Final Project Progress Report
Reporting Period: From May 1, 2014, To April 30, 2018

National Oceanic and Atmospheric Administration (NOAA)
Collaborative Science, Technology, and Applied Research (CSTAR) Program

Understanding and Improving the Full Hydrometeorological Forecasting Chain Using Multimodel Ensembles

Principal Investigator (PI):
Alfonso I. Mejia[+]
Assistant Professor of Civil and Environmental Engineering
The Pennsylvania State University
215B Sackett Building
University Park, PA 16802
Phone: (814) 865-0639
Fax: (814) 863-7304
E-mail: amejia@engr.psu.edu

Co-Principal Investigator (Co-PI):
Christopher J. Duffy
Professor of Civil and Environmental Engineering
The Pennsylvania State University
212 Sackett Building
University Park, PA 16802
Phone: (814) 863-4384
Fax: (814) 863-7304
E-mail: cxd11@psu.edu

Chapter 1: Executive Summary

This CSTAR project assembled and tested a multimodel regional hydrological ensemble prediction system (RHEPS) to make recommendations about the implementation of such system, and its various system components (e.g., weather ensembles, statistical processors and hydrological models), in operational settings. A central goal of the project was to design and implement the RHEPS to emulate realistic and relevant operational forecasting conditions. This allowed us to evaluate potential improvements to the RHEPS that are applicable to operational forecasting. The project was implemented in the Middle Atlantic River Forecast Center (MARFC) geographical domain in collaboration with the MARFC. However, one critical task of the project, namely the verification of different ensemble precipitation forecast products, was implemented over the geographical domain of the eastern US, in collaboration with the eastern RFCs and the Weather Prediction Center.

The project entailed the execution of the following three main tasks:

i) Comprehensive verification of ensemble precipitation forecasts for the eastern US
ii) Development, implementation and evaluation of both statistical weather and hydrological processors for application in operational forecasting
iii) Development, implementation and evaluation of a multimodel regional hydrological ensemble prediction system (RHEPS) consisting of: weather ensembles, relevant hydrometeorological data, statistical processors, and multiple hydrological models. The hydrological models employed were WRF-Hydro, HL-RDHM and Continous-API.

The specific results and outcomes from these tasks are detailed in the main body of this report. Chapters 2 and 3 include the verification of GEFS, SREF, and WPC-PQPF precipitation ensembles over the eastern US. Chapters 4-6 use the RHEPS to propose and evaluate different statistical processors for both weather and hydrological ensembles. Chapter 7 concludes with the implementation and evaluation of the multimodel RHEPS. Each chapter includes recommendations relevant to operational hydrometeorological forecasting.

The main outcome of the project was to demonstrate that a multimodel RHEPS is able to improve the skill of ensemble streamflow forecasts more than increases in the ensemble size of a single model. This is relevant to operational forecasting because generating many ensembles in real time is often not feasible or realistic, and may not be as effective if skill enhancements are dominated by model diversity (multimodel information). The project also found that streamflow forecasts tend to be skillful up to lead times of 7 days, with streamflow postprocessing enhancing forecast skill up to 2 days. Overall, the project results and outcomes provide numerous recommendations, as presented in each chapter, for enhancing ensemble streamflow forecasting in operational settings.

Theses, Papers, and Presentations that Resulted from this CSTAR Project

Dissertations and Theses

Siddique, R. (2017), Improving medium-range hydrological forecasting in the U.S. middle Atlantic region, PhD dissertation, The Pennsylvania State University, University Park, PA.

Xingchen, Y. (2016), A comparison between Bayesian model averaging and heteroscedastic censored logistic regression using 2012 GEFS precipitation reforecasts over the U.S. middle-Atlantic region, MS Thesis, The Pennsylvania State University, University Park, PA.

Peer-reviewed Journal Articles

Sharma*, S., R. Siddique*, P. Mendoza, S. Reed, P. Ahnert, and A. Mejia (2017). Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system. Hydrology and Earth System Sciences.

Siddique*, R., and A. Mejia (2017). Ensemble streamflow forecasting across the U.S. middle Atlantic region. Journal of Hydrometeorology, 18(7), 1905-1928, doi:10.1175/JHM-D-16-0243.1

Yang*, X., S. Sharma*, R. Siddique*, S. Greybush, A. Mejia (2017), Postprocessing of GEFS precipitation ensemble reforecasts over the U.S. middle Atlantic region, Monthly Weather Review, 145, 1641-1658, doi: 10.1175/MWR-D-16-0251.1.

Sharma*, S., R. Siddique*, N. Balderas*, J. D. Fuentes, S. Reed, P. Ahnert, R. Shedd, B. Astifan, R. Cabrera, A. Laing, M. Klein, and A. Mejia (2017), Eastern U.S. verification of ensemble precipitation forecasts, Weather and Forecasting, 32, 117–139, doi: 10.1175/WAF-D-16-0094.1.

Siddique et al. (2015), Verification of Precipitation Forecasts from Two Numerical Weather Prediction Models in the Middle-Atlantic Region of the USA: A Precursory Analysis to Hydrologic Forecasting, Journal of Hydrology, 539(3), 1390-1406.

Conferences and Seminar Presentations

Gomez, M., Sharma, S., Siddique, R., and Mejia, A. (2018), Skill of Ensemble Flood Forecast Maps, ASCE EWRI, Minneapolis, Minnesota, June 5, 2018.

Sharma, S., Siddique, R., and Mejia, A. (2018), Skill of multimodel ensemble streamflow forecasts, ASCE EWRI, Minneapolis, Minnesota, June 5, 2018.

Gomez, M., Sharma, S., Siddique, R., and Mejia, A. (2017), Ensemble Flood Forecast Maps: A Case Study in the Delaware River near Philadelphia, Mid-Atlantic Water Resources Conference, Shepherdstown, WV, October 12, 2017.

Sharma, S., Siddique, R., Gomez, M., and Mejia, A. (2017), Ensemble Flood Forecasting in the U.S. Middle Atlantic Region, Mid-Atlantic Water Resources Conference, Shepherdstown, WV, October 12, 2017.

Mejia (2017), Verifying a regional hydrological ensemble prediction system, University of Iowa, invited seminar in the IIHR-Hydroscience & Engineering, April 28, 2017.

Mejia (2016), Hydrologic ensemble forecasting across the U.S. middle Atlantic region: Demonstration of the forecasting system and of its statistical weather postprocessor, Meteo Colloquium, The Pennsylvania State University, September 21, 2016.

Siddique, R., Yang, X., Mejia, A., Sharma, S., Gomez, M., (2016), Hydrologic ensemble forecasting across the U.S. middle Atlantic region using the Pennsylvania State University Regional Hydrologic Ensemble Prediction System (Penn State-RHEPS), Mid-Atlantic Regional Water Conference, Wilmington, Delaware, September 16, 2016.

Sharma, S., R. Siddique, and A. Mejia (2016), Verification of ensemble precipitation forecasts across the eastern USA, ASCE EWRI, West Palm Beach, Florida, May 23, 2016.

Siddique, R. and A. Mejia (2016), Verification and uncertainty estimation of streamflow forecasts across different spatial scales in the middle Atlantic region of the USA, ASCE EWRI, West Palm Beach, Florida, May 23, 2016.

Sharma, S. (2016), Verification of precipitation forecast data over the eastern USA, 2016 College of Engineering Research Symposium, University Park, PA, April 5, 2016.

Balderas, N., A. Mejia, and R. Siddique (2016), Verification of GEFS Precipitation across the Eastern U.S, 15th Annual Student Conference, AMS Annual Meeting, New Orleans, LA, January 10, 2016.

Siddique, R. and A. Mejia (2015), Improving flood forecasting in the Susquehanna River basin using a probabilistic approach, 10th Annual Susquehanna River Symposium, November 13-14, 2015 at Bucknell University, poster presentation.

Siddique et al. (2015), Application of Bayesian Model Averaging for Calibration of GEFS Precipitation Reforecasts in the Middle Atlantic Region of the US, ASCE EWRI, Austin, Texas, May 19, 2015.

Mejia, A. (2015), Hydrologic modeling: Forecasting as a case study, ADAPT Center, The Pennsylvania State University, University Park, PA, May 11, 2015.

Siddique et al. (2014), Verification of Precipitation Forecasts from Two Numerical Weather Prediction Models in the Middle-Atlantic and North-Eastern Region of the USA, AGU, San Francisco, CA, December 19, 2014.

Chapter 2: Verification of the GEFS and SREF precipitation ensembles over the middle-Atlantic region

ABSTRACT

Accurate precipitation forecasts are required for accurate flood forecasting. The structures of different precipitation forecasting systems are constantly evolving, with improvements in forecasting techniques, increases in spatial and temporal resolution, improvements in model physics and numerical techniques, and better understanding of, and accounting for, predictive uncertainty. Hence, routine verification is necessary to understand the quality of forecasts as inputs to hydrologic modeling. In this study, we verify precipitation forecasts from the National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2), as well as the 21-member Short Range Ensemble Forecast (SREF) system. Specifically, basin averaged precipitation forecasts are verified for different basin sizes (spatial scales) in the operating domain of the Middle Atlantic River Forecast Center (MARFC), using multi-sensor precipitation estimates (MPEs) as the observed data. The quality of the ensemble forecasts is evaluated conditionally upon precipitation amounts, forecast lead times, accumulation periods, and seasonality using different verification metrics. Overall, both GEFSRv2 and SREF tend to overforecast light to moderate precipitation and underforecast heavy precipitation. In addition, precipitation forecasts from both systems become increasingly reliable with increasing basin size and decreasing precipitation threshold, and the 24-hourly forecasts show slightly better skill than the 6-hourly forecasts. Both systems show a strong seasonal trend, characterized by better skill during the cool season than the warm season. Ultimately, the verification results lead to guidance on the expected quality of the precipitation forecasts, together with an assessment of their relative quality and unique information content, which is useful and necessary for their application in hydrologic forecasting.

1. Introduction

Floods are one of the major natural threats to human life and property (Hegger et al., 2014; Milly et al., 2002; Pall et al., 2011; Pielke and Downton, 2000). In the United States (US) alone, floods have caused annually since 1990 more than US$5 billion worth in damages and a significant death toll (Downton et al., 2005; Morss et al., 2005; Schildgen, 1999). Globally, flood losses (adjusted for inflation) have increased from an average of US$7 billion/year in the 1980s to some US$24 billion/year in the period 2001-2011 (Kundzewicz et al., 2014). With burgeoning human population and urbanization, flood risks and vulnerabilities are increasing with time such that floods, if not properly mitigated, could become more destructive and costlier disasters in the future (Bouwer, 2010; Changnon, 1998; McCarthy, 2001; Morss et al., 2005). Since floods mostly occur for natural reasons and flood control and management with limited resources is a complex challenge, reliable and skillful flood forecasts are needed. Indeed, flood forecasts can contribute to reducing flood damages and protecting life and property (Kelsch et al., 2001; Montz and Gruntfest, 2002; Pagano et al., 2014). They can also provide necessary information for developing decision making tools and implementing improved flood control measures (Demargne et al., 2010).

y of streamflow forecasts relies heavily on a proper understanding of hydrological and meteorological uncertainties (Brown et al., 2012; Demeritt et al., 2007; Pappenberger et al., 2008). Uncertainties in the precipitation forcing can contribute significantly to uncertainties in

the streamflow forecasts (Cloke and Pappenberger, 2009; He et al., 2009; Kobold and Suselj, 2005; Schaake et al., 2007). Indeed, accurate precipitation forecasts, as well as temperature, are essential for producing skillful streamflow forecasts (Brown et al., 2012; Buizza, 2008; Cloke and Pappenberger, 2009; Demargne et al., 2010; Voisin et al., 2010). Thus, ensemble forecasts are increasingly being employed to understand and quantify forecast uncertainties (Buizza et al., 2005; Davolio et al., 2008; Demeritt et al., 2007; Epstein, 1969). Specifically, precipitation and temperature ensemble forecasts from different numerical weather prediction (NWP) models are being tested and evaluated as potential inputs to hydrologic models for improved streamflow forecasting (Adams and Ostrowski, 2010; Brown et al., 2012; Cloke and Pappenberger, 2009; He et al., 2009; Schumacher and Davis, 2010; Thielen et al., 2009). However, the structures of different meteorological forecasting systems are constantly evolving, with improvements in forecasting techniques, increases in spatial and temporal resolution, improvements in model physics and numerical techniques, and better understanding and modeling of uncertainty (Novak et al., 2013). Thus, in order to monitor and improve forecast quality, routine verification is required (Demargne et al., 2010; Jolliffe and Stephenson, 2012; Wilks, 2011).

Our primary goal with this study is to verify the forecast quality of precipitation ensembles from two NWP models in the Middle Atlantic Region (MAR) of the US, as a precursor to hydrologic forecasting. Our motivation is to help inform the development and implementation of hydrologic models and precipitation pre-processing tools in the MAR. In this context, we pose the following questions: To what extent, and how, does the quality of the precipitation ensemble forecasts vary for different forecasting systems and quality attributes? Is the pre-processing of precipitation ensembles likely to result in improved forecasts in the MAR? How does the basin size influence the quality of the precipitation forecasts? For the NWP models, we use the precipitation forecasts from the National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2) and the NCEP 21-member Short Range Ensemble Forecast (SREF) system, as these two ensemble systems are currently being proposed and tested for operational hydrologic ensemble forecasting in the US (Brown et al., 2012). For the verification, we employ the Ensemble Verification System (EVS) developed by Brown et al. (2010). We investigate the forecast quality of the GEFSRv2 and SREF conditionally upon precipitation amount, forecast lead times, different accumulation periods, and seasonality, as well for different verification metrics. We verify forecasts in the MAR because of the high frequency of floods and the ability and potential for floods to cause devastating damages within this region of the US (Choi and Fisher, 2003; Najjar et al., 2000; Neff et al., 2000).

Various verification studies have been performed for the GEFSRv2, SREF, and other similar forecasting systems (Baxter et al., 2014; Brown et al., 2012; Brown et al., 2014; Buizza et al., 2005; Charles and Colle, 2009; Demargne et al., 2010; Hamill et al., 2013; Hamill and Colucci, 1998; Hamill and Whitaker, 2006; Hamill et al., 2006; Jones et al., 2007; Pappenberger and Buizza, 2009; Schumacher and Davis, 2010; Whitaker et al., 2008; Yuan et al., 2005; Zhu, 2005). For instance, Brown et al. (2012) verified precipitation forecasts from the SREF system for basins across the US, encompassing 4 different National Weather Service (NWS) River Forecast Centers (RFCs), while Baxter et al. (2014) verified precipitation forecasts from the GEFSRv2 over the southeastern US. Here we verify the GEFSRv2 and SREF using a common verification strategy and set of metrics. We compare the GEFSRv2 and the SREF for the MAR, using multi-sensor precipitation estimates (MPEs) to verify the forecasts at multiple spatial scales, which are relevant for hydrologic forecasting.

The paper is organized as follows. In section 2, we describe the study area and datasets used. Section 3 explains the methodology employed for the verification strategy. In section 4, we present and discuss the verification results. Lastly, in section 5, we outline the main conclusions.

2. Study area and datasets

2.1 Study area

Forecast skill varies with geographic location (Baxter et al., 2014; Brown et al., 2012; Fan et al., 2014; Olsson and Lindström, 2008; Verkade et al., 2013), which implies that forecast verification should be conducted at various spatial scales and for different climate regions. Previously, precipitation forecasts from the GEFSRv2 and SREF were verified between regions in the US (Brown, 2014; Brown et al., 2012; Brown et al., 2014). To complement and expand these studies, we perform the verification within the MAR. Figure 1 shows the boundaries of the MAR which encompass the operating domain of the National Oceanic and Atmospheric Administration's Middle Atlantic River Forecast Center (MARFC).

The spatial extent of the MAR includes all of the states of Delaware and the District of Columbia along with parts of Maryland, New York, New Jersey, Pennsylvania, Virginia, and West Virginia (Polsky et al., 2000; Sinnott and Cushing, 1978). Figure 1 illustrates the geographic extent of the MAR. Although the MAR accounts only for approximately 5 percent of the total land mass of the US, 41 million people or 10% of the nation's population reside here (United States Census Bureau, 2013). In fact, some of the largest metropolitan areas in the US are located in the MAR (see Figure 1). Further, the MAR encompasses the drainage basin of the following four major rivers in the US: the Delaware, Susquehanna, Potomac, and James River (Sinnott and Cushing, 1978). It is composed of several physiographic provinces, including the Coastal Plain, Appalachian Plateaus, Ridge and Valley, and Piedmont province (Fenneman and Johnson, 1946). Each of these provinces represents a unique combination of regional terrain texture, rock type, and geologic structure. These unique physical features result in characteristic hydrologic responses for each province, making the hydrology of the MAR diverse (Ator et al., 2005; Rutledge and Mesko, 1996; Wardrop et al., 2005). In terms of climatic conditions, the MAR is relatively humid, having an average annual temperature of 11^{o}C and mean annual precipitation of approximately 900-1000 mm (Polsky et al., 2000). Over the last century, the MAR has seen a notable increase in the frequency of extreme precipitation events and variability of climatic patterns (Karl et al., 1996). This, coupled with high levels of urbanization (Figure 1 illustrates the extent of urbanization in the MAR), makes the MAR extremely flood prone (Najjar et al., 2000).

2.2 GEFSRv2

GEFSRv2 are the retrospective forecasts produced using the same atmospheric model and initial conditions used by NCEP's operational Global Ensemble Forecast System (GEFS) (Hamill et al., 2013). For both GEFS and GEFSRv2, the perturbations of the initial conditions are made by the Ensemble Forecast System, but the GEFS model runs are initiated four times daily whereas the GEFSRv2 is initiated only once per day, at 00 UTC (Hamill et al., 2013; Wei et al., 2008). The forecast lead times of GEFSRv2 extend from 1 to 16 days and each forecast cycle consists of forecasts valid for 3 hourly accumulations from days 1 to 3 and 6 hourly accumulations from days 4 to16. Two different model runs with separate horizontal resolution

are available. The first model runs (T254L42) comprise reforecasts for the first 8 days of the forecast cycle while the second model runs (T190L42) comprise reforecasts for the last 8 days of the cycle. The native grid resolution for the first and second model runs are 0.5^0 (~55 km) and 0.67^0 (~73 km) Gaussian grid spacing, respectively. Figure 1a illustrates the 0.67^0 GEFSRv2 grid. Approximately 25 years of GEFSRv2 data have been archived. Here we used a total of 12 years, from January 2002 to December 2013. This period was selected because the MPE data from 2002 to 2013 is considered to be of a relatively consistent quality as compared to the earlier years. Table 1 summarizes the main characteristics of the GEFSRv2. Additional information about the GEFSRv2 is provided by Hamill et al. (2013).

