With SLS, the following objective function $J$ is optimized:

$$J = \sqrt{\sum_{i=1}^{m} (Y_i - S_i(\Phi))^2}, \quad (1)$$

where $Y_i$ and $S_i$ denote the daily observed and simulated flows, respectively; $\Phi$ is the parameter set being estimated; and $m$ is the total number of days used for calibration.

We use 3 years (2004–06) and 1 year (2004) of streamflow data to calibrate the small and large basins, respectively. A short calibration period is used to ameliorate computational demand. To assess the model performance during calibration, the correlation coefficient $R$, modified correlation coefficient $R_m$, percent bias (PB), and Nash–Sutcliffe efficiency (NSE) are used (see appendix for their mathematical definition). To clarify, these metrics are used during the manual phase of the calibration process to assess the model performance and then to assess the final results from the implementation of the SLS, but the actual implementation of the SLS is based on the objective function in Eq. (1). We note that, with our calibration approach, most of the performance gain was achieved during the manual phase of the calibration, while the use of SLS only yielded marginal improvements. Thus, the use of Eq. (1) is not unduly biasing the calibration results toward the high flows.

b. Statistical postprocessor

To statistically postprocess the ensemble streamflow forecasts (i.e., to quantify the streamflow uncertainty and adjust forecast biases), the so-called hydrological model output statistics (HMOS) approach is implemented (Regonda et al. 2013). Similar approaches are common and widely used in weather forecasting (Glahn and Lowry 1972; Hamill et al. 2004; Wilks 2015). HMOS is selected since it can be applied in an operational setting, but other more sophisticated postprocessors are available. The general goal with HMOS is to statistically...
correct or improve current forecasts by treating them as predictands in a regression model that depends on variables associated with past forecasts and simulations. As the HMOS postprocessor, we use a first-order autoregressive model in normal space, with a single exogenous variable, similar to the approach by Regonda et al. (2013).

For each ensemble member, the postprocessing model is as follows:

\[ Z_{k+1}^0 = (1 - b_{k+1})Z_k^0 + b_{k+1}Z_{k+1}^0 + E_{k+1}, \]  

where \( Z_k^0 \) and \( Z_{k+1}^0 \) denote the normalized observed flow at times \( k \) and \( k + 1 \), respectively; \( Z_{k+1}^0 \) is the normalized forecast flow at time \( k + 1 \); \( b_{k+1} \) denotes the weight given to the forecast at time \( k + 1 \); and \( E_{k+1} \) denotes the residual error at time \( k + 1 \). For the above model, assuming that there is significant correlation between \( E_{k+1} \) and \( Z_{k+1}^0 \)

\[ E_{k+1} = \frac{\sigma_{E_k}}{\sigma_{E_k}} \rho(E_{k+1}, E_k)E_k + W_{k+1}, \]  

where \( \sigma_{E_k} \) and \( \sigma_{E_{k+1}} \) denote the standard deviation of \( E_k \) and \( E_{k+1} \), respectively; \( \rho(E_{k+1}, E_k) \) denotes the serial correlation between \( E_{k+1} \) and \( E_k \); and \( W_{k+1} \) is a random error generated from the normal distribution \( \mathcal{N}(0, \sigma_{W_{k+1}}^2) \).

To estimate the parameter \( \sigma_{W_{k+1}}^2 \), we use the following:

\[ \sigma_{W_{k+1}}^2 = [1 - \rho^2(E_{k+1}, E_k)]\sigma_{E_k}^2. \]  

The step-by-step procedure for implementing the postprocessor [Eqs. (2)–(4)] is as follows:

(i) Past forecasts for each lead time and corresponding observations are assembled and transformed into standard normal deviates using the normal quantile transformation (NOT; Krzysztofowicz 1997). In this study, 8 years of forecast and observation data are used as the training period, following a leave-one-out approach (i.e., the entire 10 years of forecasts are postprocessed by selecting 8 years to train and 2 to postprocess/verify).

(ii) Ten equally spaced values of \( b_{k+1} \) within 0.1–0.9 are selected.

(iii) For each \( b_{k+1} \), \( \sigma_{W_{k+1}}^2 \) is calculated from Eq. (4) using the training data to estimate the parameters in Eq. (4).

(iv) Variable \( W_{k+1} \) is generated from \( \mathcal{N}(0, \sigma_{W_{k+1}}^2) \) and \( E_{k+1} \) is calculated from Eq. (3).

(v) A trace of \( Z_{k+1}^0 \) from Eq. (2) is generated and transformed back to real space using the inverse NOT.

(vi) Steps iii–v are repeated to generate \( N \) number of postprocessed ensemble traces, 10 traces per each raw streamflow ensemble.

(vii) By repeating steps iii–vi, ensemble streamflow forecasts are generated for all the selected \( b_{k+1} \) values. The mean continuous ranked probability score (CRPS; see appendix for mathematical definition) is calculated separately for each \( b_{k+1} \), and the value of \( b_{k+1} \) that produces the smallest mean CRPS is selected.

The above postprocessing procedure is applied at each individual lead time. For lead times beyond the initial one (day 1), 1-day-ahead predictions are used as the observed streamflows. For the cases where \( Z_{k+1}^0 \) falls beyond the historical maxima or minima, extrapolation is used to model the tails of the forecast distribution. For the upper tail (high flows), a hyperbolic distribution (Journel and Huijbregts 1978) is used while linear extrapolations are used for the lower tail (low flows).