2.3 SREF

For the verification analysis, 4 years of operational SREF forecasts were available, ranging from January 2008 to February 2010 and from January 2012 to November 2013. A continuous dataset or reforecast was not available. The SREF forecasts are initiated 4 times a day at 0300, 0900, 1500 and 2100 UTC for the period of record used here. However, the archived data for the period between 2008 and 2009 only included model runs for 0300 and 2100. We are aware that using two model runs instead of four might add systematic variability but we chose to include the period 2008-2009 to reduce sampling uncertainty, as much as possible. Thus, we only used these two runs when performing the verification for the period 2008-2009, while the four daily runs were used for the period 2012-2013. Each forecast cycle comprises lead times from 3 to 87 hours and the forecast for each lead time is valid for 3 hourly accumulations of precipitation. For example, a model run at 0300 UTC will produce 3 hourly forecasts that are valid for the next 3, 6, 9, 12,…, and 87 hours and each of them is the accumulation of the preceding 3 hours of precipitation. The horizontal resolution of the SREF has been increased several times since it first became operational in 2001 (Du et al., 2009). The operational SREF started with a horizontal resolution of 40 km and it changed to 32 km in 2009. The last major change was made in 2012, which increased the horizontal resolution of all members to 16 km. Figure 1b illustrates the 32 km SREF grid. Table 1 summarizes some of the main characteristics of the SREF. Note that Table 1 does not include information about the evolution of model cores and physics in the SREF system but this and other relevant information about the SREF system can be found elsewhere (see, e.g., Du et al. (2009), Du et al. (2003), Du et al. (2006)).

2.4 MPEs

For this study, the GEFS and SREF forecasts were verified against MPEs. MPEs are currently the highest quality and most accurate estimates of spatiotemporal precipitation that are available, having the least amount of bias and showing maximum correlation with reference to independent gauge observations in the MAR (Breidenbach and Bradberry, 2001). The MPEs are high resolution, hourly gridded precipitation data products, which are produced by combining multiple radar estimates and rain gauge measurements (Seo and Breidenbach, 2002; Seo et al., 2010; Young et al., 2000). The MPEs are available as Hydrologic Rainfall Analysis Project grids (Greene et al., 1979) in the polar stereographic map projection at a resolution of approximately 4 x 4 km^2 (Breidenbach and Bradberry, 2001). The development of the MPE products requires the implementation of several steps. Initially, gauge-only hourly schemes are produced using the Thiessen polygon method (Thiessen, 1911), together with an optimal estimation (OE) technique (Seo, 1998a; Seo, 1998b). Then, a radar-only scheme is produced by mosaicking two-dimensional Digital Precipitation Array data from multiple radars (Seo et al., 2010; Young et al.,

2000). Finally, hourly radar and gauge schemes are merged into the final MPE products by optimally combining them using an OE technique. For each hour, forecasters at MARFC examine the results of automated MPE procedures, quality control gauge data, correct erroneous radar data, and return the multi-sensor grid generation procedures as needed. Recently, MARFC has started using the next generation multisensor quantitative precipitation estimates, also known as Q2, radar-only fields (http://www.nssl.noaa.gov/projects/q2/) as an optional input to MPE processes. This final MPE product is used operationally by RFCs and local NWS offices for different hydrologic, meteorological, and water resources applications. For the verification analysis, we accumulated the MPEs into 6 and 24 hourly accumulations to match the temporal scales commonly used in operational hydrologic forecasting. Ultimately, we utilized the MPEs to compute the observed mean areal precipitation and verification metrics for selected basin sizes.

3. Methodology

Meteorological ensemble forecasts are important to scientists, administrators, and decision makers (Brier and Allen, 1951; Cloke and Pappenberger, 2009; Jolliffe and Stephenson, 2012; Ramos et al., 2010), among other groups of users. Depending on the application, different measures of forecast quality may be preferred or emphasized. Murphy (1993) identifies three attributes for assessing the 'goodness' of forecasts: consistency, quality, and socio-economic value. Our focus here is on the quality attribute, i.e. the degree of correspondence between the forecasts and observations. Wilks (2011) recommends and outlines a series of 'scalar' attributes, e.g., accuracy, reliability, resolution, discrimination, and sharpness, to determine the forecast quality and compare different forecast systems. Ultimately, to identify the key sources of errors in the forecasts and to provide supporting information to different users (Demargne et al., 2010; Pappenberger et al., 2008; Wilks, 2011), a pool of verification metrics needs to be employed. For this study, we selected verification metrics on the basis of previous results (Brown, 2014; Brown et al., 2012; Jolliffe and Stephenson, 2012; Wilks, 2011) as well as the needs of hydrological forecasters. We selected a mixed pool of deterministic and probabilistic metrics to facilitate the verification of the ensemble forecasts. Deterministic verification metrics evaluate the ensemble mean forecast, while probabilistic metrics help to evaluate the forecast probabilities, including the ensemble spread (Brown et al., 2010). We describe next the metrics used in this study.

We use the relative mean error (RME) and correlation coefficient as the deterministic metrics. The correlation coefficient measures the degree of linear association between the observed and forecast variable (Murphy and Epstein, 1989). It is also a good summary measure of their joint probability distribution (Murphy, 1993; Murphy and Winkler, 1987). However, the correlation coefficient is unable to provide any direct information about the bias in the forecasts (Brier and Allen, 1951). Hence, we use the RME to quantify the average error between the ensemble mean forecast and their corresponding observations as a fraction of the average observed value. This error metric serves to explore the bias of a forecast system (Wilks, 2011). The RME is given by

$$RME = \frac{\sum_{i=1}^{n} (\overline{X_i} - Y_i)}{\sum_{i=1}^{n} Y_i} \quad (1)$$

where $\bar{X}_i = 1/m \sum_{k=1}^{m} X_{i,k}$, 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.

Besides deterministic metrics, we also employ probabilistic metrics, which are described next, to investigate the probabilistic attributes of the selected forecasting systems. To measure the skill of the forecasts, we use the Brier Skill Score (BSS) which is derived from the Brier Score (BS). The BS is a common verification metric and analogous to the mean square error (MSE) (Wilks, 2011). However, unlike the MSE, the BS is defined for a discrete probability forecast. The BS can be useful for verifying heavy precipitation because it employs the probability of occurrence of an event rather than focusing on the magnitude of the error between the forecasts and observations (Brown et al., 2010; Wilks, 2011). The BS can be expressed as follows:

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

where $F_{X_i}(q) = \Pr[X_i > q]$ or the probability of X_i to exceed q, q is a fixed threshold, n is the number of verification pairs, and

$$F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Together with equations (2)-(3), the Brier Skill Score (BSS) can be computed as:

$$BSS = 1 - \frac{BS_{\text{main}}}{BS_{\text{reference}}} \tag{4}$$

where BS_{main} and $BS_{\text{reference}}$ are the BS values for the main and reference forecasting system, respectively. Thus, any positive BSS value between [0, 1] indicates that the main forecasting system (i.e., the system to be evaluated) performs better than the reference forecasting system. In this study, climatology is used as the reference forecasting system for BSS. When factoring by the forecast probability of occurrence, the BS can be further decomposed into contributions from (lack of) reliability, resolution, and uncertainty. However, instead of using the decomposed BS to quantify the reliability and resolution of the forecasts, we use the so-called reliability diagram.

The reliability diagram plots the average observed probability of occurrence of an event given the forecast probability, against its forecast probability of occurrence . The reliability diagram serves to explore the full joint distribution of forecasts and observations for a discrete event, and conveys information about the quality of forecasts in different ranges of the forecast probability distribution (Wilks, 2011). A forecasting system is perfectly reliable when the forecast indicates that the probability of occurrence of an event is p and it is actually observed with a relative frequency of p of those occasions on which such forecasts are issued. If the forecast probabilities are divided into k bins and a forecast event is defined by the exceedance of a threshold, q, then the average probability of the forecasts that fall in the kth bin, B_k, is given by

$$\bar{F}_{X_k}(q) = \frac{1}{|I_k|} \sum_{I_k} F_{X_i}(q), \text{ where } I_k = \{i : i \in B_k\}. \tag{5}$$

The corresponding average probability of the observations is given by

$$\bar{F}_{Y_k}(q) = \frac{1}{|I_k|}\sum_{I_k} F_{Y_i}(q), \text{ where } F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The reliability diagram plots $\bar{F}_{X_k}(q)$ against $\bar{F}_{Y_k}(q)$ for a total number of forecasts $|I_k|$ in each bin B_k.

The Continuous Ranked Probability Score (CRPS) measures the integral square error of an ensemble forecasting system (Wilks, 2011). Specifically, the CRPS is computed by integrating the squared difference between the cumulative distribution function (cdf) of the forecasts and observed values,

$$CRPS = \int_{-\infty}^{\infty} [F_X(q) - F_Y(q)]^2 \, dq, \tag{7}$$

where $F_X(q)$ and $F_Y(q)$ are the cdfs of the forecast and observed variables, respectively. To verify a set of forecasts, the CRPS is averaged over the total number of pairs of forecast and observed values for the main forecasting system ($\overline{CRPS}_{\text{main}}$). The mean Continuous Ranked Probability Skill Score (CRPSS) measures the fractional improvement in CRPS of the main forecast system against a reference forecasting system ($\overline{CRPS}_{\text{reference}}$),

$$CRPSS = 1 - \frac{\overline{CRPS}_{\text{main}}}{\overline{CRPS}_{\text{reference}}}, \tag{8}$$

where positive scores indicate that the skill of the main forecasting system is higher than the reference forecasting system. In this study, we use sample climatology as the reference forecasting system in equation (8).

Event discrimination is another important attribute of forecast quality. Discrimination is concerned with the ability of the forecasts to distinguish between occurrences and non-occurrences of an event by forecasting a different set of probabilities for different observed outcomes (Jolliffe and Stephenson, 2012). Here, we use the relative operating characteristic (ROC) (Green and Swets, 1966) to measure event discrimination. The ROC measures the trade-off between the fraction of forecasts that correctly predict the occurrence of an event (probability of detection) and the fraction that incorrectly predict its occurrence (probability of false detection, i.e., the risk of "crying wolf"). For a given threshold (event), the PoD is given by

$$PoD = \frac{\sum_{i=1}^{n} I_{X_i}(F_{X_i}(q) > d \,|\, Y_i > q)}{\sum_{i=1}^{n} I_{Y_i}(Y_i > q)}, \tag{9}$$

where d denotes the probability threshold at which the event triggers some action (i.e., the forecast is deemed an occurrence) and I denotes the indicator function. Using the same notation as in equation (9), the PoFD can be expressed as

$$PoFD = \frac{\sum_{i=1}^{n} I_{X_i}(F_{X_i}(q) > d \,|\, Y_i \le q)}{\sum_{i=1}^{n} I_{Y_i}(Y_i \le q)}. \tag{10}$$

The relationship between the PoD and PoFD is assumed to be bivariate normal such that

$$PoD = \phi\{a + b\phi^{-1}(PoFD)\}\ , \text{where } a = \frac{\mu_{PoD} - \mu_{PoFD}}{\sigma_{PoDa}} \text{ and } b = \frac{\sigma_{PoFD}}{\sigma_{PoD}}. \quad (11)$$

ϕ is the cdf of the standard normal distribution, μ_{PoD} and μ_{PoFD} are the means while σ_{PoD} and σ_{PoFD} denote the standard deviations of the PoD and PoFD, respectively. The ROC curve plots the PoD (fraction of true alarms) against the PoFD (fraction of false alarms) for all possible values of the decision threshold, d [0, 1], noting that an ensemble forecast is essentially a step function, with as many possible values of d as the number of ensemble members. This metric is useful for evaluating the information content in the forecasts for large observed events or flood warnings, where important decisions need to be made.

Throughout the verification analysis, we used three different precipitation thresholds to represent light, moderate, and heavy precipitation. The thresholds were fixed throughout the analysis and were determined using the entire sample of observed precipitation (i.e., from 2002 to 2013). The thresholds are 1 mm (light), 5 mm (moderate), and 15 mm (heavy) for the 6 hourly accumulations, and 1, 10, and 30 mm for the 24 hourly accumulations. These thresholds correspond to sample climatological non-exceedance probabilities approximately equal to 0.5, 0.9, and 0.99 for light, moderate, and heavy precipitation, respectively.

We also used for the verification analysis three different areal extents, together with their corresponding observed mean areal precipitation. These areal extents are used to represent small (100 x 100 km^2), intermediate (300 x 300 km^2), and large (500 x 500 km^2) basin sizes. To determine the precipitation forecasts for these areal extents, we used the areal average of the forecast cells within the square basin size. Further, to reduce the sampling uncertainty when computing the verification metrics, we selected three or more basins of the same size from different locations within or in close proximity to the MAR boundaries via spatial pooling. Thus, for a given basin size, each basin was treated as a separate verification unit that provides a new set of forecast and observation pairs. Then, we aggregated the verification results obtained from the different verification units for the same basin size and an average performance was obtained for the selected metric. To determine the 90% confidence intervals, we applied the block bootstrapping technique using a minimum of 2000 samples (Politis and Romano, 1994).

4. Results and discussion

In this section, we use the metrics described previously to assess the quality of the precipitation ensemble forecasts from the GEFSRv2 and SREF. This section is divided into three subsections. The first and second subsections describe and discuss the verification results for the GEFSRv2 and SREF, respectively. The verification analysis for the GEFSRv2 is conducted for the period between 2002 and 2013. The verification analysis for the SREF is for the period between January 2008 and January 2010 and between January 2012 and November 2013. The third subsection presents and discusses verification results for both the GEFSRv2 and SREF using a common period of analysis (i.e., that of the SREF).

4.1 Verification of the GEFSRv2 precipitation forecasts

Figures 2a-d show box plots of the precipitation forecast errors (forecast minus observed) for the GEFSRv2, arranged by increasing amount of observed precipitation, for lead times of 1, 3, 5, and 7 days, respectively. The box plots (Figures 2a-d) are plotted for 24 hourly accumulations and a 100 x 100 km^2 basin size. Figures 2a-d indicate that the GEFSRv2 overestimates light to moderate precipitation events while it consistently underestimates heavy

precipitation events. This trend is similar across lead times. The underestimation of heavy precipitation is quite noticeable in Figure 2c where the forecasts fail to capture the largest events by a considerable margin. The results for other basin sizes (not shown) are qualitatively similar to the results shown in Figure 2. Overall, Figure 2 indicates the presence of conditional biases (i.e., the forecast error depends on the value of the observed precipitation) in the GEFSRv2 precipitation forecasts. Further investigation with the NWP models is required to better understand and assess the reasons behind this conditional bias or the tendency to underestimate large precipitation events.

In Figures 3a-c and 3d-f, we illustrate the RME for the 6 and 24 hourly GEFSRv2 ensemble mean forecasts, respectively, against the forecast lead time. The RMEs are shown for different combinations of basin sizes and precipitation thresholds. The most salient feature in Figure 3 is the general tendency of the GEFSRv2 to underforecast across lead times (note that a negative RME indicates underforecasting whereas a positive RME indicates overforecasting in the ensemble mean). The underforecasting tends to increase with both the lead time and the precipitation threshold but decreases with the basin size. Contrasting the 6 (Figures 3a-c) and 24 (Figures 3d-f) hourly accumulations, the RMEs tend to be relatively similar for moderate and heavy precipitation; light precipitation shows somewhat larger negative RMEs for the 24 hourly accumulations. Further, the increase in the negative RMEs across lead times is rather smooth for the 24 hourly accumulations (Figure 3b) whereas the 6 hourly (Figure 3a) accumulations tend to exhibit a daily oscillation. One reason for this oscillation may be due to the lesser ability of the GEFSRv2 model to predict convective phenomena (Baxter et al., 2014), since the peaks occur at lead times of 6, 12, 30, 36, 54, 60, …, 366, and 372 hours, which correspond to midnight and early morning hours in standard time. However, this oscillation could be related to other factors, for which a more careful investigation is warranted.

As another measure of forecast quality, we show in Figures 4a-c and 4d-f the correlation coefficient between the GEFSRv2 ensemble mean forecast and the corresponding observed precipitation values versus the forecast lead time for the 6 and 24 hourly accumulations, respectively. Note that a higher correlation coefficient implies a lower mean squared error, other factors being equal (Wilks, 2011). Overall, in Figure 4, the correlations associated with a given combination of precipitation threshold and basin size tend to decrease with forecast lead time. Specifically, at lead times below approximately 180 hours (7.5 days), the correlations in Figure 4 tend to increase both as the precipitation threshold decreases and as the basin size increases. At lead times greater than 180 hours, the individual curves converge towards each other and approach 0, thus indicating a lesser sensitivity of the correlations to both the precipitation threshold and basin size for the longer lead times, where the forecasts lack any predictive ability. With regard to the aggregation period, the correlation coefficients are smaller for the 6 (Figure 4a-c) than 24 (Figure 4d-f) hourly accumulations across the different forecast lead times. The results in Figure 4 suggest the potential for the GEFSRv2 to provide skillful forecasts for lead times of up to 7 days.