### c. Verification strategy

The verification of the ensemble streamflow forecasts is done using the Ensemble Verification System (EVS; Brown et al. 2010). The EVS is a comprehensive and modular verification tool developed by Brown et al. (2010) for the NWS to facilitate the verification of different ensemble forecast variables (Brown 2014; Brown et al. 2014; Sharma et al. 2017; Siddique et al. 2015). We use for the verification 6-hourly streamflow forecasts and daily observed streamflows at the outlet of each of the eight selected basins. The verification is conditioned upon the lead time, streamflow threshold, and season. The relative mean error (RME), Brier skill score (BSS), and continuous ranked probability skill score (CRPSS) are employed as the verification metrics (see appendix for their mathematical definition).

For the verification analysis, we generate and verify two different datasets of ensemble streamflow forecasts, namely, raw (without postprocessing) calibrated and postprocessed. To verify the raw calibrated ensemble forecasts across lead times of 1–7 days, 10 years of data (2004–13) are used. To verify the postprocessed ensemble forecasts, as indicated before, a leave-one-out approach is implemented by using 8 years of data to train the postprocessor and the remaining 2 years to verify the forecasts. Both streamflow forecast datasets, raw and postprocessed, are verified against observed and simulated streamflows to assess and contrast different sources of uncertainty. Note that hydrological model runs forced with meteorological forecasts contain both meteorological and hydrological uncertainties while simulated streamflows (i.e., model runs forced with the observed forcings) are characterized by hydrological uncertainty. In other words, ensemble streamflow forecasts verified against simulated...
streamflows provide a measure of meteorological uncertainty, as opposed to the total uncertainty, which is measured here by ensemble streamflow forecasts that are verified against observed streamflows. Note that our definition of hydrological uncertainty includes the overall uncertainty associated with the observed forcings and the hydrological model (initial conditions, model structure, and parameters).

4. Results

This section is subdivided into the following three main subsections: verification of simulated streamflows, verification of raw ensemble streamflow forecasts, and verification of postprocessed ensemble streamflow forecasts. The results associated with each subsection are separated into low to moderate and high flows to verify the performance of the RHEPS under these two different flow conditions. The low- to moderate-flow category represents flows with a nonexceedance probability (Pr) of 0.50, while the high-flow category is for Pr = 0.90 (i.e., flows with exceedance probability less than 0.1 are denoted as high). These flow thresholds are selected to represent baseflow conditions (in the case of the low to moderate flows) and to represent a range of flood conditions (in the case of the high flows).

a. Verification of the simulated streamflows

The main results associated with the performance of the simulated streamflows for the entire period of analysis (2004–13) and the eight selected basins are summarized in Table 2. Further, the results in Table 2 are based on various performance metrics ($R$, $R_m$, PB, and NSE, as defined in the appendix) and are separated according to uncalibrated and calibrated simulation runs, as well as low- to moderate- and high-flow conditions. Note that the simulated, as opposed to forecasted, streamflows are obtained by forcing HL-RDHM with observed precipitation and near-surface temperature data.

1) LOW TO MODERATE FLOWS

The correlation coefficient between the simulated and observed low to moderate flows tends to be greater for the larger basins. For example, $R = 0.81$ in the large basin of the Delaware River (TREN4) for the uncalibrated simulation run while the small basin (WALN6) has a coefficient of 0.76 for the same run (Table 2). The overall improvement in $R$, averaged across the eight basins, between the uncalibrated and calibrated simulation runs is ~7%, but it can be as high as 40% in the case of SHBN6 (Table 2). The modified correlation coefficient is also computed since it is better than $R$ in

<table>
<thead>
<tr>
<th>Performance statistic</th>
<th>Model run</th>
<th>Flow condition</th>
<th>Delaware</th>
<th>Susquehanna (north branch)</th>
<th>Potomac</th>
<th>James</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>WALN6</td>
<td>SHBN6</td>
<td>WVYN6</td>
<td>DAWM2</td>
</tr>
<tr>
<td>$R$</td>
<td>Uncalibrated</td>
<td>Low to moderate</td>
<td>0.76</td>
<td>0.59</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.74</td>
<td>0.76</td>
<td>0.55</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>0.86</td>
<td>0.85</td>
<td>0.68</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>0.86</td>
<td>0.87</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Calibrated</td>
<td>Low to moderate</td>
<td>0.76</td>
<td>0.87</td>
<td>0.82</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>0.88</td>
<td>0.93</td>
<td>0.86</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>0.88</td>
<td>0.93</td>
<td>0.86</td>
<td>0.92</td>
</tr>
<tr>
<td>$R_m$</td>
<td>Uncalibrated</td>
<td>Low to moderate</td>
<td>0.74</td>
<td>0.59</td>
<td>0.46</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
<td>0.50</td>
<td>0.39</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>0.60</td>
<td>0.61</td>
<td>0.46</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>0.60</td>
<td>0.74</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Calibrated</td>
<td>Low to moderate</td>
<td>0.74</td>
<td>0.80</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>0.67</td>
<td>0.89</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>0.70</td>
<td>0.74</td>
<td>0.84</td>
<td>0.90</td>
</tr>
<tr>
<td>$\text{PB}$</td>
<td>Uncalibrated</td>
<td>Low to moderate</td>
<td>8.94</td>
<td>−5.67</td>
<td>20.01</td>
<td>27.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>−23.39</td>
<td>−4.94</td>
<td>−34.65</td>
<td>15.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>−3.99</td>
<td>−0.74</td>
<td>4.82</td>
<td>20.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>6.95</td>
<td>−4.74</td>
<td>−0.67</td>
<td>2.99</td>
</tr>
<tr>
<td></td>
<td>Calibrated</td>
<td>Low to moderate</td>
<td>−21.73</td>
<td>−5.57</td>
<td>−5.15</td>
<td>−12.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>−4.52</td>
<td>−1.52</td>
<td>0.88</td>
<td>−1.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>0.52</td>
<td>−0.21</td>
<td>−0.49</td>
<td>−0.50</td>
</tr>
<tr>
<td></td>
<td>$\text{NSE}$</td>
<td>Uncalibrated</td>
<td>0.44</td>
<td>−0.07</td>
<td>0.18</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.71</td>
<td>0.42</td>
<td>0.46</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Calibrated</td>
<td>Low to moderate</td>
<td>0.52</td>
<td>0.69</td>
<td>0.67</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>0.58</td>
<td>0.57</td>
<td>0.66</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>0.77</td>
<td>0.78</td>
<td>0.71</td>
<td>0.85</td>
</tr>
</tbody>
</table>
accounting for hydrograph shape and size (McCuen and Snyder 1975; Smith et al. 2004). Based on the value of $R_m$ for the selected basins (Table 2), the improvement after calibration is on average ~17%.

In terms of the NSE, there is a large gain in performance between the calibrated and uncalibrated simulation runs. For example, the largest improvement is seen in SHBN6, where the NSE increases from 0.21 to 0.69 after calibration (Table 2). This large improvement seems related to difficulties in obtaining reliable a priori parameters for this basin, which is likely affected by karst geology (Reed et al. 2006; Tang et al. 2007). During the calibration process, other than the routing parameters, the upper-zone free water capacity (UZFWM) and the fractional daily upper-zone free water withdrawal rate (UZK) are the two parameters that showed the most variability from basin to basin. In the case of SHBN6, both parameters (UZFWM and UZK) showed great ability to improve the simulation performance, more than in the other basins. This may be because of the presence of karst in SHBN6, since karst systems act like underground pipes to transport subsurface water, particularly affecting interflow conditions that in HL-RHDM are controlled by the parameters UZFWM and UZK.

The smallest gain in the NSE is seen in RMDV2. In this case, after calibration, the NSE changes from 0.53 to 0.57 (Table 2), indicating that the historical data used for calibration may not be informative enough to affect the parameter response. It is also possible that one or more parameters in this basin are seasonal in nature, affecting the ability of the calibration to achieve improved parameter values. In one of the eight basins, WVYN6, the NSE decreases slightly after calibration for the low to moderate flows (from 0.49 to 0.41), but the same basin shows overall (i.e., including all flows) a gain in performance (from 0.66 to 0.74). Thus, the decrease in the value of NSE for the low to moderate flows in WVYN6 is likely due to tradeoffs in some of the parameter values. Indeed, for the eight selected basins, the performance of the calibrated simulation runs is overall (i.e., including all flows) satisfactory, with NSE values ranging from 0.68 to 0.86 (Table 2). Additionally, there is a large improvement in the percent bias of the calibrated simulation runs. For instance, the PB for the uncalibrated runs in the Potomac River basin are 20.01% for DAWM2 and 27.67% for BRKM2 (Table 2), which are reduced to −0.67% and 2.99%, respectively, in the calibrated runs.

2) High flows

Comparing the performance of the calibrated runs against the uncalibrated ones for the high flows, the values of $R$ and $R_m$ mostly improve, in a few cases they stay relatively the same. For example, the value of $R_m$ in the Potomac River basin increased from 0.29 to 0.79 at DAWM2 and from 0.70 to 0.87 at BRKM2, while the value of $R$ stayed nearly the same at PYAV2 in the James River basin (Table 2). Using the value of $R_m$ for the calibrated runs to contrast the performance of the high and low to moderate flows, the high flows tend to outperform the low to moderate flows, but in a few basins (WALN6, SHBN6, and RMDV2 in Table 2) the low to moderate flows perform better. To further understand this, we examined the simulation runs (i.e., individual hydrographs) and noticed that some of the high-flow events, mostly during the winter months (November–April), are somewhat underestimated. Thus, incorporating additional data when implementing SNOW-17 could, in the future, contribute to improving the performance of the winter high flows in these basins.

The NSE value for the high flows, averaged across all the selected basins, improves from 0.31 with the uncalibrated runs to 0.58 with the calibrated ones. However, the NSE values for the overall flow conditions (i.e., including all flows) are higher; they improve on average from 0.64 to 0.77. Based on the NSE values, the high flows perform better than the low to moderate flows, which is expected since the NSE is more favorable to high flows than low flows. Further, as was the case with the low to moderate flows, the uncalibrated runs for the high flows tend to show some unusually high PB values. For example, PB = −34.65% at DAWM2 in the Potomac River basin, which after calibration is reduced to −5.15%. Ultimately, the PB values for the calibration runs are satisfactory, ranging from 0.88% to −5.54% (Table 2). The overall performance of the calibrated simulation runs in this study compare well with results from previous studies using the same model and region (Tang et al. 2006; van Werkhoven et al. 2008) as well as with HL-RDHM performance statistics in other regions (Mejia and Reed 2011; Reed et al. 2004).

b. Verification of the raw ensemble streamflow forecasts

This subsection presents and discusses the verification results for the raw (without postprocessing) ensemble streamflow forecasts, generated by forcing the calibrated HL-RDHM with the GEFS precipitation and near-surface temperature reforecasts. To verify the raw ensemble streamflow forecasts, the RME, CRPSS, and BSS (see appendix) are used as the verification metrics for the period 2004–13. For each of the three verification metrics used, two different versions of the same metric are computed, one using observed flows as the reference and another one using simulated flows.
1) LOW TO MODERATE FLOWS