To examine the skill of the GEFSRv2 precipitation forecasts, we plot in Figures 5a-c and 5d-f the CRPSS (relative to sample climatology) versus the forecast lead time for the 6 and 24 hourly accumulations, respectively. In Figure 5, at lead times less than approximately 180 hours (7.5 days), the forecasts tend to show, for a given precipitation threshold, higher skill as the basin size increases, although the differences in forecast skill are very small between the 300 x 300 km^2 (Figures 5b and 5e) and 500 x 500 km^2 (Figures 5c and 5f) basin size. At lead times greater than 180 hours, the CRPSS behaves differently; the skill increases as the precipitation threshold

decreases while the basin size plays a lesser role. Nonetheless, at these larger lead times (beyond 180 hours), the nominal values and sampling uncertainty of the CRPSS tend to show zero or negative skill, with the exception of light precipitation. For the light precipitation amounts, the CRPSS remains positive in Figure 5 at all forecast lead times. Relative to the accumulation period, the results are relatively similar for the 6 (Figure 5a-c) and 24 (Figure 5d-f) hourly accumulations, although the 24 hourly accumulations tend to be somewhat more skillful.

To assess the reliability of the GEFSRv2 precipitation forecasts, we plot in Figures 6a-i reliability diagrams for different lead times (i.e., 1, 3, and 6 days) and basin sizes (i.e., 100 x 100, 300 x 300 and 500 x 500 km^2). Overall, the precipitation forecasts in Figure 6 are somewhat overconfident at high forecast probabilities and underconfident at low forecast probabilities. To some degree, the biases depend on the precipitation threshold, lead time, and basin size. For example, the biases are generally smaller for light precipitation amounts in the large basin size (Figure 6i). Overall, Figure 6 highlights the potential for post-processing as a way to reduce these conditional biases (see, e.g., Gneiting et al. (2005) and Raftery et al. (2005)).

To explore the seasonal quality of the forecasts, we plot in Figures 7a and 7b the BSS against the calendar months for light and heavy precipitation, respectively, using 24 hourly accumulations and a basin size of 500 x 500 km^2. The BSS for each calendar month is calculated using the sample climatology as the reference. The sample climatology is determined using the entire period of analysis. Notwithstanding the large sampling uncertainty in Figure 7 as indicated by the wide confidence intervals, it shows that the skill of the precipitation forecasts is generally higher for the winter months (December-February) and that it tends to decline for the remaining months, reaching the lowest value in the summer months (June-August). The fall months (September-November) tend to have a slightly higher skill than the spring months (March-May). This seasonal pattern in the BSS is accentuated somewhat as the lead time increases. For instance, at a lead time of 7 days, the BSS shows little or no skill for both light (Figure 7a) and heavy (Figure 7b) precipitation in the month of August whereas at a lead time of 1 day the forecasts tend to be better than sample climatology across all months. Similar seasonal behavior to that identified here has been reported by others for the GEFSRv2 precipitation forecasts (Baxter et al., 2014; Brown, 2014; Hamill et al., 2013).

In summary, the results indicate that the GEFSRv2 provides skillful precipitation forecasts up to lead times of 6 or 7 days. Beyond these lead times, the precipitation forecasts for heavy precipitation show little or no skill while light to moderate precipitation tends to show some skill. Further, the forecasts of heavy precipitation are strongly biased and the bias can vary markedly depending on the basin size, accumulation period, and lead time.

4.2 Verification of the SREF precipitation forecasts

To examine the accuracy of the SREF precipitation forecasts, Figures 8a and 8b show box plots of the forecast error (forecast minus observed) against increasing amounts of observed precipitation for lead times of 1 and 3 days, respectively. The box plots are for 24 hourly accumulations and a 100 x 100 km^2 basin size. Overall, the box plots indicate that light precipitation events may be slightly overestimated while heavy precipitation events are clearly underestimated.

As another measure of bias, Figures 9a-c plots the RME of the SREF ensemble mean forecast against the forecast lead time for different basin sizes and 6 hourly accumulations. In Figures 9a-c, the RMEs tend to decrease (i.e., it becomes more negative) with both increasing lead time and decreasing basin size for moderate and heavy precipitation while it remains

relatively constant across lead times for light precipitation. The RMEs for the 24 hourly accumulations (not shown) are qualitatively similar to the RMEs in Figure 9a-c. Both Figures 8 and 9a-c indicate that the bias of the SREF precipitation forecasts declines for the larger precipitation thresholds. In terms of the correlation coefficient (Figures 9d-f), the basin size seems more dominant than the precipitation threshold in determining the quality of the precipitation forecasts, e.g., the correlation coefficient is lowest in Figure 9d for the 100 x 100 km^2 basin size, independently of the precipitation threshold. Further, the correlation coefficient increases for all precipitation thresholds as the basin size increases. However, the CRPSS indicates (Figure 9g-i) that both the basin size and precipitation threshold influence the forecast skill, as the CRPSS is sensitive to the bias as well as the correlation (and other attributes of forecast quality). Indeed, the SREF precipitation forecasts for moderate to heavy precipitation and a large basin size (500 x 500 km^2) are the most skillful in this case (Figure 9i).

Regarding the reliability of the SREF precipitation forecasts, they consistently show overforecasting across lead times, precipitation thresholds, and basin sizes (Figures 10a-i). The overforecasting diminishes somewhat for the smaller forecast probabilities and as the precipitation threshold decreases. By examining the inset plots in Figure 10a-i, the forecasts for heavy precipitation exhibit less confidence (i.e., are less sharp) than the forecasts for light to moderate precipitation. As the slope of the estimated reliability curves in Figure 10 tends to be less than the 1:1 reference line, the precipitation forecasts are overconfident. This means that, in calibrating the forecasts, the highest forecast probabilities will need to be adjusted downwards. In terms of the seasonal skill, the SREF precipitation forecasts reveal both differences and similarities in the monthly BSS between light (Figure 11a) and heavy precipitation (Figure 11b). For light precipitation (Figure 11a), there is little to no skill in the month of July while the remaining months show a relatively constant level of skill, with the fall months being slightly less skillful than the winter months. For heavy precipitation (Figure 11b), there is little to no skill during the late spring and early summer months (warm season, April-July) while the cool season is consistently skillful across the months of August-March. We note that qualitatively similar patterns have been identified before for precipitation forecasts from the SREF (Brown et al., 2012).

In summary, the SREF precipitation forecasts show more skill during the cool season than the warm season. The skill for moderate to heavy precipitation tends generally to be better than for light precipitation. Nonetheless, depending on the basin size and lead time, there are strong biases in the precipitation forecasts at any precipitation threshold.

4.3 Verification of the GEFSRv2 and SREF precipitation forecasts over the same time period of analysis

Here we present and discuss the verification results for the GEFSRv2 and SREF over a consistent period of record (i.e., the shorter period of the SREF). This is useful to understand and assess the relative quality of the two forecasting systems. To this end, we show in Figures 12a-c and 12d-f the skill of both systems using the BSS and CRPSS, respectively, plotted against precipitation thresholds (i.e., climatological non-exceedance probabilities) for 24 hourly accumulations and a basin size of 500 x 500 km^2. Also, note in Figure 12 that the forecast skill can only be compared for lead times out to 87 hours since this is the forecast cycle of the SREF. Thus, in terms of the BSS (Figure 12a), the GEFSRv2 is somewhat more skillful than the SREF for light (Pr=0.5) to heavy (Pr=0.99) precipitation amounts and at lead times of 1-3 days. However, the skill of both systems drops relatively quickly for climatological non-exceedance

probabilities larger than 0.9. Furthermore, the bootstrap confidence intervals for the GEFSRv2 and SREF in Figure 12a-c tend to overlap each other, thereby suggesting that the differences in forecast skill may not be significant. In terms of the CRPSS (Figure 12d-f), the skill of both systems shows a gradual increase with increasing precipitation threshold. However, the SREF tends to show better CRPSS across all precipitation thresholds, notwithstanding some slightly lower skill at a lead time of 3 days for the lower precipitation thresholds. The results shown in Figure 12 are qualitatively similar (not shown) for the other basins sizes.

To assess the ability of the forecasting systems to discriminate between precipitation events, we plot in Figures 13a-i the ROC curves for the GEFSRv2 and SREF for different lead times (i.e., 1-3 days) and precipitation thresholds (i.e., light, moderate, and heavy precipitation). All the ROC curves are for 24 hourly accumulations and a 100 x 100 km^2 basin size. Note that ROC curves show the ability of the forecasts to discriminate between the occurrence and non-occurrence of a precipitation event across different forecast probability thresholds. All the curves in Figure 13 lie between the climatological curve (45 degree diagonal line connecting the points [0,0] and [1,1]) and that associated with a perfect forecasting system (perpendicular line connecting the points [0,0], [0,1], and [1,1]), thus emphasizing that the ROC curves have some ability to discriminate between precipitation events. For the GEFSRv2, the forecasts for heavy precipitation show relatively better discrimination than the forecasts for light to moderate precipitation for the larger probabilities of false detection. This trend holds true for all the lead times. On the other hand, the SREF shows more or less similar discrimination for all the precipitation thresholds at a given lead time. However, the GEFSRv2 exhibits significantly better discrimination than the SREF for light, moderate, and heavy precipitation at all lead times. The ROC curves in Figure 13 are qualitatively similar (not shown) for the other basins sizes.

Overall, both the GEFSRv2 and SREF show similar skill in forecasting moderate and heavy precipitation at lead times of 1-3 days, when the same period of analysis is considered. The GEFSRv2 is better at discriminating between the occurrence and non-occurrence of a given precipitation amount, including small and large amounts, than the SREF.

5. Summary and conclusions

In this study, we assessed the quality of precipitation forecasts in the MAR from two ensemble forecasting systems, namely the GEFSRv2 and SREF. Using various verification metrics (e.g., box plots of the error between forecast and observed precipitation, RME, correlation coefficient, CRPSS, BSS, reliability diagram, and ROC), the forecast quality of these two systems was evaluated conditionally upon precipitation amounts, basin size, forecast lead times, different accumulation periods, and seasonality. Throughout the verification analysis, we used 3 precipitation thresholds to represent light (Pr=0.5), moderate (Pr=0.9), and heavy precipitation (Pr=0.99), as well as 3 different basin sizes to represent small (100 x 100 km^2), intermediate (300 x 300 km^2), and large (500 x 500 km^2) basins. On the basis of the verification results obtained, we emphasize the following:

- The GEFSRv2 ensemble forecasts show good forecast skill for light to moderate precipitation while the forecasts for heavy precipitation are consistently too low. This trend is apparent for both the 6 and 24 hourly accumulations, although the 24 hourly forecasts are slightly more skillful. For heavy precipitation amounts in the MAR, the GEFSRv2 forecast skill becomes relatively small after lead times of 5 or 6 days and negligible (i.e., approximately equal to climatology) after 9 days. The GEFSRv2 forecasts become more reliable with increasing basin size and decreasing precipitation

threshold. However, the general tendency is for the GEFSRv2 to underestimate the smaller forecast probabilities and overestimate the larger forecast probabilities.

- As with the GEFSRv2, the SREF also tends to overforecast light to moderate precipitation amounts while it largely underforecasts the heavy precipitation amounts. The magnitude of the SREF forecast errors increase with increasing lead time and precipitation threshold.
- Similar to the GEFSRv2, the SREF forecasts become more reliable with increasing basin size and decreasing precipitation threshold. Generally, the SREF forecasts are more reliable than the equivalent GEFSRv2 forecasts. Nonetheless, the overall tendency is for the SREF to overestimate the moderate and high forecast probabilities.
- Seasonal trends are visible in both the GEFSRv2 and SREF forecasts. Generally, the forecasts from both systems exhibit more skill during the cool season than the warm season. This trend tends to be similar across forecast lead times, basin sizes, and precipitation thresholds.

The verification results from this study compare well against previous findings for the same forecasting systems (Brown et al, 2012; Brown, 2014). Overall, we find that the precipitation forecasts from the GEFSRv2 and SREF show comparable quality and skill for the short-range forecasts (i.e., lead times $\leq$ 3 days). However, the CRPSS indicates that the SREF forecasts are slightly more skillful than the GEFSRv2 and the GEFSRv2 reveals better discrimination than the SREF for moderate and heavy precipitation. It thus seems plausible that an optimal combination of these two systems could contribute to improving the overall skill of the precipitation forecasts. Our verification analysis (e.g., the reliability diagrams) also indicates that the quality of the precipitation forecasts could be further improved by employing statistical post-processing techniques. This should be further investigated using different post-processing techniques, as this could be an important source for gaining additional forecast skill and reliability. Both the GEFSRv2 and SREF show higher forecast skill for the larger basins (500 x 500 km^2) than the small (100 x 100 km^2) and intermediate (300 x 300 km^2) basins, irrespective of the precipitation threshold. This suggests the possibility of generating better streamflow forecasts for large basins than small ones, other factors being equal, but this will need to be established through streamflow hindcasting and verification. Further, the quality of streamflow forecasts across basin sizes will depend on the performance and type of hydrologic model, e.g., spatially lumped or distributed, used for generating the forecasts. This also will need to be investigated.

To continue advancing this research, we plan to explore and evaluate various forecasting scenarios to assess the benefits of integrating the outputs from different precipitation forecasting systems (e.g., GEFSRv2 and SREF), application of post-processing techniques, and different hydrologic model structures to potentially improve flood forecasting across spatiotemporal scales.

References

Adams, T., Ostrowski, J., 2010. Short Lead-Time Hydrologic Ensemble Forecasts from Numerical Weather Prediction Model Ensembles, World Environmental and Water Resources Congress 2010. American Society of Civil Engineers, pp. 2294-2304. DOI:doi:10.1061/41114(371)237 10.1061/41114(371)237

Ator, S.W., Denver, J.M., Krantz, D.E., Newell, W.L., Martucci, S.K., 2005. A surficial hydrogeologic framework for the Mid-Atlantic Coastal Plain, US Department of the Interior, US Geological Survey.

Baxter, M.A., Lackmann, G.M., Mahoney, K.M., Workoff, T.E., Hamill, T.M., 2014. Verification of Quantitative Precipitation Reforecasts over the Southeastern United States. Weather and Forecasting, 29(5): 1199-1207. DOI:10.1175/WAF-D-14-00055.1

Bouwer, L.M., 2010. Have Disaster Losses Increased Due to Anthropogenic Climate Change? Bulletin of the American Meteorological Society, 92(1): 39-46. DOI:10.1175/2010BAMS3092.1

Breidenbach, J.P., Bradberry, J.S., 2001. Multisensor precipitation estimates produced by National Weather Service River Forecast Centers for hydrologic applications.

Brier, G.W., Allen, R.A., 1951. Verification of weather forecasts. Compendium of meteorology: 841-848.

Brown, J., 2014. Verification of temperature, precipitation and streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service: an evolution of the medium-range forecasts with forcing inputs from NCEP's Global Ensemble Forecast System (GEFS) and a comparison to the frozen version of NCEP's Global Forecast System (GFS), Hydrologic Solutions Limited, Subcontract Agreement 2013-09 with LEN Technologies Inc.

Brown, J., Demargne, J., Seo, D.-J., Liu, Y., 2010. The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations Environmental Modeling and Software, 25(2010): 854-872.

Brown, J.D., Seo, D.-J., Du, J., 2012. Verification of Precipitation Forecasts from NCEP's Short-Range Ensemble Forecast (SREF) System with Reference to Ensemble Streamflow Prediction Using Lumped Hydrologic Models. Journal of Hydrometeorology, 13(3): 808-836. DOI:10.1175/JHM-D-11-036.1

Brown, J.D. et al., 2014. Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. Journal of Hydrology(0). DOI:http://dx.doi.org/10.1016/j.jhydrol.2014.05.028

Buizza, R., 2008. The value of probabilistic prediction. Atmospheric Science Letters, 9(2): 36-42. DOI:10.1002/asl.170

Buizza, R. et al., 2005. A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. Monthly Weather Review, 133(5): 1076-1097. DOI:10.1175/MWR2905.1

Changnon, S.A., 1998. Letters to the editor-Comments on Secular Trends of Precipitation Amount, Frequency, and Intensity in the United States. Bulletin of the American Meteorological Society, 79(11): 2550-2551

Charles, M.E., Colle, B.A., 2009. Verification of extratropical cyclones within the NCEP operational models. Part II: The short-range ensemble forecast system. Weather and Forecasting, 24(5): 1191-1214

Choi, O., Fisher, A., 2003. The impacts of socioeconomic development and climate change on severe weather catastrophe losses: Mid-Atlantic Region (MAR) and the US. Climatic Change, 58(1-2): 149-170

Cloke, H.L., Pappenberger, F., 2009. Ensemble flood forecasting: A review. Journal of Hydrology, 375(3–4): 613-626. DOI:http://dx.doi.org/10.1016/j.jhydrol.2009.06.005

Davolio, S. et al., 2008. A meteo-hydrological prediction system based on a multi-model approach for precipitation forecasting. Natural Hazards and Earth System Science, 8(1): 143-159

Demargne, J. et al., 2010. - Diagnostic verification of hydrometeorological and hydrologic ensembles. - 11(- 2): - 122

Demeritt, D. et al., 2007. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. Environmental Hazards, 7(2): 115-127. DOI:http://dx.doi.org/10.1016/j.envhaz.2007.05.001

Downton, M., Miller, J., Pielke, R., 2005. Reanalysis of U.S. National Weather Service Flood Loss Database. Natural Hazards Review, 6(1): 13-22. DOI:10.1061/(ASCE)1527-6988(2005)6:1(13)

Du, J. et al., 2009. NCEP short-range ensemble forecast (SREF) system upgrade in 2009

Du, J., DiMego, G., Tracton, S., Zhou, B., 2003. NCEP Short-Range Ensemble Forecasting (SREF) System: Multi-IC, Multi-model and Multi-physics approach

Du, J. et al., 2006. New dimension of NCEP short-range Ensemble forecasting (SREF) system: inclusion of WRF members

Epstein, E.S., 1969. Stochastic dynamic prediction1. Tellus, 21(6): 739-759. DOI:10.1111/j.2153-3490.1969.tb00483.x

Fan, F.M., Collischonn, W., Meller, A., Botelho, L.C.M., 2014. Ensemble streamflow forecasting experiments in a tropical basin: The São Francisco river case study. Journal of Hydrology(0). DOI:http://dx.doi.org/10.1016/j.jhydrol.2014.04.038

Fenneman, N.M., Johnson, D.W., 1946. Physical divisions of the United States. US Geological Survey.