We use the RME to quantify the flow forecast error (see appendix). A negative RME indicates the presence of an underforecasting bias while a positive RME indicates overforecasting bias. For our selected basins, the RME exhibits mostly a negative bias whose absolute value increases with the lead time (Fig. 2). This result is in agreement with previous findings (Siddique et al. 2015) that demonstrate that precipitation forecasts from the GEFSRv2 across the MAR are consistently underforecasted for 1–7 days of lead time. Moreover, this is the case for both RMEs (i.e., relative to observed flows as well as relative to simulated flows), with the exception of the Potomac River basin (BRKM2, Fig. 2e) that shows a positive bias that increases with the lead time relative to the simulated flows. Further, most of the basins show a relatively small difference between the two RMEs (except WALN6 and BRKM2 in Figs. 2c and 2e, respectively). This indicates that the effect of the hydrological uncertainty on the RME of low- to moderate-flow forecasts is for the most part small compared to the effect of the meteorological uncertainty. This can be explained further from a modeling perspective. In each grid cell of the model, the lower-zone soil is mostly responsible for the generation of base flow (approximately similar to our low to moderate flows), which is slow responding in nature and is generally well represented in the model structure.

FIG. 2. RME of the mean raw forecasts: low- to moderate-flow conditions. The RME is shown for the raw mean ensemble streamflow forecasts vs the forecast lead time, with reference to both observed (solid line) and simulated (dashed line) flows. Results are shown for (a)–(d) small and (e)–(h) large basins, under low- to moderate-flow conditions (flows with nonexceedance probability of 0.5).
Raw forecasts: Low–moderate flow conditions
Small basins (Area < 2000 km²)

DAWM2 (Potomac)  SHBN6 (Susquehanna)  WALN6 (Delaware)  PYAV2 (James)

Large basins (Area > 12000 km²)

BRKM2 (Potomac)  WVYN6 (Susquehanna)  TREN4 (Delaware)  RMDV2 (James)

FIG. 3. CRPSS of the raw forecasts: low- to moderate-flow conditions. The CRPSS is shown for the raw ensemble streamflow forecasts vs the forecast lead time, with reference to both the observed (solid line) and simulated (dashed line) climatological flows. Results are shown for (a)–(d) small and (e)–(h) large basins, under low- to moderate-flow conditions (flows with nonexceedance probability of 0.5).

To measure the skill of the raw ensemble flow forecasts, the CRPSS is used (see appendix). A CRPSS value of zero means no skill (i.e., same skill as the reference system) and a CRPSS value of one indicates maximum skill. As was the case with the RME, the CRPSS values are computed with reference to both the observed and simulated climatological flows (Fig. 3). In all the selected basins, except DAWM2 (Fig. 3a), the CRPSS shows that the skill of the low- to moderate-flow forecasts, with reference to the simulated flows, is high for the initial lead times, but it gradually declines as the lead time increases. The CRPSS may be low for DAWM2, among other possibilities, because this is a small basin and meteorological forecast skill tends to decrease considerably with decreasing spatial scale or basin size (Sharma et al. 2017; Siddique et al. 2015). Furthermore, by comparing the two versions of the CRPSS metric (i.e., the solid line against the dashed line in Fig. 3) for all the basins, we find that hydrological uncertainty is relatively dominant for the initial lead times (days 1–3), but it becomes less dominant as the lead time increases (days 6–7). The relatively large hydrological uncertainty at the initial lead times may be related to the value of the model parameters. It is well known that distributed hydrological models are affected by the lack of direct parameter observability, which leads to issues of parameter identifiability and equifinality. This will have an impact on the model initial conditions.

Also, at the initial lead times, the skill of the forecasts with reference to the observed flows is generally low.
relative to the simulated flows. For instance, at a lead time of 1 day in Fig. 3d, the CRPSS is only \(0.3\) with reference to the observed flows but jumps to \(0.9\) with reference to the simulated flows. This highlights the skill of the meteorological ensembles at the initial lead times, whose uncertainty becomes dominant at the longer lead times, as suggested by the tendency of the two CRPSS metrics to converge toward each other at the 7-day lead time (see, e.g., Figs. 3g,h).

To assess the skill of the raw seasonal flow forecasts, the BSS is determined for the “dry” (including the months of June–November) and “wet” (including the months of December–May) season for each of the eight selected basins, under low- to moderate-flow conditions (Fig. 4). The BSS is computed from the Brier score (BS), which is analogous to the mean-square error of the forecasts (see appendix). A BSS of one implies perfect skill and a BSS of zero implies no skill. As was the case with the other metrics, the BSS is shown here with reference to both simulated and observed climatological flows (Fig. 4). As expected from our previous results (e.g., Fig. 3), the skill of the seasonal forecasts tends to decline with increasing lead time (Fig. 4). The skill declines more rapidly when measured relative to the simulated flows, as opposed to observed, highlighting that hydrological uncertainty strongly affects forecast skill at
the initial lead times and, at longer lead times, meteorological uncertainty becomes a more dominant factor in determining seasonal forecast skill. For example, for the large basin in the Potomac River (BRKM2, Fig. 4e) and the wet season, the BSS has a value of 0.9 and 0.45 at a lead time of 1 day relative to the simulated and observed flows, respectively, but these values decrease to 0.4 and 0.35, respectively, at a lead time of 7 days (Fig. 4e). The dry season forecasts tend to display similar (e.g., PYAV2 and BRKM2 in Figs. 4d and 4e, respectively) or slightly better skill than the wet season ones relative to both observed and simulated flows. Overall, the seasonal values of the BSS are similar across the selected basins. DAWM2 (Fig. 4a) seems to be the only exception, exhibiting a notably low skill relative to the other basins, suggesting that basin size may be an important factor in determining seasonal skill. Note that DAWM2 is the basin with the smallest drainage area, 262 km², out of the eight basins considered (Table 1).