Gneiting, T., Raftery, A.E., Westveld Iii, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Monthly Weather Review, 133(5): 1098-1118

Green, D.M., Swets, J.A., 1966. Signal detection theory and psychophysics, 1. Wiley New York

Greene, D.R., Hudlow, M.D., Farnsworth, R.K., 1979. A multiple sensor rainfall analysis system, pp. 44-53

Hamill, T.M. et al., 2013. NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. Bulletin of the American Meteorological Society, 94(10): 1553-1565. DOI:10.1175/BAMS-D-12-00014.1

Hamill, T.M., Colucci, S.J., 1998. Evaluation of Eta–RSM Ensemble Probabilistic Precipitation Forecasts. Monthly Weather Review, 126(3): 711-724. DOI:10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2

Hamill, T.M., Whitaker, J.S., 2006. Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. Monthly Weather Review, 134(11): 3209-3229. DOI:10.1175/MWR3237.1

Hamill, T.M., Whitaker, J.S., Mullen, S.L., 2006. Reforecasts: An Important Dataset for Improving Weather Predictions. Bulletin of the American Meteorological Society, 87(1): 33-46. DOI:10.1175/BAMS-87-1-33

He, Y. et al., 2009. Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. Meteorological Applications, 16(1): 91-101. DOI:10.1002/met.132

Hegger, D.T. et al., 2014. Assessing Stability and Dynamics in Flood Risk Governance. Water Resources Management, 28(12): 4127-4142. DOI:10.1007/s11269-014-0732-x

Jolliffe, I.T., Stephenson, D.B., 2012. Forecast verification: a practitioner's guide in atmospheric science. John Wiley & Sons.

Jones, M.S., Colle, B.A., Tongue, J.S., 2007. Evaluation of a Mesoscale Short-Range Ensemble Forecast System over the Northeast United States. Weather and Forecasting, 22(1): 36-55. DOI:10.1175/WAF973.1

Karl, T.R., Knight, R.W., Easterling, D.R., Quayle, R.G., 1996. Indices of Climate Change for the United States. Bulletin of the American Meteorological Society, 77(2): 279-292. DOI:10.1175/1520-0477(1996)077<0279:IOCCFT>2.0.CO;2

Kelsch, M., Caporali, E., Lanza, L., 2001. Hydrometeorology of Flash Floods. In: Gruntfest, E., Handmer, J. (Eds.), Coping With Flash Floods. NATO Science Series. Springer Netherlands, pp. 19-35. DOI:10.1007/978-94-010-0918-8_4

Kobold, M., Suselj, K., 2005. Precipitation forecasts and their uncertainty as input into hydrological models. Hydrology and Earth System Sciences Discussions, 9(4): 322-332.

Kundzewicz, Z.W. et al., 2014. Flood risk and climate change: global and regional perspectives. Hydrological Sciences Journal, 59(1): 1-28

McCarthy, J.J., 2001. Climate Change 2001: Impacts, Adaptation, and Vulnerability: Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Climate change 2001. Cambridge University Press

Milly, P.C.D., Wetherald, R.T., Dunne, K.A., Delworth, T.L., 2002. Increasing risk of great floods in a changing climate. Nature, 415(6871): 514-517

Montz, B.E., Gruntfest, E., 2002. Flash flood mitigation: recommendations for research and applications. Global Environmental Change Part B: Environmental Hazards, 4(1): 15-22. DOI:http://dx.doi.org/10.1016/S1464-2867(02)00011-6

Morss, R.E., Wilhelmi, O.V., Downton, M.W., Gruntfest, E., 2005. Flood Risk, Uncertainty, and Scientific Information for Decision Making: Lessons from an Interdisciplinary Project. Bulletin of the American Meteorological Society, 86(11): 1593-1601. DOI:10.1175/BAMS-86-11-1593

Murphy, A.H., 1993. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. Weather and Forecasting, 8(2): 281-293. DOI:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

Murphy, A.H., Epstein, E.S., 1989. Skill Scores and Correlation Coefficients in Model Verification. Monthly Weather Review, 117(3): 572-582. DOI:10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2

Murphy, A.H., Winkler, R.L., 1987. A General Framework for Forecast Verification. Monthly Weather Review, 115(7): 1330-1338. DOI:10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

Najjar, R.G. et al., 2000. The potential impacts of climate change on the mid-Atlantic coastal region. Climate Research, 14(3): 219-233

Neff, R. et al., 2000. Impact of climate variation and change on Mid-Atlantic Region hydrology and water resources. Climate Research, 14(3): 207-218

Novak, D.R. et al., 2013. Precipitation and Temperature Forecast Performance at the Weather Prediction Center. Weather and Forecasting, 29(3): 489-504. DOI:10.1175/WAF-D-13-00066.1

Olsson, J., Lindström, G., 2008. Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. Journal of Hydrology, 350(1–2): 14-24. DOI:http://dx.doi.org/10.1016/j.jhydrol.2007.11.010

Pagano, T.C. et al., 2014. Challenges of Operational River Forecasting. Journal of Hydrometeorology, 15(4): 1692-1707. DOI:10.1175/JHM-D-13-0188.1

Pall, P. et al., 2011. Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. Nature, 470(7334): 382-385. DOI:http://www.nature.com/nature/journal/v470/n7334/abs/10.1038-nature09762-unlocked.html#supplementary-information

Pappenberger, F., Buizza, R., 2009. The Skill of ECMWF Precipitation and Temperature Predictions in the Danube Basin as Forcings of Hydrological Models. Weather and Forecasting, 24(3): 749-766. DOI:10.1175/2008WAF2222120.1

Pappenberger, F., Scipal, K., Buizza, R., 2008. Hydrological aspects of meteorological verification. Atmospheric Science Letters, 9(2): 43-52. DOI:10.1002/asl.171

Pielke, R.A., Downton, M.W., 2000. Precipitation and Damaging Floods: Trends in the United States, 1932–97. Journal of Climate, 13(20): 3625-3637. DOI:10.1175/1520-0442(2000)013<3625:PADFTI>2.0.CO;2

Politis, D.N., Romano, J.P., 1994. The Stationary Bootstrap. Journal of the American Statistical Association, 89(428): 1303-1313. DOI:10.1080/01621459.1994.10476870

Polsky, C., Allard, J., Currit, N., Crane, R., Yarnal, B., 2000. The Mid-Atlantic Region and its climate: past, present, and future. Climate Research, 14(3): 161-173

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 133(5): 1155-1174

Ramos, M.-H., Mathevet, T., Thielen, J., Pappenberger, F., 2010. Communicating uncertainty in hydro-meteorological forecasts: mission impossible? Meteorological Applications, 17(2): 223-235. DOI:10.1002/met.202

Rutledge, A.T., Mesko, T.O., 1996. Estimated hydrologic characteristics of shallow aquifer systems in the Valley and Ridge, the Blue Ridge, and the Piedmont physiographic provinces based on analysis of streamflow recession and base flow. US Geological Survey Professional Paper (USA)

Schaake, J.C., Hamill, T.M., Buizza, R., Clark, M., 2007. HEPEX: The Hydrological Ensemble Prediction Experiment. Bulletin of the American Meteorological Society, 88(10): 1541-1547. DOI:10.1175/BAMS-88-10-1541

Schildgen, B., 1999. Unnatural Disasters: Areas that suffer repeat flooding yet continue to rebuild. Sierra

Schumacher, R.S., Davis, C.A., 2010. Ensemble-Based Forecast Uncertainty Analysis of Diverse Heavy Rainfall Events. Weather and Forecasting, 25(4): 1103-1122. DOI:10.1175/2010WAF2222378.1

Seo, D.-J., Breidenbach, J.P., 2002. Real-Time Correction of Spatially Nonuniform Bias in Radar Rainfall Data Using Rain Gauge Measurements. Journal of Hydrometeorology, 3(2): 93-111. DOI:10.1175/1525-7541(2002)003<0093:RTCOSN>2.0.CO;2

Seo, D.-J., Seed, A., Delrieu, G., 2010. Radar and Multisensor Rainfall Estimation for Hydrologic Applications, Rainfall: State of the Science. American Geophysical Union, pp. 79-104. DOI:10.1029/2010GM000952

Seo, D.J., 1998a. Real-time estimation of rainfall fields using rain gage data under fractional coverage conditions. Journal of Hydrology, 208(1–2): 25-36. DOI:http://dx.doi.org/10.1016/S0022-1694(98)00140-1

Seo, D.J., 1998b. Real-time estimation of rainfall fields using radar rainfall and rain gage data. Journal of Hydrology, 208(1–2): 37-52. DOI:http://dx.doi.org/10.1016/S0022-1694(98)00141-3

Sinnott, A., Cushing, E.M., 1978. Summary appraisals of the nation's ground-water resources--Mid-Atlantic Region [USA]. Professional Papers-US Geological Survey (USA). no. 813-I.

Thielen, J. et al., 2009. Monthly-, medium-, and short-range flood warning: testing the limits of predictability. Meteorological Applications, 16(1): 77-90. DOI:10.1002/met.140

Thiessen, A.H., 1911. Precipitation averages for large areas. Monthly Weather Review, 39(7): 1082-1089. DOI:10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2

United States Census Bureau, 2013. U.S. Department of Commerce

Verkade, J.S., Brown, J.D., Reggiani, P., Weerts, A.H., 2013. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. Journal of Hydrology, 501(0): 73-91. DOI:http://dx.doi.org/10.1016/j.jhydrol.2013.07.039

Voisin, N., Schaake, J.C., Lettenmaier, D.P., 2010. Calibration and Downscaling Methods for Quantitative Ensemble Precipitation Forecasts. Weather and Forecasting, 25(6): 1603-1627. DOI:10.1175/2010WAF2222367.1

Wardrop, D.H. et al., 2005. Use of landscape and land use parameters for classification and characterization of watersheds in the mid-Atlantic across five physiographic provinces. Environmental and Ecological Statistics, 12(2): 209-223. DOI:10.1007/s10651-005-1042-5

Wei, M., Toth, Z., Wobus, R., Zhu, Y., 2008. Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. Tellus A, 60(1): 62-79. DOI:10.1111/j.1600-0870.2007.00273.x

Whitaker, J.S., Hamill, T.M., Wei, X., Song, Y., Toth, Z., 2008. Ensemble Data Assimilation with the NCEP Global Forecast System. Monthly Weather Review, 136(2): 463-482. DOI:10.1175/2007MWR2018.1

Wilks, D.S., 2011. Statistical methods in the atmospheric sciences, 100. Academic press

Young, C.B., Bradley, A.A., Krajewski, W.F., Kruger, A., Morrissey, M.L., 2000. Evaluating NEXRAD Multisensor Precipitation Estimates for Operational Hydrologic Forecasting. Journal of Hydrometeorology, 1(3): 241-254. DOI:10.1175/1525-7541(2000)001<0241:ENMPEF>2.0.CO;2

Yuan, H. et al., 2005. Verification of Probabilistic Quantitative Precipitation Forecasts over the Southwest United States during Winter 2002/03 by the RSM Ensemble System. Monthly Weather Review, 133(1): 279-294. DOI:10.1175/MWR-2858.1

Zhu, Y., 2005. Ensemble forecast: A new approach to uncertainty and predictability. Advances in Atmospheric Sciences, 22(6): 781-788. DOI:10.1007/BF02918678

Table 1. Summary and main characteristics of the datasets used in this study.

	Horizontal Resolution [km^2]	**Total ensemble members**	**Lead time [hour]**	**Models**	**Period of analysis [years]**	**Projection system**
GEFSRv2	~55 x 55 (0.5^ox0.5^o)	11	1-192	1	2002-2013	Geographic coordinate system
	~73 x 73 (0.67^ox0.67^o)	11	193-384	1	2002-2013	Geographic coordinate system
SREF	~40 x 40	21	1-87	3/4	2008-2009	Lambert conic projection
	~32 x 32	21	1-87	3/4	2009 and 2011-2012	Lambert conic projection
	~16 x 16	21	1-87	3/4	2012-2013	Lambert conic projection
MPE	4 x 4	N/A	N/A	N/A	Same as forecasts	Polar stereographic

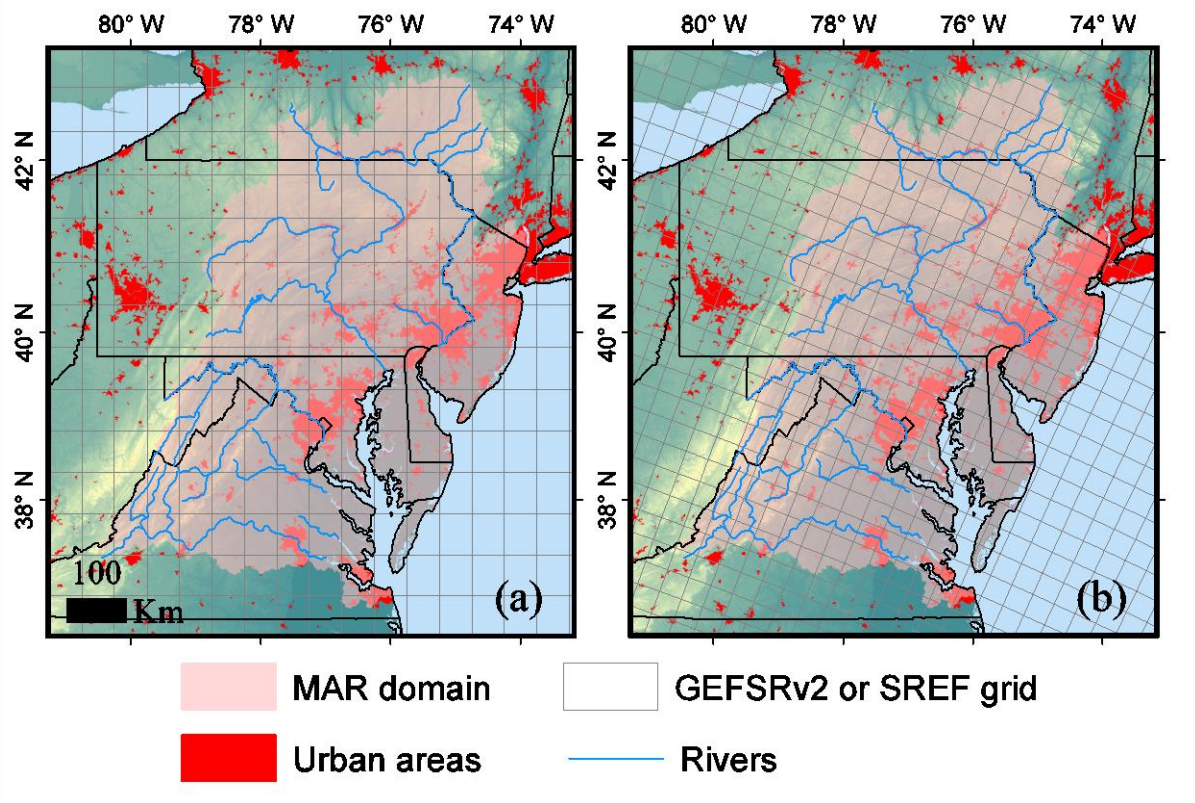


Figure 1. Illustration of the boundary of the MAR together with the grids for the (a) GEFSRv2 and (b) SREF. The boundary of the MAR corresponds in this study to the operating domain of the MARFC.

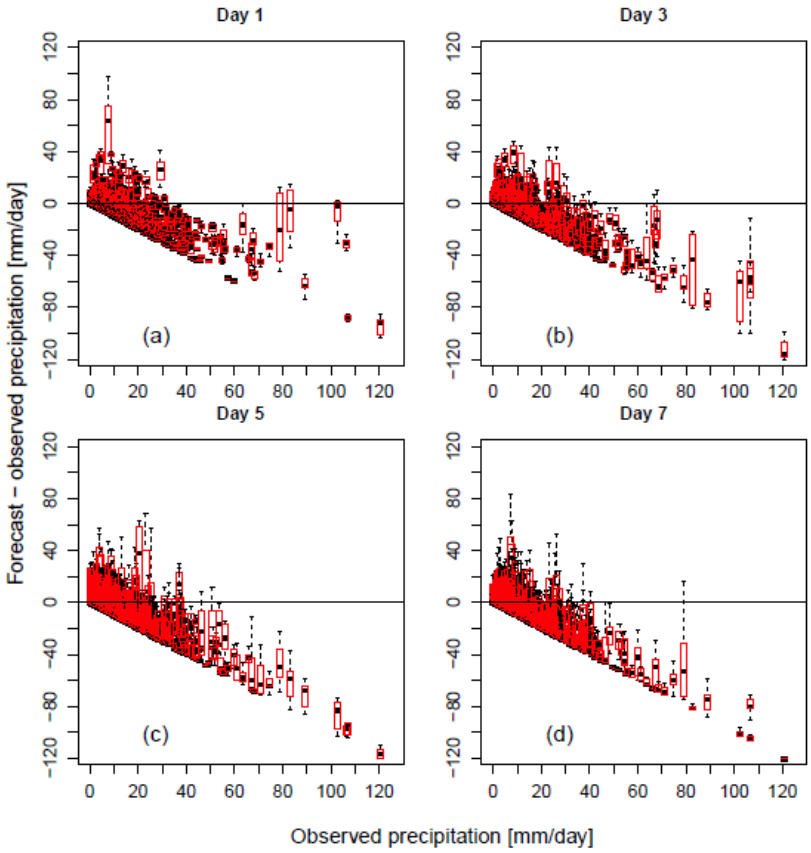


Figure 2. Box plots of errors in the GEFSRv2 precipitation forecasts, arranged according to the observed values, for lead times of (a) 1, (b) 3, (c) 5, and (d) 7 days. The box plots are for 24 hourly accumulations and a 100 x 100 km^2 basin size.

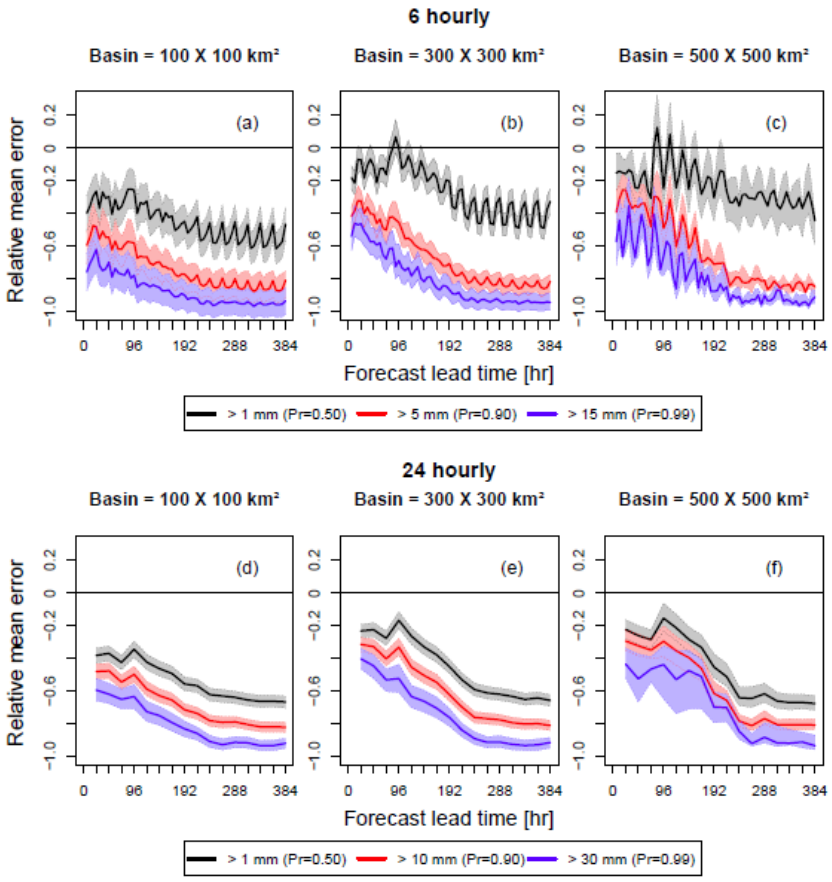


Figure 3. RME of the GEFSRv2 ensemble mean forecast versus the forecast lead time for (a)-(c) 6 and (d)-(f) 24 hourly accumulations. The RMEs are shown for different combinations of basins sizes and precipitation thresholds.

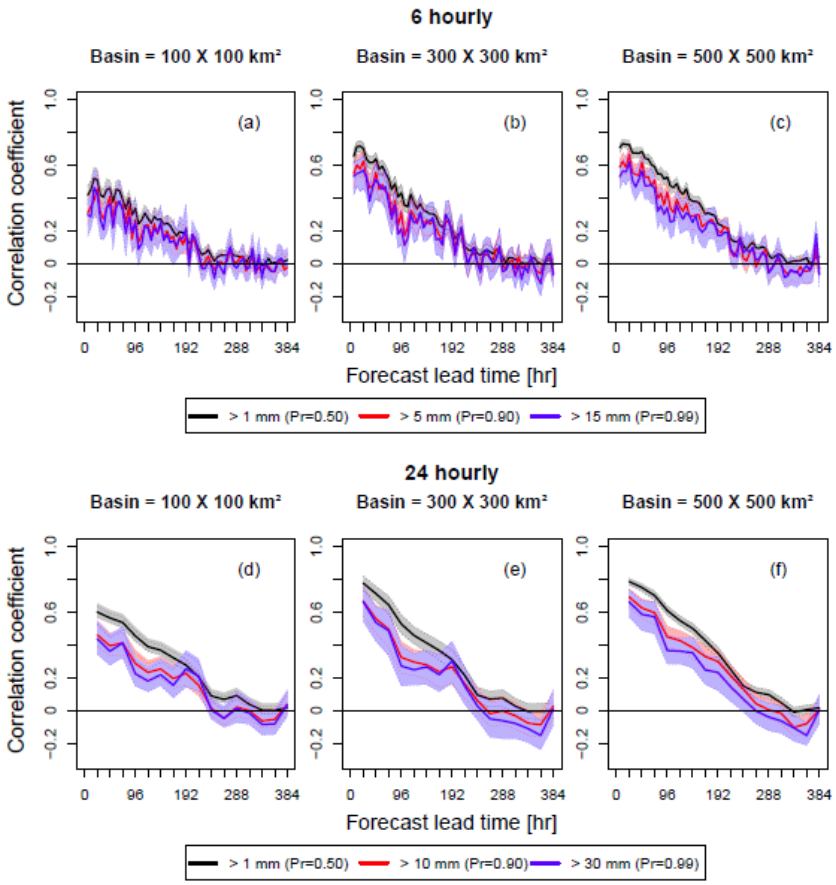


Figure 4. Correlation coefficient between the GEFSRv2 mean ensemble forecast and the corresponding observed precipitation values as a function of the forecast lead time for (a)-(c) 6 and (d)-(f) 24 hourly accumulations. The correlation coefficients are shown for different combinations of basin sizes and precipitation thresholds.

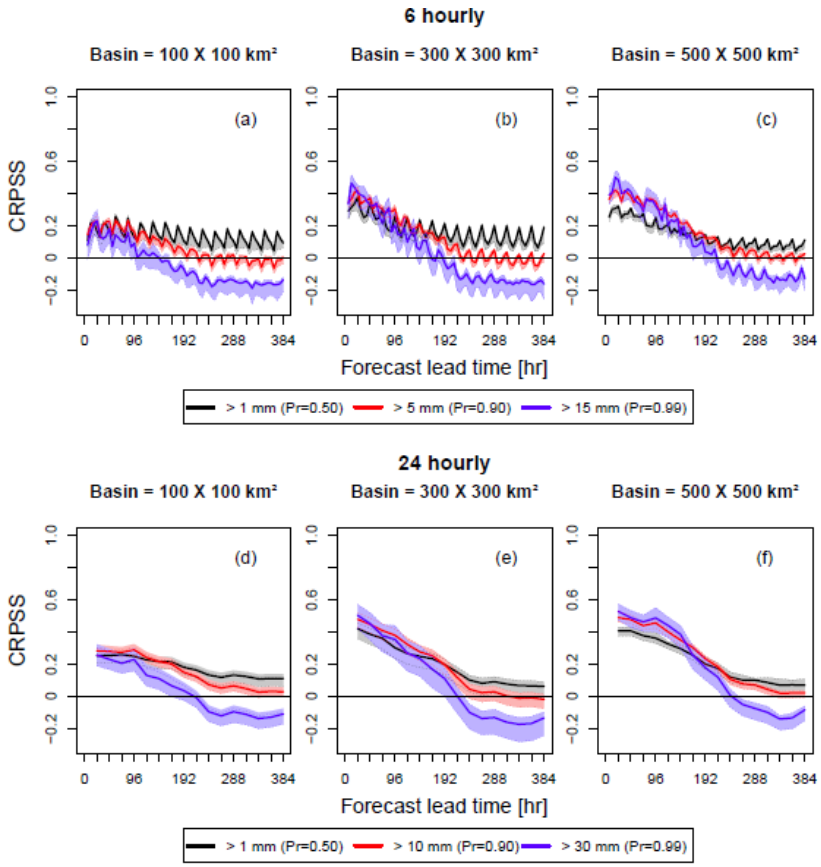


Figure 5. GEFSRv2 mean CRPSS versus the forecast lead time for (a)-(c) 6 and (d)-(f) 24 hourly accumulations. The values of the CRPSS are shown for different combinations of basin sizes and precipitation thresholds.

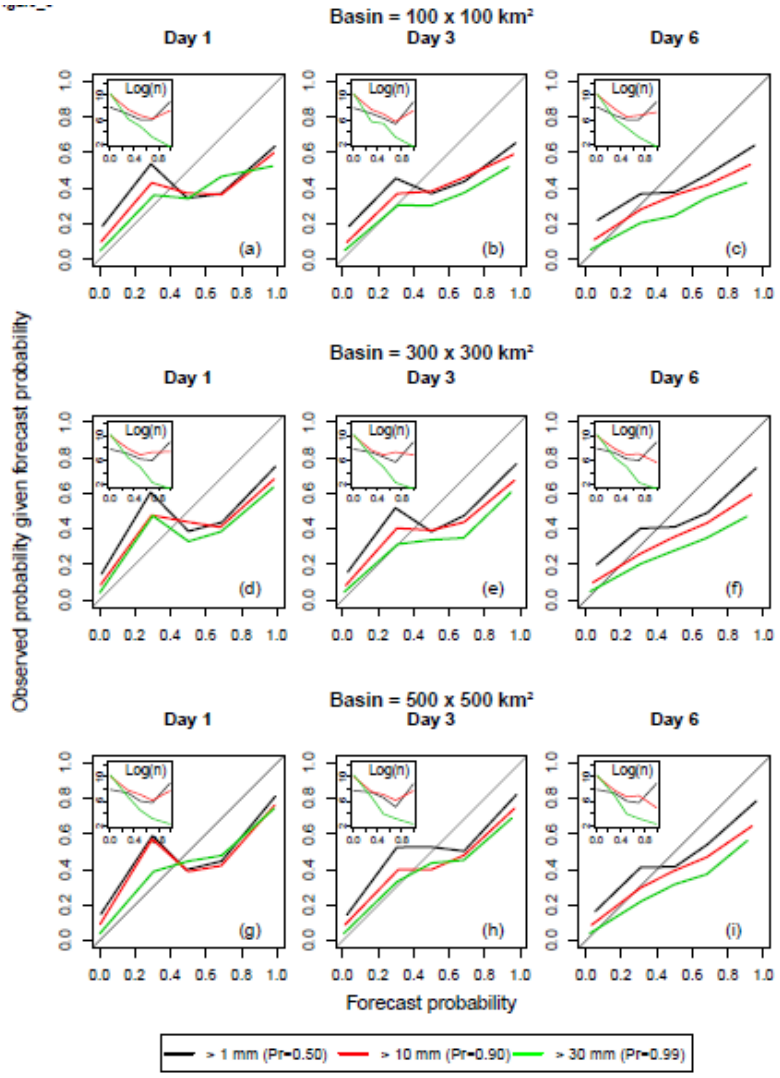


Figure 6. Reliability diagrams for the 24 hourly GEFSRv2 precipitation forecasts for lead times of (a) 1, (b) 3, (c) 6 days and a 100 x 100 km^2 basin size; (d) 1, (e) 3, (f) 6 days for a 300 x 300 km^2 basin size; and (g) 1, (h) 3, (i) 6 days for a 500 x 500 km^2 basin size. The insets show the sample size in logarithmic scale of the different forecast probability bins.

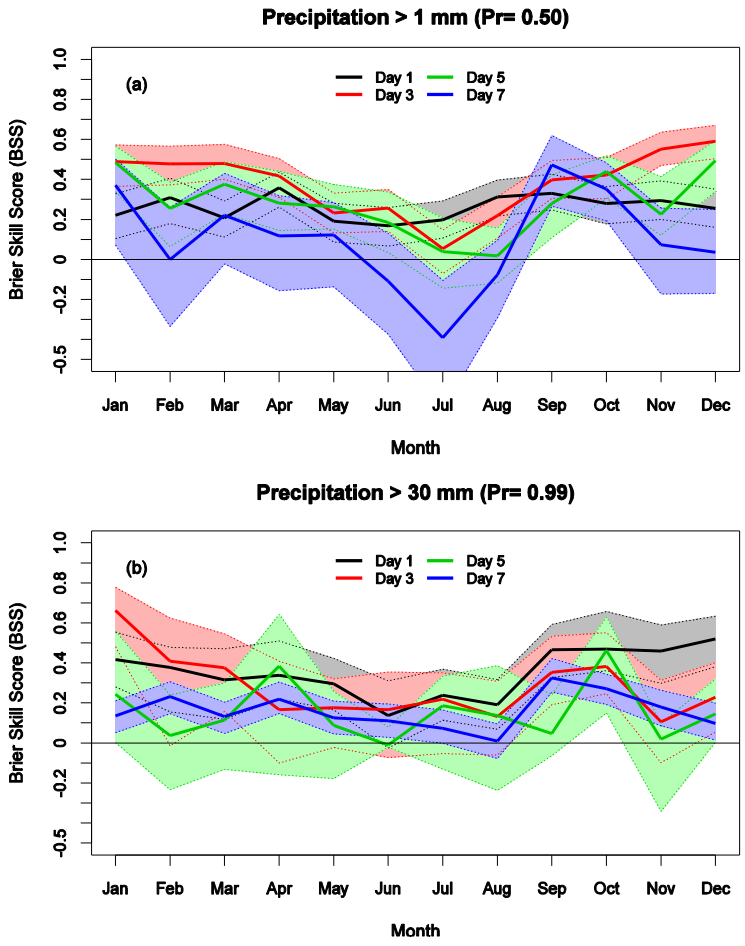


Figure 7. Monthly BSS values for the GEFSRv2 precipitation forecasts for (a) light and (b) heavy precipitation. The results are shown for various lead times (i.e., 1, 3, 5, and 7 days), 24 hourly accumulations, and a basin size of 500 x 500 km^2.

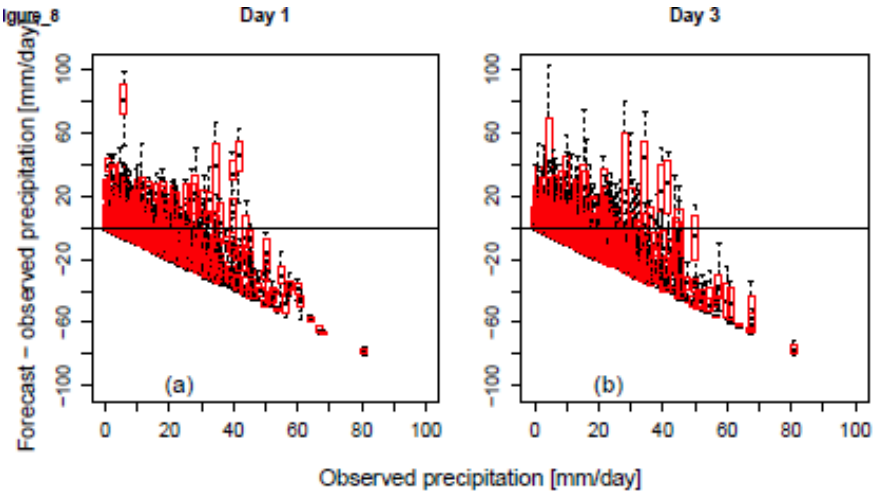


Figure 8. Box plots of errors in the SREF precipitation forecasts arranged according to the observed values for lead times of (a) 1 and (b) 3 days. The box plots are for 24 hourly accumulations and a 100 x 100 km2 basin size.

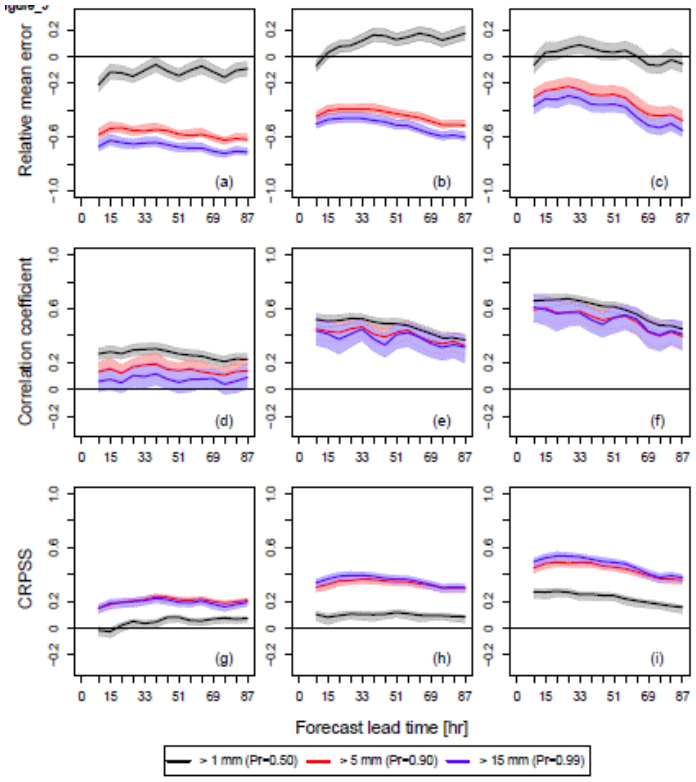


Figure 9. (a)-(c) RME, (d)-(f) correlation coefficient, and (g)-(i) CRPSS of the SREF ensemble mean forecast versus the forecast lead time for 6 hourly accumulations and different combinations of basin sizes (i.e., 100 x 100, 300 x 300, and 500 x 500 km^2) and precipitation thresholds (i.e., 1, 5, and 15 mm).

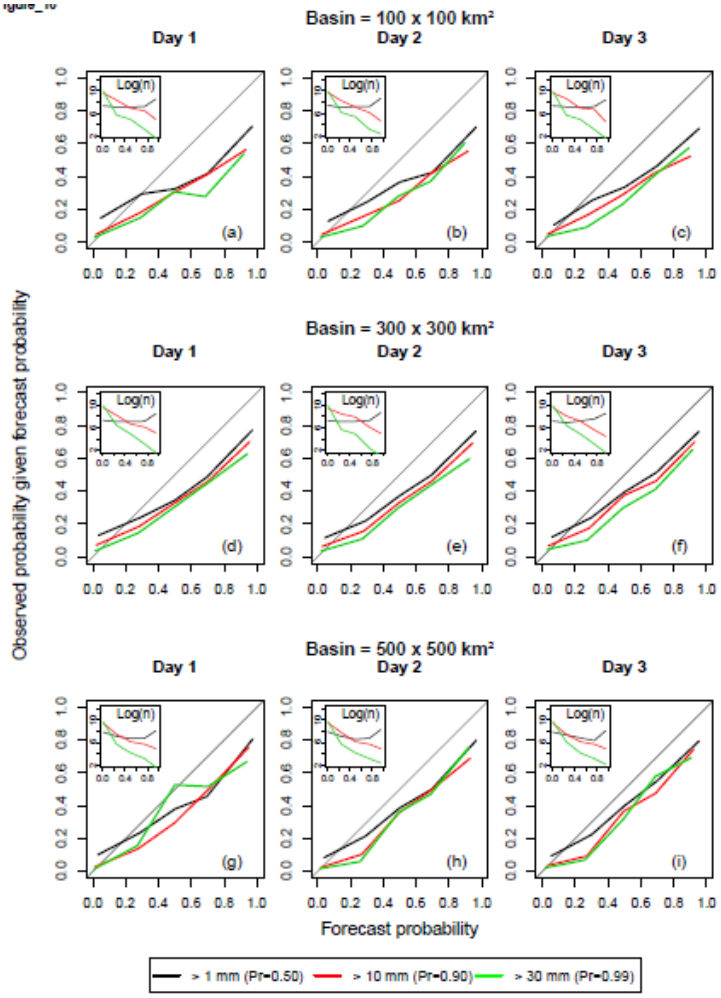


Figure 10. Reliability diagrams for the 24 hourly SREF precipitation forecasts for lead times of (a) 1, (b) 3, (c) 6 days and a 100 x 100 km^2 basin size; (d) 1, (e) 3, (f) 6 days for a 300 x 300 km^2 basin size; and (g) 1, (h) 3, (i) 6 days for a 500 x 500 km^2 basin size. The insets show the sample size in logarithmic scale of the different forecast probability bins.