2) HIGH FLOWS

In relation to the high flows, the RME indicates mainly underforecasting across the selected basins (Fig. 5). BRKM2 (Fig. 5e) is the only exception, which is characterized by a strong positive bias when the simulated flows are used as reference. The underestimation is <30% across the initial lead times (days 1–3) and in some basins is much smaller than that; however, it increases with increasing lead time. The high flows, irrespective of the basin size, exhibit a greater influence of
meteorological uncertainty than the low to moderate flows. For instance, in Fig. 5c, the RME for the high flows (relative to the simulated ones) is approximately −55% at the day 7 lead time. In contrast, the RME for the same basin and lead time during low- to moderate-flow conditions is only −20% (Fig. 2c). Thus, the propagation of the meteorological uncertainties to the hydrological predictions has greater potential to influence the quality of the high-flow forecasts than the low to moderate ones. This is not surprising since the high flows result from the direct response of the basin to the precipitation events, whereas the low to moderate flows are in this case dominated by subsurface processes and only indirectly by the precipitation events. One difference between the high- and low- to moderate-flow forecasts is that the RME for the high flows increases more rapidly with increasing lead time compared to the low to moderate flows. Specifically, the bias in the high-flow forecasts increases rapidly during the longer lead times (Fundel et al. 2013). This implies that the high-flow forecasts can particularly benefit from a good representation of model states and skillful initial conditions.

As expected, the high-flow forecasts are less skillful as the forecast lead time increases (Fig. 6). As was the case with the low to moderate flows, at the initial lead times (1–3 days), the CRPSS shows that the major source of uncertainty for the high flows is hydrological (Fig. 6). Hydrological uncertainty becomes less pronounced and meteorological uncertainty starts to dominate as the lead time grows (>3 days). Overall, the skill of the high-flow forecasts across the selected basins is similar. For instance,
with the exception of DAWM2 (Fig. 6a) and PYAV2 (Fig. 6d), and relative to the observed flows, the CRPSS tends to be between 0.4 and 0.6 at a lead time of 1 day and between 0.1 and 0.3 at a lead time of 7 days. The results also show that, in most cases, meteorological uncertainty has a greater effect on the small basins compared to the large ones. Accordingly, there is a tendency for the large basins to show slightly better flow forecast skill than the small ones. For instance, in the Delaware River, the large basin (Fig. 6g) has a skill of 0.6 at a lead time of 1 day and the small basin (Fig. 6c) has a skill of 0.43 at the same lead time. This gain in skill with basin size is, however, due to both improvements in the performance of the meteorological forecasts (Siddique et al. 2015) and hydrological model (Table 2). Interestingly, the only selected basins that do not follow this scaling trend are the ones in the Susquehanna River basin (Figs. 6b,f), where the overall skill (i.e., relative to the observed flows) is higher in the small basin than in the large one. This indicates that the large basin is in this case subject to high hydrological uncertainty. This uncertainty may be due to inaccurate model initial conditions and large parametric uncertainty.

The skill of the raw, seasonal ensemble flow forecasts is illustrated in Fig. 7. In Fig. 7, the BSS values are shown for high-flow conditions, according to the dry (June–November) and wet (December–May) months, and with
reference to both observed and simulated climatological flows. In the wet season, relative to the observed flows (i.e., accounting for the total uncertainty), the forecasts tend to be as or slightly more skillful than in the dry season, except for SHBN6 and PYAV2 (Figs. 7b,d), which are both small basins. Relative to the simulated flows (i.e., emphasizing meteorological uncertainty), the situation reverses and the forecasts exhibit slightly better skill in the dry season compared to the wet season (Fig. 7). This suggests that hydrological uncertainty tends, particularly for the large basins, to be greater in the dry season than in the wet one. In some cases, this is more evident at the longer lead times (e.g., 6-day lead time in Fig. 7g where the distance between the dashed lines is wider than between the solid lines).

In general, the CRPSS indicates that streamflow ensembles in the MAR are more skillful than the deterministic forecasts (Fig. 8). The CRPSS values in Fig. 8 are computed with reference to the deterministic (unperturbed GEFSRv2 member) forecasts. The improvement is small across the initial lead times (<3 days); however, it increases with the lead time. This is the case for both low to moderate and high flows (Fig. 8). For instance, the CRPSS for both the low to moderate as well as the high flows in the large basin of the Potomac River (BRKM2, Fig. 8e) is slightly higher than zero at the 1-day lead time but rises by ~20% at the 7-day lead time. This highlights the fact that ensemble forecasting is particularly beneficial for medium-range predictions. The overall gain in skill between the ensemble and
Raw and postprocessed forecasts: Low–moderate and high flow conditions

Small basins (Area < 2000 km²)

Deterministic forecasts, from the day 1 to the day 7 lead time, is 10%–20% and 15%–40% for the low to moderate and high flows, respectively. It is also relevant to note that the ensemble forecasts show consistent improvements across all the basins in the MAR, despite their differences in hydroclimatic, landscape, and subsurface conditions.

c. Verification of the postprocessed ensemble streamflow forecasts

1) LOW TO MODERATE FLOWS

To verify the postprocessed ensemble forecasts for the low- to moderate-flow conditions, the RME is plotted for both the postprocessed and raw (without postprocessing) ensemble mean (Fig. 9). For the low to moderate flows, the RME indicates that postprocessing tends to reduce the forecast bias mostly in the large basins. For example, PYAV2 (Fig. 9d), BRKM2 (Fig. 9e), WVYN6 (Fig. 9f), and RMDV2 (Fig. 9h) demonstrate improved RME values relative to the raw forecast values. The most noticeable improvements are seen in WVYN6 and PYAV2, where the postprocessed forecasts are nearly unbiased across all the lead times. At lead times longer than 3 days, the postprocessor is unable to reduce the RME for DAWM2 (Fig. 9a), WALN6 (Fig. 9c), and BRKM2 (Fig. 9e). Besides identifying limitations in the postprocessor, this serves to diagnose flow conditions that...
could potentially benefit from improved hydrological modeling. For instance, the bias in the low flows could be partly due to the fact that low flows are regulated in some of the selected basins, for example, the large basin of the Delaware River, TREN4. If that turns out to be the case, it may be necessary to account for low-flow regulations to further improve the RME.