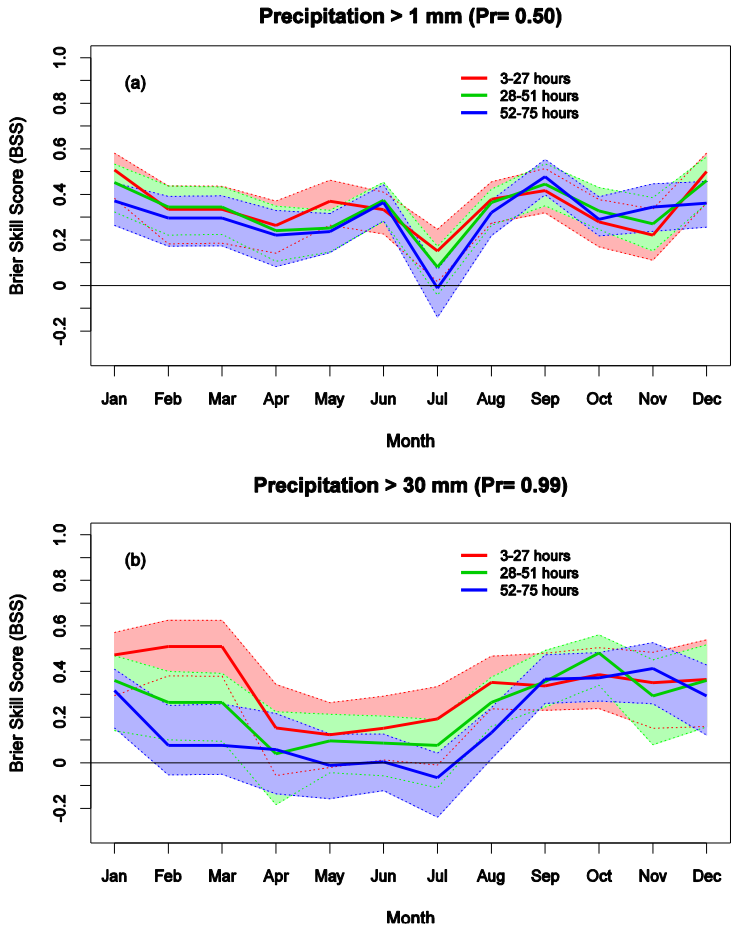


Figure 11. Monthly BSS values for the SREF precipitation forecasts for (a) light and (b) heavy precipitation. The results are shown for various lead times (i.e., 1, 3, 5, and 7 days), 24 hourly accumulations, and a basin size of 500 x 500 km^2.

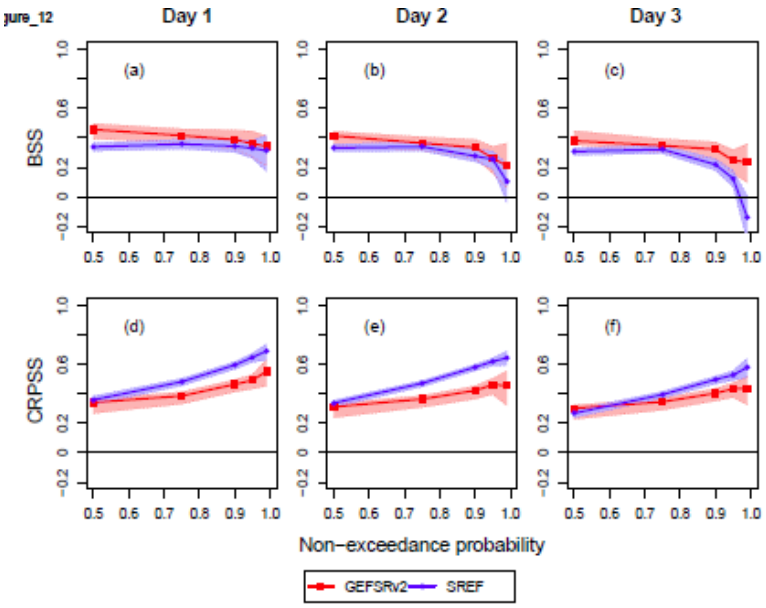


Figure 12. (a)-(c) BSS and (d)-(f) CRPSS versus the climatological non-exceedance probability of precipitation thresholds for both the GEFSRv2 and SREF precipitation forecasts. The skill metrics (BSS and CRPSS) are computed for 24 hourly accumulations and a 500 x 500 km^2 basin size.

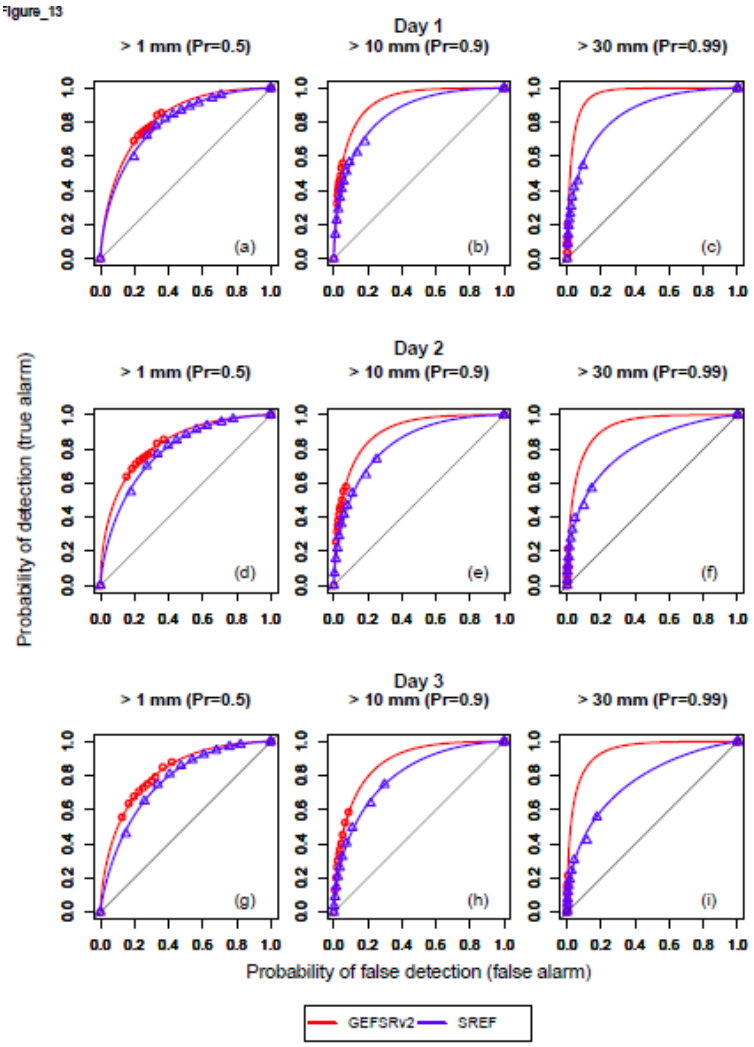


Figure 13. ROC curves for the GEFSRv2 and SREF precipitation forecasts at lead times of (a)-(c) 1, (d)-(f) 2, and (g)-(i) 3 days. The symbols represent the sample values of the PoD and PoFD, and the lines represent the values fitted under the binormal approximation. All the ROC curves shown are for the 24 hourly accumulations and a 100 x 100 km^2 basin size.

Chapter 3: Verification of precipitation ensembles from the GEFS, SREF, and WPC-PQPF over the eastern U.S.

ABSTRACT

The quality of ensemble precipitation forecasts across the eastern United States (U.S.) is investigated; specifically, the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System Reforecast version 2 (GEFSRv2) and Short Range Ensemble Forecast (SREF), as well as the NCEP's Weather Prediction Center probabilistic quantitative precipitation forecast (WPC-PQPF) guidance. The forecasts are verified using multi-sensor precipitation estimates. The verification is done using various metrics conditioned upon seasonality, precipitation threshold, lead time, and spatial aggregation scale. The forecasts are verified over the geographic domain of each of the four eastern River Forecasts Centers (RFCs) in the U.S. We first verify forecasts from i) all the three systems or guidance using a common period of analysis (2012-2013) for lead times from 1 to 3 days, and then ii) for the GEFSRv2 alone, using a longer period (2004-2013) and lead times from 1 to 16 days.

The verification results indicate that, across the eastern U.S., precipitation forecast bias decreases and the skill and reliability improve as the basin size increases; however, all the forecasts exhibit underforecasting bias. The skill of the forecasts is appreciably better in the cool season than in the warm one. The WPC-PQPFs tend to show some gains in the correlation coefficient, relative mean error, and forecast skill relative to both the GEFSRv2 and SREF, but the gains vary with the RFC and lead time. Based on the GEFSRv2, we find that medium-range precipitation forecasts tend to have skill up to approximately day 7 relative to sampled climatology.

1. Introduction

a. Motivation and objectives

Precipitation is a key forcing of interest in many forecasting applications (Cherubini et al. 2002; Ebert and McBride 2000; Ebert et al. 2003; Fritsch et al. 1998; Hall et al. 1999; Voisin et al. 2008; Zhu and Luo 2015). Precipitation forecasts are used to issue severe weather warnings (Messner et al. 2014); forecast floods and other hydrologic variables (Kim and Barros 2001); support the operation of water supply reservoirs (Demargne et al. 2014; Pagano et al. 2001); inform decision-making in the transportation (Antolik 2000; Cools et al. 2010; Hwang et al. 2015; Vislocky and Fritsch 1995), industrial (Kolb and Rapp 1962), and agricultural sectors (Jones et al. 2000); and manage ecosystems (Sene 2016); among other applications. In all of these applications, it is critical to understand and characterize the quality of the precipitation forecasts. For example, the accuracy of both severe weather warnings and flood forecasts depends strongly on the accuracy of the precipitation forecasts (Brown et al. 2012; Demargne et al. 2010; Messner et al. 2014). In the case of flood forecasts, the accuracy of precipitation forecasts can significantly contribute to preventing flood-related damages to human life, infrastructure, property, and agriculture (Knebl et al. 2005; Montz and Gruntfest 2002).

Despite recent advances in weather forecasting from operational numerical weather prediction (NWP) models, accurate prediction of precipitation remains a critical issue and challenge (Cuo et al. 2011; Ralph et al. 2010; Röpnack et al. 2013). Uncertainty in precipitation forecasts may be due to shortcomings in the initial conditions and model physics, as well as the chaotic nature of the atmosphere (Berner et al. 2015; Grimit and Mass 2002). Precipitation

forecast uncertainty tends to increase with the magnitude of the expected precipitation amounts (Scheuerer and Hamill 2015) and is typically larger for convective than synoptic-scale events (Röpnack et al. 2013). In flood forecasting, precipitation uncertainty is often the largest contributor to the overall deficiency of the streamflow forecasts (Yu et al. 2016).

To understand and quantify the uncertainty of precipitation forecasts, ensemble techniques are increasingly being employed (Charron et al. 2010; Schaake et al. 2007; Shrestha et al. 2015; Yu et al. 2016). As ensemble forecasting systems evolve, the need arises to monitor and verify the quality of the evolving forecasting systems (Brown and Seo 2010). Ensemble verification not only provides information needed to understand forecasting errors and biases but it can assist with making decisions about future enhancements to the forecasting systems (Davis et al. 2006; Ebert et al. 2013; Murphy and Winkler 1987). Indeed, forecast verification is a fundamental aspect of forecasting (Casati et al. 2008; Cherubini et al. 2002; Davis et al. 2006; Rossa et al. 2008; Welles et al. 2007; Wernli et al. 2008). It is required to assess and compare the performance of different forecasting systems (Mason and Weigel 2009; Murphy and Winkler 1987), and to provide meaningful information to administrators, scientists, and forecast users (Murphy 1993).

Recently, various weather verification strategies have been developed to better incorporate datasets (e.g., high-resolution NWP outputs, spatial or gridded observations, etc.) and to account for the spatial distribution and scale dependency of weather variables (Casati et al. 2004; Davis et al. 2006; Ebert 2008; Roberts 2008). Here, to verify precipitation ensembles in the eastern U.S., we employ the Ensemble Verification System (EVS) (Brown et al. 2010), following the implementation strategy of Siddique et al. (2015). Among other key conditions, Siddique et al. (2015) account for spatial scale by verifying areal-averaged precipitation across different basin sizes. This areal-averaged approach is meaningful in this case because, partly, the motivation for performing this verification is to inform future hydrologic forecasting, research strategies. The areal-averaged approach can be viewed as representative of the aggregative hydrologic response of a river basin to the precipitation forcing.

We verify the ensemble precipitation forecasts within the geographic domain of each of the four eastern River Forecast Centers (RFCs) in the U.S. The four eastern RFCs are the Middle Atlantic River Forecast Center (MARFC), Northeast River Forecast Center (NERFC), Ohio River Forecast Center (OHRFC), and Southeast River Forecast Center (SERFC). We selected these RFCs because i) they collectively represent one of the most active geographic regions in the U.S. for extreme precipitation events (Hitchens et al. 2013; Moore et al. 2015); ii) they contain several major U.S. cities that can be particularly vulnerable to the impacts associated with damaging weather events and severe flooding; iii) they generally contain good quality of precipitation observations due to relatively dense networks of point observations and good radar coverage in most areas; and iv) there is a general interest in understanding the quality of different forecasting systems to support on-going forecasting operational efforts.

For the verification of the precipitation ensembles, we use precipitation outputs from the National Oceanic and Atmospheric Administration's (NOAA's) National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2) (Hamill et al. 2013) and the 21-member Short Range Ensemble Forecast (SREF) system (Du et al. 2009), as well as the NCEP's Weather Prediction Center probabilistic quantitative precipitation forecasts (WPC-PQPFs) (WPC 2016). We select these three forecasting systems or guidance for various reasons. They are either operational or similar to operational systems available and familiar to forecasters. They encompass various relevant and

interesting forecasting conditions, including different model resolutions and number of ensemble members, human-generated forecasts and, in the case of the GEFSRv2, a statistically consistent long-term dataset.

Our primary objective here is to verify and compare the ensemble precipitation forecasts from the GEFSRv2, SREF, and WPC-PQPF in the geographic domains of the RFCs in the eastern U.S. We verify the forecasts using multi-sensor precipitation estimates (MPEs) as the observed forcing (Breidenbach and Bradberry 2001; Fulton et al. 1998). We use a variety of deterministic and probabilistic metrics for the verification, conditioned upon the forecast lead time, seasonality, precipitation threshold, and spatial aggregation scale. With this study, we want to gain insight into the following questions: How does the performance of the different forecasting systems or guidance compare against each other? How does the quality of the forecasting systems vary within and between the RFCs? Does the spatial aggregation scale affect the quality of the precipitation forecasts? Are these RFCs likely to benefit from statistical postprocessing techniques?

b. Background on relevant verification studies

Several verification studies have been conducted to assess the quality of precipitation forecasts from the forecasting systems or guidance selected here (Baxter et al. 2014; Brown et al. 2012; Hamill et al. 2013; Novak et al. 2014; Siddique et al. 2015; Stensrud and Yussouf 2007; Sukovich et al. 2014). Hamill et al. (2013) verified the calibrated ensemble precipitation forecasts from the GEFSRv2 over the Continental U.S. (CONUS). They found that the GEFSRv2 is more skillful than its predecessor. Also, using the GEFSRv2, Baxter et al. (2014) performed a detailed verification of precipitation forecasts over the southeastern U.S. They found that the precipitation forecasts have some skill up to a lead time of 5.5 days. Both Brown et al. (2012) and Stensrud and Yussof (2007) analyzed precipitation forecasts from the SREF. Brown et al. (2012) found that the skill and reliability of precipitation forecasts from the SREF vary with the U.S. geographic region, lead time, precipitation threshold, and season.

Siddique et al. (2015) compared precipitation forecasts from the GEFSRv2 and SREF against MPEs in the geographic domain of the MARFC. They found that generally the two systems show similar skill and reliability over the MARFC but some differences in performance were also noted. The analysis of WPC-PQPF guidance has been limited. Indeed, recent analysis has been focused on the deterministic WPC quantitative precipitation forecasts (WPC-QPFs) (Novak et al. 2014; Sukovich et al. 2014). These studies of WPC-QPFs highlight the ability of human-generated forecasts to improve upon the accuracy of NWPs and of forecasters to learn from improved and evolving forecasting systems. Additionally, Sukovich et al. (2014) demonstrated how the accuracy of extreme WPC-QPFs vary with the U.S. geographic region and seasonality.

2. Study area

We perform the verification analysis separately in each of the four RFCs considered. Figure 1 illustrates the RFCs. We provide next a brief description of the geographic domain encompassed by each RFC.

a. MARFC

The spatial extent of the MARFC, hereafter referred to as the middle Atlantic region (MAR), includes New Jersey, Maryland, Delaware, District of Columbia, as well as parts of New

York, Pennsylvania, Virginia, and West Virginia (Fig. 1). The MAR contains a massive and complex network of build infrastructure, which makes severe weather and flooding hazards particularly relevant. It is home to several major U.S. cities, including Philadelphia and Washington D.C., and of many defining cultural and historical landmarks. The population in the MAR is approximately 41 million or 10% of the total U.S. population (United States Census Bureau 2015) but it only accounts for ~4% of total land mass.