The CRPSS is used to investigate the skill of the postprocessed forecasts (Fig. 10). The CRPSS is computed in this case with reference to the raw ensemble forecasts. Overall, postprocessing improves the skill of the low to moderate flows at the initial lead times (<4 days) across basins; however, the level of improvement varies from basin to basin. It can be as low as 2% (WVYN6 at a lead time of 1 day, Fig. 10f) and as high as 25% (BRKM2 at a lead time of 1 day, Fig. 10e).

2) HIGH FLOWS

Postprocessing is more effective for the high flows than the low to moderate flows. For the high-flow conditions, the postprocessed ensemble forecasts show significant improvements relative to the raw ensemble forecasts across lead times (Figs. 9, 10). For instance, the raw ensemble forecast mean for DAWM2 (Fig. 9a) underestimates the observed one by ~60% at the 7-day lead time, while this underestimation is reduced to 15% after postprocessing. Similar improvements are seen in other basins, although improvements in the small basins tend to be smaller compared to the large basins. In terms
of the skill, the CRPSS shows that the skill of the high flows are consistently improved across lead times after postprocessing (Fig. 10), with the only exceptions being DAWM2 (Fig. 10a) and PYAV2 (Fig. 10d) at the longest lead time considered, 7 days. Postprocessing demonstrates, overall, significant potential for improving flow forecasts, particularly for the high-flow conditions.

5. Discussion

In this study, we demonstrate the use of weather ensembles from the GEFS to force a distributed hydrological model and generate regional ensemble streamflow forecasts. Such capabilities could be implemented in the United States by NOAA’s River Forecast Centers. Additionally, we clearly demonstrate the ability to improve the raw streamflow forecasts through postprocessing, particularly for the high flows, and the ability of the probabilistic streamflow forecasts to improve over the deterministic forecasts, particularly at the medium-range lead times. Our results show that, overall, hydrological uncertainty tends to dominate at the initial lead times (1–3 days) compared to the meteorological uncertainty. These results are dependent on the utilization of a hydrological model that has been calibrated prior to its implementation in forecasting mode. Although the results are not shown here for brevity, we forced the uncalibrated HL-RDHM with the GEFSRv2 reforecasts to find that, for most of the basins, the additional uncertainty associated with the uncalibrated forecast runs tended to be greater than the total uncertainty of the calibrated forecast runs. Recognizing that all hydrological models require parameterizations of one form or another, parameter calibration is relevant for improving and evaluating streamflow forecasts.

In terms of the modeling performed, the only human alteration that is explicitly accounted for in HL-RDHM is land-use/land-cover (LULC) change. Although LULC change can be an important driver of hydrological conditions, other human processes can play a part as well, such as water diversions, irrigation, urban storm water management, and reservoir operations, among others. In this case, water diversions seem less critical than in other U.S. regions where significant amounts of water are diverted for irrigation purposes. In our geographic region, agriculture is mainly fed by rain. Storm water management could be relevant for highly urbanized basins. In this study, only the small subbasin in the Potomac River basin, DAWM2, is highly urbanized. As mentioned earlier, it is likely that to improve streamflow forecasts in DAWM2, further information about its urban drainage patterns and storm water management infrastructure may be necessary. The benefits of including this additional information, however, need to be carefully balanced against the availability of high-quality, high-resolution weather ensembles.

Reservoir operations affect all of our large basins, that is, they all contain reservoirs that are used for water supply and water management purposes. Although the impact of reservoirs on our streamflow forecasts remains uncertain, we do not believe that they are significant. First, the majority of the reservoirs in the study area are mainly operated for water supply purposes, not for flood control. Thus, the reservoirs do little to affect or modify flood conditions. Second, in each of the large basins used, the total drainage area of the basin is much larger than the drainage area feeding the reservoirs. Thus, the effects of the reservoirs on streamflow are likely dissipated by the time the reservoir flows reach the overall outlet of the selected, large basins. Perhaps the only exception is the Delaware River basin, TREN4. There are several reservoirs in this basin that are specifically operated to manage low-flow conditions near the main outlet of TREN4. The main outlet of TREN4 is just upstream of one of the major water supply intakes for the city of Philadelphia. When riverine flows are too low in TREN4, additional water is released from some of the upstream reservoirs since the intake can otherwise be severely affected by high salinity concentrations from downstream tidal waters. It thus seems likely that our low- to moderate-flow forecasts for TREN4 could potentially benefit from considering reservoir operations. This should be investigated further in the future.

Our results highlight that improving ensemble weather predictions can go a long way in improving hydrological forecasts. For example, in terms of the RME, streamflow forecasts are consistently underforecasted, irrespective of the basin size and flow condition. Further, meteorological uncertainty has a greater impact on the RME compared to hydrological uncertainty for both low- to moderate- and high-flow conditions, particularly at the longer lead times (>3 days). In terms of skill, during both low- to moderate- and high-flow conditions, the CRPSS values (compared to simulated) for the small basins are lower than for the large basins. This is an indication that meteorological uncertainty is greater than the hydrological uncertainty in the smaller basins. This may be explained, in part, by the fact that the variability of the meteorological forcing is difficult to capture with meteorological forecasts that have a much coarser native resolution (~55 × 55 km²) than the size of the small basins considered (~262–1719 km²). Another relevant source of forecast uncertainty may be related to the season or time of the year. Results from the seasonal analysis varied widely for the small basins without a distinguishable trend or pattern. However, in the case of the large basins, the
forecasts tended to exhibit similar or better forecast skill during the wet season compared to the dry season, particularly for the high flows. The lower skill in the dry season could partly be due to challenges in predicting convective precipitation (Siddique et al. 2015), suggesting another area where improved weather predictions may contribute to enhancing streamflow forecasts.

An important aspect of forecasting that is not addressed in this study is the effect of uncertain observations on streamflow forecasts and their verification. In particular, MPEs are considered here as representative of the truth, but they are, in reality, subject to errors and systematic biases, particularly during the winter months. Furthermore, interpolation is used to map the GEFSRv2 data onto the finer $2 \times 2$ km$^2$ grid resolution of the hydrological model. This could be another source of systematic biases. Future studies are needed to quantify the impact of these sources of uncertainty on streamflow forecasts.

Based on our study results and somewhat relative to the United States, a key question that remains to be answered is which source of uncertainty or system component is more relevant, or should be prioritized, to improve operational forecasts. Recognizing that this is a question that cannot be answered by one study alone, some ideas are suggested. The short answer, although general, is that for the shorter lead times improvements in hydrological conditions appear particularly important, while for the longer lead times (>3 days) improvements in meteorological conditions seem relevant. Specifically, we highlight that having a poorly calibrated hydrological model will significantly undermine benefits (skill boost) from both weather ensembles and post-processing. We suggest that post-processing should always be applied since it does not need to be complicated, as demonstrated here, to reap benefits. Additionally, post-processing not only serves as a bias correction step, but it allows the quantification of forecast uncertainty. We believe the latter is a necessary step to better diagnose forecast quality and address, in more systematic ways, improvements to the hydrological forecasting system. The use of weather ensembles seems in this case particularly useful, in terms of skill gain, for the longer lead times. Thus, it would seem to follow that, as more emphasis is placed on the longer lead times, weather ensembles should play a bigger role in operational streamflow forecasting.

6. Summary and conclusions

In this study, using the RHEPS, we generated short- to medium-range (1–7 days) ensemble streamflow forecasts over the MAR. The RHEPS consisted of a distributed hydrological model, namely HL-RDHM, forced by GEFSRv2 ensemble reforecasts (precipitation and near-surface temperature). The ensemble streamflow forecasts were generated for a 10-yr period (2004–13) in eight river basins, encompassing some of the major rivers (Delaware, James, Potomac, and Susquehanna) in the MAR. For each of these rivers, we chose one large basin and a smaller, nested subbasin to consider the effect of spatial scale on the performance of the streamflow forecasts. To account for different sources of forecast uncertainty, the streamflow forecasts were verified relative to both simulated and observed flows. On the basis of the present implementation of the RHEPS, the following main conclusions are emphasized:

- The RME shows that the raw mean ensemble forecasts mostly underestimate the observed and simulated mean across lead times of 1–7 days, under both low- to moderate- and high-flow conditions. The underestimation increases with increasing lead time and the RME is lower in the large basins compared to the small ones.
- The CRPSS values for the raw ensemble streamflow forecasts imply that the skill of the meteorological forcing is relatively high for the initial lead time (1 day), but it decreases as the lead times increases. Thus, at lead times of 1–3 days, the raw ensembles seem largely affected by hydrological uncertainty. Across longer lead times (>3 days), hydrological uncertainty becomes less pronounced and meteorological uncertainty dominates. This trend is apparent in both the low to moderate and high flows.
- The raw ensemble streamflow forecasts exhibit weak seasonal trends. For the high-flow conditions, forecasts tend to have slightly better or similar skill in the wet season compared to the dry one. For the high-flow conditions and large basins, hydrological uncertainty seems to have a greater impact on forecasts in the dry season than the wet one.
- The smaller basins reveal greater meteorological uncertainty than the large ones, whereas hydrological uncertainty varies widely across basin sizes, even though the performance of the hydrological simulations is somewhat improved in the large basins. The latter highlights the need to benchmark both simulation and forecasting outputs from hydrological models, as done in this study, to fully understand and assess model performance.
- The raw ensemble streamflow forecasts show more skill than the deterministic (unperturbed GEFSRv2 member) forecasts across lead times of 1–7 days. The improvement is small at the initial lead time (day 1), but gradually increases with increasing forecast lead times.
- Results also show that postprocessing can improve the skill of streamflow forecasts over the raw streamflow forecasts.

Unauthenticated | Downloaded 05/30/22 01:52 PM UTC
ensembles. After postprocessing, the skill of the streamflow forecasts for high-flow conditions are mostly improved across the entire 7-day forecast cycle. The improvements in low- to moderate-flow forecasts are mainly seen across the short-range lead times (<4 days).

Acknowledgments. We gratefully acknowledge the suggestions and comments provided by the two anonymous reviewers. We also acknowledge the funding support provided by the NOAA/NWS through Award NA14NWS4680012. We are thankful to Peter Ahnert and Seann Reed of NOAA’s MARFC for providing the observed forcing data and many valuable suggestions regarding this research.