The physical geography of the MAR is characterized by a relatively flat coastal plain on the eastern edge, followed towards the west by the Piedmont and Ridge and Valley zones, and ending with the Appalachian Plateau on the western edge (Polsky et al. 2000). This makes the MAR physically and ecologically diverse (Jones et al. 1997). Land cover and land use vary within the MAR among forested, agricultural, and urbanized landscapes (Herlihy et al. 1998), with forested areas being the most predominant (Jones et al. 1997). Additionally, the MAR comprises several major U.S. river basins including the Delaware, Susquehanna, Potomac, and James River (Siddique et al. 2015). The climate in the MAR is relatively humid, with a mean annual temperature of ~11^{o}C over the period of 1895-1997 (Polsky et al. 2000). Precipitation is relatively uniform throughout the year, with the total mean annual precipitation being ~1009 mm (Neff et al. 2000). Located among the NERFC, OHRFC, and SERFC, the MAR shares many of the hydrometeorological complexities of these other RFCs.

b. NERFC

The NERFC geographic domain, hereafter referred to as the northeast region (NER), is comprised by the states of Maine, New Hampshire, Vermont, Massachusetts, Connecticut, Rhode Island, and the majority of New York (Fig. 1). It occupies only ~3% of the U.S. landmass, however, the region is densely populated and accounts for approximately 32 million people, or 9% of the total U.S. population (United States Census Bureau 2015). Precipitation forecasting is challenging over the NER in part because of the combination of physical features that contribute to landscape and boundary complexity such as the Great Lakes, the Appalachian Mountains, and the irregular coastlines (Colle et al. 2003).

Extreme precipitation events, heat waves, and coastal flooding often affect the region (Melillo et al. 2014). During the winter, a positive North Atlantic Oscillation phase can generally result in increased precipitation amounts and occurrence of snow (Durkee et al. 2007). Over the period of 1948-2007, the mean annual precipitation was recorded to be ~1040 mm (Spierre and Wake 2010). Additionally, there is a significant increasing trend in the frequency and intensity of extreme precipitation events over the NER (Kunkel et al. 2013). Since 1970, temperature has increased by almost 0.25^{o}C per decade (Hayhoe et al. 2006), consequently decreasing the snow to precipitation ratio at selected stations (Huntington et al. 2004). Generally, the climate in the NER can be classified as warm, humid summers and snowy, cold winters with frozen precipitation.

c. OHRFC

The OHRFC domain, hereafter referred to as the Ohio region (OHR), consists of the Ohio River basin above Smithland Lock and Dam, the Cumberland River Basin, and tributaries to Lake Erie that terminate in Ohio and Pennsylvania. This region includes the entire state of Ohio, and the majority of the states of Indiana, Kentucky, and West Virginia, along with parts of several surrounding states (Illinois, Maryland, New York, North Carolina, Pennsylvania, and Tennessee) (Fig. 1). The OHR encompasses an area of ~5% of the U.S. land mass and includes

~9% (30 million people) of the total U.S. population (United States Census Bureau 2015). The Ohio River basin is the third largest by discharge (with a mean discharge of ~8,733 m^3/sec) in the U.S. (White et al. 2005) as well as the largest tributary by volume to the Mississippi River. The major tributaries of the Ohio River include the Kentucky, Cumberland, Wabash, and Kanawha Rivers. The eastern portion of the OHR is located in the Blue Ridge, Valley and Ridge, and Appalachian plateaus provinces of the Appalachian highlands, and is dominated by forest cover. Agricultural lands and some urban centers dominate the land cover in the western half of the OHR, except for some prairies in the north and west.

Riverine floods are common occurrences on the OHR (White et al. 2005). Flooding during the spring season is often associated with atmospheric circulation anomalies from the tropical Atlantic Ocean and the Gulf of Mexico that can result in heavy precipitation (Nakamura et al. 2013). Additionally, during La Niña winters, heavy precipitation events are accompanied by extreme high temperature events (Gershunov and Barnett 1998). The climate in the OHR can be classified as humid and temperate, with cool moist winters and warm summers (Voisin et al. 2011). Precipitation in the OHR is well distributed throughout the year, with the mean annual amount varying from ~1000 to 1200 mm per year (O'Donnell et al. 2000). The geographic patterns of glaciations, mountains, and the interaction of atmospheric systems with a complex geography, make the hydrometeorological behavior of the OHR diverse.

d. SERFC

The geographic domain of the SERFC, hereafter referred to as the southeast region (SER), includes all of the state of Florida, the majority of the states of Alabama, Georgia, South Carolina, North Carolina, and small portions of southern Virginia and western Mississippi (Fig. 1). The SER accounts for ~8% of the U.S. land mass and is home to 15% of the total U.S. population (47 million people) (United States Census Bureau 2015). Agriculture is a major sector of this region's economy and improved precipitation forecasts are believed to contribute to the region's economic well-being (Adams et al. 1995). Precipitation forecasts are particularly challenging in the SER because of the interaction among an extensive coastal line, tropical storms, sea breezes, and topography (Baxter et al. 2014). The location and unique physical geography of the SER means that precipitation events are related to a variety of sources, such as tropical cyclones, extra tropical baroclinic waves, mesoscale convective systems, or localized diurnal convection (Moore et al. 2015).

The climate in the SER is relatively humid (Yilmaz et al. 2005). The average monthly maximum temperatures over the region is about ~25.2°C, with a mean annual precipitation of ~1350 mm (Nam and Baigorria 2015). The SER shows a diurnal cycle of precipitation, with strong ocean and land linkage, characterized by greater afternoon precipitation on land and morning precipitation over the ocean (Prat and Nelson 2014). The region is often affected by El Niño Southern Oscillation activity (Ropelewski and Halpert 1987). During winter, El Niño years tend to be wet, whereas La Niña years are dry (Hansen et al. 1998). In spring, El Niño events tend to show higher precipitation amounts throughout the region while La Niña tends to show above average temperature in Georgia, northern Florida, and South Carolina (Jones et al. 2000).

3. Datasets

a. GEFSRv2

The GEFSRv2 datasets are based on the same atmospheric model and initial conditions as the 2012 NOAA GEFS, version 9.0.1 (Hamill et al. 2013). The reforecast model was run at

T254L42 (~0.50^{o} Gaussian grid spacing) and T190L42 (~0.67^{o} Gaussian grid spacing) resolutions for the first and second 8 days, respectively. The 11-member reforecasts are initiated only once daily at 00 UTC. The GEFSRv2 forecast cycle consists of 3 hourly accumulations for the first 72 hours (days 1-3) and 6 hourly accumulations for days 4-16. Table 1 summarizes the main characteristics of the GEFSRv2. In this study, we use 10 years of GEFSRv2 data, from 2004 to 2013. This period was mainly selected to match the available period for higher quality MPEs.

b. SREF

The NCEP's SREF system is a multi-analysis, multi-model, and multi-physics regional ensemble prediction system, currently initiated 4 times a day at 0300, 0900, 1500, and 2100 Coordinated Universal Time (UTC). Each forecast cycle comprises lead times of up to 87 hours and the forecast for each lead time is valid for 3 hourly precipitation accumulations. The SREF system was operationally implemented in ~2001 and initially consisted of a 10-member ensemble, 5 members from each the Eta and Regional Spectral Model, with a 48-km horizontal resolution (Du and Tracton 2001). Subsequent updates increased the 10-member system to 26 members, and the horizontal resolution changed from 48 to 16 km (Du et al. 2015). Here, we use 2 years of operational 21-member SREF forecasts, from January 2012 to November 2013. The SREF runs that we use include two different core models with horizontal grid spacing of 16 and 32 km, respectively. Table 1 summarizes the SREF datasets used in this study.

c. WPC-PQPFs

The WPC-PQPFs are derived, for lead times of 1 to 3 days and at 4 x 4 km^2 horizontal resolution, by incorporating forecast uncertainty information into 6-hour deterministic WPC-QPFs (WPC 2016). Specifically, a 62-member ensemble is obtained by grouping members from various forecasting systems, including the operational GEFS and SREF, NCEP's Global Forecasting System, and the European Center for Medium-Range Weather Forecasts. These ensembles are then used to estimate the variance of a binormal probability distribution function (pdf), whose mode is given by the value of the WPC-QPF. The binormal pdf is then sampled to produce the WPC-PQPFs.

The WPC probabilistic forecasts are provided in two different formats (WPC 2016): i) probabilities of exceeding a threshold, and ii) percentile accumulations, where lower percentile values are associated with smaller accumulations than are higher percentile values. Here, we use the percentile accumulation format for the 6-hour WPC-PQPFs, for lead times of 1 to 3 days released twice per day at 00 and 12 UTC. The percentile accumulations represent the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentile of the fitted pdf. We treat these 7 percentile accumulations as different precipitation ensemble members. Tables 1 summarizes key information about the WPC-PQPFs. We use WPC-PQPFs for the years 2012 and 2013.

d. MPEs

We use MPEs as the observed forcing when verifying the ensemble precipitation forecasts. For the MPEs, we use datasets provided by each of the RFCs considered in this study. These datasets are similar to the NCEP stage-IV MPEs (Moore et al. 2015; Prat and Nelson 2015). As with the NCEP stage-IV dataset, the MPEs provided by the RFCs represent a continuous time series of hourly, high-resolution gridded precipitation observations at 4 x 4 km^2 cells, over each of the four eastern RFCs. We aggregate the MPEs to the temporal (6 hourly) and

spatial scale necessary for the verification analysis. We use here MPEs over the period of 2004-2013 (Table 1).

4. Verification strategy

We use for the verification analysis different metrics, including both deterministic and probabilistic measures. Specifically, we consider the following 6 verification metrics: correlation coefficient, relative mean error (RME), Brier skill score (BSS), continuous ranked probability skill score (CRPSS), reliability diagram, and relative operating characteristic (ROC) curve. The mathematical definition of each of these metrics is provided in the Appendix. Additional details about the verification metrics can be found elsewhere (e.g., Wilks 2011; Jolliffe and Stephenson 2012). We use the EVS for the verification analysis (Brown et al. 2010).

When verifying the forecasts, we condition the forecasts and observed datasets upon different variables (e.g., forecast lead time, seasonality, precipitation threshold, and spatial aggregation scale) to account for various relevant scenarios. We use 6 hourly precipitation accumulations and focus our verification on moderate to heavy precipitation amounts. For this, we select precipitation amounts greater than that implied by a non-exceedance probability, in the sampled climatological probability distribution, of 0.9. To account for the effect of the spatial aggregation scale, we verify areal-averaged precipitation, as opposed to individual grid cells, across different basin sizes. We select three different basin sizes: small (100 x 100 km^2), intermediate (300 x 300 km^2), and large (500 x 500 km^2). For a particular basin size, we compute the different verification metrics by aggregating the verification results from three or more basins of the same size. The latter is done to account for sampling variability.

To perform the verification analysis, we work with two main case studies. In the first case study, we verify 6 hourly precipitation accumulations from the GEFSRv2, SREF, and WPC-PQPFs, over their common period of two years (2012-2013), for forecast lead times of 1 to 3 days. In the second case study, we verify 6 hourly precipitation accumulations from the GEFSRv2 alone, over the period of 2004-2013, for forecast lead times of 1 to 16 days, with the exception of the SER that is verified from 2006 to 2013. The verification, in both case studies, is done separately for each of the four eastern regions using the three different basin sizes considered (100 x 100, 300 x 300, and 500 x 500 km^2).

5. Verification of short-range GEFSRv2, SREF, and WPC-PQPF forecasts (days 1-3)

a. Correlation coefficient and RME

We use the correlation coefficient and RME as the deterministic metrics of forecast quality. We show in Fig. 2, for the three different basin sizes considered, the correlation coefficient as a function of the forecast lead time (days 1-3) for the GEFSRv2 (Figs. 2a-c), SREF (Figs. 2d-f), and WPC-PQPF (Figs. 2g-i). The overall trend in Fig. 2 is for the correlation coefficient to decline as the forecast lead time increases, meaning that the forecasts become more dissimilar to the observed values with larger forecast lead times, and to rise as the basin size increases. For instance, regarding the latter, the values of the correlation coefficient for the GEFSRv2 tend to be larger in Fig. 2c (large basin size), across regions and forecast lead times, than in Fig. 2a (small basin size). This behavior is similar for the SREF and WPC-PQPF in Fig. 2.

Relative to the other forecasting systems, the GEFSRv2 shows the most variability in the values of the correlation coefficient and the values do not indicate that one particular region performs worse or better than the other (Figs. 2a-c). The variability in the GEFSRv2 curves tend

to follow a diurnal cycle of higher predictability in the late morning and early afternoon hours than in the late night and early morning hours. A similar diurnal cycle to that identified here has been reported by others for the GEFSRv2 precipitation forecasts (Siddique et al. 2015), as well as cloud and visibility forecasts (Verlinden and Bright 2016). We investigate the diurnal cycle further in the next subsections using the other verification metrics. For the SREF (Figs. 2d-f), the curves associated with each region are, for the most part, close to each other, thus indicating that the performance of the SREF may be somewhat similar across the different eastern regions. The WPC-PQPFs (Figs. 2g-i) are also characterized by curves that are similar to each other but the curve for the MAR seems to be consistently higher than the other ones, potentially suggesting improved quality in the MAR for the WPC-PQPF. Since the WPC-PQPFs are issued from the MAR, it is possible that the forecasters' familiarity with the MAR may play a role in the performance of the WPC-PQPF in this region.

To examine the bias associated with the mean ensemble forecast, we plot in Fig. 3 the RME versus the forecast lead time for the precipitation forecasts from the GEFSRv2 (Figs. 3a-c), SREF (Figs. 3d-f), and WPC-PQPF (Figs. 3g-i). Generally, the trend in Fig. 3 is for the three forecasting systems or guidance to underforecast moderate to heavy precipitation, i.e., the tendency is for a negative bias across the forecast lead times and basin sizes. The bias increases, in most cases, with the forecast lead time and decreases some as the basin size increases. Comparing the three forecasting systems or guidance against each other, the WPC-PQPF seems to have the least bias of the three. For example, the bias for the NER at a lead time of 12 hours and the largest basin size considered (500 x 500 km^2) is ~0 for the WPC-PQPF (Fig. 3i) and ~-0.1 for the SREF (Fig. 3f).

As was the case in Fig. 2, the curves for the GEFSRv2 show again the most variability across forecast lead times and are marked by a diurnal cycle of oscillating RME values. Generally, there also seems to be a tendency in Fig. 3 for the RME to show consistently less bias (closer to zero) in the NER and more bias (farther from zero) in the SER than in the other regions. The latter is particularly noticeable for the SREF (Figs. 3d-f) and WPC-PQPF (Figs. 3g-i). For instance, in Figs. 3g-i, the curves for the SER are always below all the other curves, indicating that the SER has a stronger underforecasting bias than the other regions. One reason for this underforecasting bias may be due to the greater uncertainty in convective precipitation, which is more common in the SER, compared with the other eastern regions.

b. Skill

To investigate the probabilistic attributes of the selected forecasting systems and guidance, we examine the BSS and CRPSS associated with the precipitation forecasts and observations pairs. When computing the BSS and CRPSS, we use sampled climatology as the reference system. In Fig. 4, we show the BSS as a function of the forecast lead time for the cool (October-March) and warm (April-September) season. The BSS in Fig. 4 is computed using 6 hourly accumulations, a 500 x 500 km^2 basin size, and both light to moderate precipitation events (Pr=0.5) as well as moderate to heavy precipitation events (Pr=0.9). Overall, the results in Fig. 4 indicate that the forecast skill of the GEFSRv2 (Figs. 4a-d), SREF (Figs. 4e-h) and WPC-PQPF (Figs. 4i-j) declines with increasing forecast lead time, and it is generally higher in the cool season than in the warm one across all the regions. Additionally, the WPC-PQPF is generally more skillful than the GEFSRv2 and SREF, independently of the forecast lead time, region, and precipitation threshold. For example, the WPC-PQPF tends to remain skillful across lead times and regions while the GEFSRv2 (e.g., Fig. 4a) does not. Also, in some cases, the skill

of the SREF declines quickly for moderate to heavy precipitation events (e.g., the MAR and NER in Fig. 4f) and it can have a relatively wider spread in skill among the regions than the GEFSRv2 and WPC-PQPF, especially for the cool season (e.g., Fig. 4f).

Variations in the BSS among the regions are also evident in Fig. 4. During the cool season, the skill from all three forecasting systems or guidance, for light to moderate precipitation, is relatively better within the OHR than in the other regions (Figs. 4a, 4e, and 4i). This is particularly noticeable for the GEFSRv2 (Fig. 4a). The MAR shows the least skill among all the regions for the SREF (Figs. 4e-h) but, in contrast, a comparable skill for the WPC-PQPF guidance (Figs. 4i-l). The GEFSRv2 forecasts are characterized by a strong diurnal cycle, with the cycle being somewhat stronger in the SER (e.g., Fig. 4c) than in the other regions. During the warm season and for moderate to heavy precipitation events, the NER seems to have slightly greater skill than the other regions (e.g., Figs. 4d, 4h, and 4l), which may be due to the influence of the jet stream and extratropical cyclones on precipitation in this most poleward of the study domains . We note that the lack of a diurnal cycle in the precipitation forecasts from the SREF and WPC-PQPF may be due to the fact that these systems issue or release forecasts at least twice a day.

In Fig. 5, we plot the CRPSS (relative to sampled climatology) against the forecast lead time for precipitation forecasts from the GEFSRv2 (Figs. 5a-c), SREF (Figs. 5d-f), and WPC-PQPF (Figs. 5g-i). The CRPSS is computed using 6 hourly accumulations, moderate to heavy precipitation events, and different basin sizes. In general, the CRPSS decreases with increasing forecast lead time but increases with increasing basin size, independently of the region. There is also a slight tendency for the CRPSS to increase from the GEFSRv2 to the SREF and from the SREF to the WPC-PQPF. The regional variations in the CRPSS are largest in the GEFSRv2 and least in the WPC-PQPF, where the forecasts seem to exhibit similar skill independently of the region, particularly for the large basin size (e.g., Fig. 5i). As was the case in Fig. 4, the GEFSRv2 shows in Fig. 5 a strong diurnal cycle, potentially signaling the lesser ability of the GEFSRv2 to capture and resolve convective events.