APPENDIX

Verification Metrics

a. Correlation coefficient

The correlation coefficient $R$ represents the linear association between two variables (observed and simulated flows in this study). The correlation coefficient is defined as

$$ R = \frac{\sum_{i=1}^{N} S_i Y_i - \sum_{i=1}^{N} S_i \sum_{i=1}^{N} Y_i}{\sqrt{\left[ \sum_{i=1}^{N} S_i^2 - \left( \sum_{i=1}^{N} S_i \right)^2 \right] \left[ \sum_{i=1}^{N} Y_i^2 - \left( \sum_{i=1}^{N} Y_i \right)^2 \right]^{1/2}}} $$

where $S_i$ and $Y_i$ denote the simulated and observed flow, respectively, at time $i$, and $N$ denotes the total number of pairs of observed and simulated flows.

b. Modified correlation coefficient

The correlation coefficient only accounts for the shape but not the size of the hydrograph. In addition, it can be strongly affected by outliers. To overcome these limitations, McCuen and Snyder (1975) developed a modified version of the correlation coefficient to compare event-specific observed and simulated hydrographs. In the modified version, an adjustment factor based on the ratio of the observed and simulated flows is introduced to refine the conventional correlation coefficient. The modified correlation coefficient $R_m$ is defined as

$$ R_m = R \frac{\min(\sigma_{\text{sim}}, \sigma_{\text{obs}})}{\max(\sigma_{\text{sim}}, \sigma_{\text{obs}})} $$

where $\sigma_{\text{sim}}$ and $\sigma_{\text{obs}}$ denote the standard deviation of the simulated and observed flows, respectively.

c. Percent bias

Percent bias (PB) measures the average tendency of the simulated values to be larger or smaller than the observed. The PB is given by

$$ PB = \frac{\sum_{i=1}^{N} (Y_i - S_i)}{\sum_{i=1}^{N} Y_i} \times 100, \quad (A3) $$

where $S_i$ and $Y_i$ denote the simulated and observed flow, respectively, at time $i$.

d. Nash–Sutcliffe efficiency

The Nash–Sutcliffe efficiency (NSE) is defined as the ratio of the residual variance to the initial variance. It is widely used to measure the accuracy of the simulated flows in comparison to the observed mean. The range of the NSE can vary between $-\infty$ and 1. Any positive value close to 1 indicates a good match between the simulated and observed variable, while a negative value indicates that the observed mean is better than the simulated. The NSE is defined as

$$ NSE = 1 - \frac{\sum_{i=1}^{N} (S_i - Y_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y}_i)^2}, \quad (A4) $$

where $S_i$, $Y_i$, and $\bar{Y}_i$ are the simulated, observed, and mean observed flow, respectively, at time $i$.

e. Relative mean error

Relative mean error (RME) quantifies the average error between the ensemble mean forecasts and their corresponding observations as a fraction of the averaged observed value. RME gives an indication how good the forecast is relative to the observation. RME is expressed as follows:

$$ RME = \frac{\sum_{i=1}^{n} (\bar{X}_{i,k} - Y_i)}{\sum_{i=1}^{n} Y_i}, \quad (A5) $$

where $\bar{X}_{i,k} = 1/m \sum_{k=1}^{m} X_{i,k}$, $m$ is the number of ensemble members, $X_{i,k}$ is the forecast for member $k$ and time $i$, $Y_i$ denotes the corresponding observation at time $i$, and $n$ denotes the total number of pairs of forecasts and observed values.

f. Brier skill score

The Brier score (BS) is analogous to the mean-square error, but where the forecast is a probability and the
observation is either a 0 or 1 (Brown and Seo 2010). The BS is given by

\[
BS = \frac{1}{n} \sum_{i=1}^{n} \left[ F_{X_i}(q) - F_{Y_i}(q) \right]^2, \tag{A6}
\]

where the probability of \( X_i \) to exceed a fixed threshold \( q \) is

\[
F_{X_i}(q) = \Pr(X_i > q), \tag{A7}
\]

\( n \) is again the total number of forecast–observation pairs, and

\[
F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise} \end{cases}. \tag{A8}
\]

To compare the skill score of the main forecast system with respect to the reference forecast, it is convenient to define the Brier skill score (BSS) as

\[
BSS = 1 - \frac{BS_{\text{main}}}{BS_{\text{reference}}}, \tag{A9}
\]

where \( BS_{\text{main}} \) and \( BS_{\text{reference}} \) are the BS values for the main forecast system (i.e., the system to be evaluated) and reference forecast system, respectively. Any positive values of the BSS, from 0 to 1, indicate that the main forecast system performs better than the reference forecast system. Thus, a BSS of 0 indicates no skill and a BSS of 1 indicates perfect skill.

g. Mean continuous ranked probability skill score

Continuous ranked probability score (CRPS), which is less sensitive to sampling uncertainty, is used to measure the integrated square difference between the cumulative distribution function (cdf) of a forecast \( F_{x}(q) \) and the corresponding cdf of the observation \( F_{y}(q) \). The CRPS is given by

\[
\text{CRPS} = \int_{-\infty}^{\infty} \left[ F_{x}(q) - F_{y}(q) \right]^2 dq. \tag{A10}
\]

To evaluate the skill of the main forecast system relative to the reference forecast system, the associated skill score, the mean continuous ranked probability skill score (CRPSS), is defined as

\[
\text{CRPSS} = 1 - \frac{\text{CRPS}_{\text{main}}}{\text{CRPS}_{\text{reference}}}, \tag{A11}
\]

where the CRPS is averaged across \( n \) pairs of forecasts and observations to calculate the mean CRPS of the main forecast system (CRPS_{main}) and reference forecast system (CRPS_{reference}). The CRPSS ranges from \(-\infty \) to 1, with negative scores indicating that the system to be evaluated has worse CRPS than the reference forecast system, while positive scores indicate a higher skill for the main forecast system relative to the reference forecast system, with 1 indicating perfect skill.

REFERENCES