Contrasting the regions against each other in Fig. 5, we find that the MAR shows again the least skill at the initial lead times with the SREF, but a comparable skill with the GEFSRv2 and WPC-PQPF. The NER seems to consistently have a slightly larger skill than the other regions with the SREF and WPC-PQF. While the SER and OHR behave similarly in regards to their skill (e.g., Figs. 5b and 5d).

c. Reliability

We examine the reliability of the GEFSRv2, SREF, and WPC-PQPF across the four eastern regions in Fig. 6. To compute the reliability curves, we focus on moderate to heavy precipitation events and a large basin size (500 x 500 km^2). For the GEFSRv2, the forecasts tend to be underconfident at low forecast probabilities and overconfident at high forecast probabilities at the day 1 forecast lead time (Fig. 6a). At the day 2 (Fig. 6b) and day 3 (Fig. 6c) forecast lead time, the GEFSRv2 forecasts for the low forecast probabilities become less underconfident. These trends, regarding the reliability of the GEFSRv2, are similar across all the regions. The SREF is consistently overconfident across forecast lead times and regions (Figs. 6d-f), with the MAR and NER being slightly more unreliable than the OHR and SER (e.g., Fig. 6d). The WPC-PQPFs are underconfident at high forecast probabilities. Nevertheless, the WPC-PQPFs seem relatively more reliable than the GEFSRv2 and SREF. Additionally, the SER seems to be somewhat more reliable than the other regions. Overall, the three forecasting systems or

guidance exhibit conditional biases across regions, thus indicating that postprocessing may be beneficial to all the regions.

d. ROC curves

To examine the ability of the different forecasting systems and guidance to discriminate between occurrences versus non-occurrences of a precipitation event, we plot in Fig. 7 the ROC curve for each region. We note that the ROC curve actually plots the probability of detection (PoD) of an event (or true alarm) versus the probability of false detection (PoFD) (or false alarm) using a set of different probability thresholds. Additionally, a larger area under the ROC curve and above the 45° line from the origin (i.e., the ROC area) represents a more skillful forecast, with more ability to discriminate between precipitation events. We use 6 hourly accumulations, a 500 x 500 km^2 basin size, and moderate to heavy precipitation events to determine the ROC curves.

We show the ROC curves for the MAR, NER, OHR, and SER in Figs. 7a-d, respectively. Overall, the GEFSRv2 and WPC-PQPF show better discrimination ability than the SREF across regions, although these differences can be very small in some regions, e.g., SER (Fig. 7d). The MAR shows a poor ability to discriminate different events with the SREF but a comparable ability with the GEFSRv2 and WPC-PQPF (Fig. 7a). Overall, in Fig. 7, the WPC-PQPF consistently shows better discrimination across regions; however, the GEFSRv2 exhibits somewhat better discrimination than the WPC-PQPF for moderate to heavy precipitation events across the MAR

6. Verification of short- to medium-range GEFSRv2 forecasts (days 1-16)

a. Correlation and RME

We now focus our attention on short- to medium-range precipitation forecasts from the GEFSRv2, where we consider forecasts for the period of 2004-2013. In Figs. 8a-c, we show the correlation coefficient between the GEFSRv2 mean ensemble forecast and the corresponding observed precipitation values as a function of the forecast lead time for small, intermediate, and large basin sizes, respectively. The correlation coefficient declines with increasing forecast lead time and increases slightly with the basin size. This behavior is similar across regions. Fig. 8 also suggests that, at forecast lead times beyond 8 days, there is little to no predictability in the GEFSRv2 across regions.

Figs. 8d-f plot the RME of the GEFSRv2 mean ensemble forecast against the forecast lead time for small, intermediate, and large basin sizes, respectively. For all the regions, the RME shows a strong negative bias that increases with the forecast lead time, although it seems to stabilize at ~12 days, and decreases slightly with increasing basin size. Additionally, the RME does not vary greatly among regions. The two most salient differences are the larger unconditional bias for SER at forecast lead times of less than 3-4 days and the stronger daily oscillations for SER. Regarding the latter, with convective events likely being more dominant in SER than in the other regions, it is perhaps not surprising that GEFSRv2 forecasts show a stronger daily variation in this region.

Contrasting the results in Figs. 8a-c against those in Figs. 2a-c, we find that the correlation coefficients based on the 2012-2013 GEFSRv2 dataset (Figs. 2a-c) are similar to those from the longer dataset for the period 2004-2013 (Figs. 8a-c). We reach a similar conclusion by contrasting the RME values in Figs. 8d-f against those in Figs. 3a-c.

b. Skill

We show in Figs. 9a-c the BSS as a function of the calendar month for forecast lead times of 1, 3, and 5 days, respectively. The BSS in Fig. 9 are computed using 6 hourly accumulations, 500 x 500 km^2 basin sizes, and moderate to heavy precipitation events. In Fig. 9, the GEFSRv2 shows overall less skill in the summer months than in the winter months across regions and forecast lead times. The month of July seems to generally have the lowest skill. However, the NER can have its lowest skill in May (Fig. 9a) and OHR in August (Fig. 9b). Additionally, the skill tends to decrease with the forecast lead time, as expected, so that BSS values in Fig. 9c (day 5) tend to be lower than in Fig. 9a (day 1) across months and regions.

In terms of the CRPSS, the skill decreases with increasing forecast lead time, as expected, and increases somewhat with increasing basin size across regions (Fig. 10). In Fig. 10, we compute the CRPSS using 6 hourly accumulations and moderate to heavy precipitation events. Indeed, the skill as a function of the forecast lead time tends to be similar across the different eastern regions (Fig. 10). The MAR seems, however, to consistently have slightly better skill up to day 7 than the SER across basin sizes (Figs. 10a-c). These results, which span the period of 2004-2013, are consistent with our previous findings for years 2012-2013 (Figs. 5a-c). Overall, Fig. 10 shows that GEFSRv2 tends to remain skillful across the eastern regions up to a lead time of 7 days, after which the skill becomes similar to sampled climatology.

c. Reliability

The reliability diagrams in Fig. 11 show that the GEFSRv2 is slightly underconfident at low forecast probabilities and strongly overconfident at large probabilities for day 1 forecast lead time across basin sizes (Figs. 11a, 11d, and 11g). At longer forecast lead times (day 3 in Figs. 11b, 11e, and 11h, and day 6 in Figs. 11c, 11f, and 11i), the GEFSRv2 mainly overpredicts the forecast probabilities across basins sizes. This makes the GEFSRv2 overconfident across the eastern regions at longer forecast lead times. Indeed, the trends in the reliability diagrams in Fig. 11 are, for the most part, similar across regions. Nevertheless, in some cases, the MAR and NER show somewhat less reliability (Fig. 11h) and the SER more reliability (Fig. 11g) than the other regions. In terms of the forecast sharpness, assessed by inspecting the insets in Fig. 11, the SER is relatively less sharp in some of the cases in Fig. 11 (e.g., Figs. 11f and 11i). Fig. 11 further confirms and supports our previous results based on Figs. 6a-c, which underscore the potential for statistical postprocessing to improve the raw ensemble forecasts from the GEFSRv2.

7. Summary and conclusions

In this study, we verified the quality of ensemble precipitation forecasts from the GEFSRv2, SREF, and WPC-PQPF. We selected these three forecasting systems or guidance because they are operational, have multiyear data available, and/or represent conditions of interest to forecasters. The verification was conducted for 6 hourly accumulations and mostly for moderate to heavy precipitation events across four eastern regions. The regions represent the geographic domains of the eastern U.S. RFCs.

Based on the three forecasting systems or guidance analyzed, the verification results indicate that, across the eastern U.S., precipitation forecast bias decreases and the skill and reliability improve as the basin size increases. However, all the forecasts exhibit a strong underforecasting bias. Additionally, the skill of the forecasts is appreciably better in the cool season than in the warm one. Overall, the WPC-PQPFs tend to show some gains in the correlation coefficient, relative mean error, and forecast skill relative to both the GEFSRv2 and

SREF, but the gains vary with the region and forecast lead time. For the regions considered, according to the short- to medium-range GEFSRv2 outputs, we find that: the precipitation forecasts tend to have some skill up to approximately day 7, beyond that the skill is similar to that from sampled climatology; 0- to 7-day forecast bias grows with the forecast lead time; and the analysis based on reliability diagrams indicates that forecasts tend, for the most part, to be overconfident.

Relative to the other regions considered, the MAR shows unusually low skill in the warm season with the SREF and a noticeable skill gain with the WPC-PQPF, relative to both the GEFSRv2 and SREF. The SER shows a pronounced daily cycle, more pronounced than in the other eastern regions, characterized by greater skill in the late morning than in the late evening. The SREF seems to perform better in the NER than in the other eastern regions, i.e. forecasts are the least biased and have the most skill for days 1-3.

For this study, we verified ensemble precipitation forecasts from the GEFSRv2, SREF, and WPC-PQPF since precipitation is a key forcing of interest in many weather-related applications. Our verification was based on selected metrics (see the appendix) conditioned upon the precipitation threshold, forecast lead time, seasonality, and basin size. Although our verification strategy provided useful diagnosis information regarding the quality of ensemble precipitation forecasts, it did not provide direct information on how to improve the underlying NWP models. To better understand the physical and environmental conditions associated with forecast errors and skill, we will need to consider more weather variables than just precipitation, as in the approach by Moore et al. (2015). Additionally, we could focus the verification on few high-impact events or unusual weather scenarios, as suggested by Novak et al. (2014).

Acknowledgements
The first two authors gratefully acknowledge the funding support provided by the NOAA's NWS. They also acknowledge the computational support provided by the Institute for CyberScience at the Pennsylvania State University. We are also thankful to Christopher Schaffer of the Southeast River Forecast Center for providing useful comments and suggestions regarding this research.

APPENDIX

Verification metrics

a. Correlation coefficient and relative mean error (RME)

The correlation coefficient measures the degree of linear association between the pairs of mean ensemble forecasts and corresponding observations. However, the correlation coefficient does not provide any direct information about the bias in the forecast data (Brier and Allen 1951). Hence, the RME is used to explore the relative bias of a forecast system. The RME measures the mean difference between a set of forecasts and corresponding observations as a fraction of the average observed value, and can be expressed as

$$\mathrm{RME} = \frac{\sum_{i=1}^{n}\left(\overline{X}_i - Y_i\right)}{\sum_{i=1}^{n} Y_i}, \tag{1}$$

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

b. Brier Skill Score (BSS)

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

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} \left[F_{xi}(q) - F_{yi}(q) \right]^2, \tag{2}$$

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

$$F_{X_i}(q) = \mathrm{P_r}[X_i > q], \tag{3}$$

n is again the total number of forecast-observation pairs, and

$$F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In order 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):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS_{main}}}{\mathrm{BS_{reference}}}, \tag{5}$$

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

c. Reliability diagram

As suggested by Murphy (1973), the BS can be further decomposed into a reliability, resolution, and uncertainty component. In this study, instead of using the decomposed BS to quantify the reliability and resolution of the forecasts, we use the so-called reliability diagram. The reliability diagram shows the full joint distribution of forecasts and observations to reveal the reliability of the probability forecasts. For the forecast values portioned into bin B_k and defined by the exceedance of threshold *q*, the average forecast probability can be expressed as

$$\bar{F}_{X_k}(q) = \frac{1}{|I_k|} \sum_{I_k} F_{X_i}(q), \text{ where } I_k = \{i : \mathrm{X}_i \in B_k\}, \tag{6}$$

where I_k is the collection of all indices i for which X_i falls into bin B_k, and $|I_k|$ denotes the number of elements in I_k. The corresponding fraction of observations that fall in the K^{th} bin is given by

$$\bar{F}_{Y_k}(q) = \frac{1}{|I_k|} \sum_{I_k} F_{Y_i}(q), \text{ where } F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The reliability diagram plots $\bar{F}_{X_k}(q)$ against $\bar{F}_{Y_k}(q)$.

d. Mean Continuous Ranked Probability Skill Score (CRPSS)

The 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

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left[F_x(q) - F_y(q) \right]^2 dq. \tag{8}$$

To evaluate the skill of the main forecasting system relative to the reference forecast system, the associated skill score, the Mean Continuous Ranked Probability Skill Score (CRPSS), is defined as:

$$\mathrm{CRPSS} = 1 - \frac{\overline{\mathrm{CRPS}}_{\mathrm{main}}}{\overline{\mathrm{CRPS}}_{\mathrm{reference}}}, \tag{9}$$

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

e. Relative operating characteristic (ROC) curve

The ROC curve is a measure of the quality of probability forecasts that relates the probability of detection (PoD) or true alarm to the corresponding probability of false detection (PoFD) or false-alarm rate, as a decision threshold is varied across the full range of a continuous forecast quantity. For a particular threshold, the PoD is given by

$$\mathrm{PoD} = \frac{\sum_{i=1}^{n} I_{X_i}(F_{X_i}(q) > d \,|\, Y_i > q)}{\sum_{i=1}^{n} I_{Y_i}(Y_i > q)}, \tag{10}$$

where I denotes the indicator function and d denotes the probability threshold at which the event triggers some action. Similarly, the PoFD can be expressed as

$$\mathrm{PoFD} = \frac{\sum_{i=1}^{n} I_{X_i}(F_{X_i}(q) > d \,|\, Y_i \le q)}{\sum_{i=1}^{n} I_{Y_i}(Y_i \le q)}. \tag{11}$$

The relationship between PoD and PoFD is assumed bivariate normal such that

$$\mathrm{PoD} = \phi[a + b\phi^{-1}(\mathrm{PoFD})], \tag{12}$$

where $a = \frac{\mu_{\mathrm{PoD}} - \mu_{\mathrm{PoFD}}}{\sigma_{\mathrm{PoD}}}$, $b = \frac{\sigma_{\mathrm{PoFD}}}{\sigma_{\mathrm{PoD}}}$, and ϕ is the cdf of the standard normal distribution. μ_{PoD} and μ_{PoFD} are the means while σ_{PoD} and σ_{PoFD} denote the standard deviations of the PoD and PoFD, respectively. The ROC curve plots the PoD (fraction of true alarms) against the PoFD (fraction of false alarms) for all possible values of the decision threshold, *d* [0,1], noting that an ensemble forecast is essentially a step function, with as many possible values of *d* as the number of ensemble members.

References

Adams, R. M., K. J. Bryant, B. A. McCarl, D. M. Legler, J. O'Brien, A. Solow, and R. Weiher, 1995: Value of improved long-range weather information. *Contemporary Economic Policy,* **13,** 10-19.

Antolik, M. S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *Journal of Hydrology*, **239,** 306-337.

Baxter, M. A., G. M. Lackmann, K. M. Mahoney, T. E. Workoff, and T. M. Hamill, 2014: Verification of Quantitative Precipitation Reforecasts over the Southeastern United States. *Weather and Forecasting*, **29,** 1199-1207.

Berner, J., K. R. Fossell, S. Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the Skill of Probabilistic Forecasts: Understanding Performance Improvements from Model-Error Representations. *Monthly Weather Review*, **143,** 1295-1320.

Breidenbach, J. P., and J. S. Bradberry, 2001: Multisensor precipitation estimates produced by National Weather Service River Forecast Centers for hydrologic applications.

Brier, G. W., & Allen, R. A., 1951: Forecast verification. Compendium of Meteorology, 843-851.

Brown, J. D., and D.-J. Seo, 2010: A Nonparametric Postprocessor for Bias Correction of Hydrometeorological and Hydrologic Ensemble Forecasts. *Journal of Hydrometeorology*, **11,** 642-665.

Brown, J. D., D.-J. Seo, and J. Du, 2012: Verification of Precipitation Forecasts from NCEP's Short-Range Ensemble Forecast (SREF) System with Reference to Ensemble Streamflow Prediction Using Lumped Hydrologic Models. *Journal of Hydrometeorology*, **13,** 808-836.

Brown, J. D., J. Demargne, D.-J. Seo, and Y. Liu, 2010: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environmental Modelling & Software*, **25,** 854-872.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, **11,** 141-154.

Casati, B., and Coauthors, 2008: Forecast verification: current status and future directions. *Meteorological Applications*, **15,** 3-18.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System. *Monthly Weather Review*, **138,** 1877-1901.

Cherubini, T., A. Ghelli, and F. Lalaurette, 2002: Verification of Precipitation Forecasts over the Alpine Region Using a High-Density Observing Network. *Weather and Forecasting*, **17,** 238-249.

Colle, B. A., J. B. Olson, and J. S. Tongue, 2003: Multiseason Verification of the MM5. Part II: Evaluation of High-Resolution Precipitation Forecasts over the Northeastern United States. *Weather and Forecasting*, **18,** 458-480.

Cools, M., E. Moons, and G. Wets, 2010: Assessing the Impact of Weather on Traffic Intensity. *Weather, Climate, and Society*, **2,** 60-68.

Cuo, L., T. C. Pagano, and Q. J. Wang, 2011: A Review of Quantitative Precipitation Forecasts and Their Use in Short- to Medium-Range Streamflow Forecasting. *Journal of Hydrometeorology*, **12,** 713-728.

Davis, C., B. Brown, and R. Bullock, 2006: Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas. *Monthly Weather Review*, **134,** 1772-1784.

Demargne, J., J. Brown, Y. Liu, D.-J. Seo, L. Wu, Z. Toth, and Y. Zhu, 2010: Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmospheric Science Letters*, **11,** 114-122.

Demargne, J., and Coauthors, 2014: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bulletin of the American Meteorological Society*, **95,** 79-98.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Regional ensemble forecast systems at NCEP. *Paper 2A.5 of the 27th Conf. on WAF and 23rd Conf. on NWP, Chicago, IL, 29 June – 3 July, 2015*.

Du, J., and Coauthors, June 2009: NCEP short-range ensemble forecast (SREF) system upgrade in 2009. In Preprints, 19th Conf. on Numerical Weather Prediction and 23rd Conf. on Weather Analysis and Forecasting, Omaha, Nebraska, Amer. Meteor. Soc. Weather and Forecasting.

Du, J., and M. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: An update. Preprints. *Ninth Conf. on Mesoscale Processes*, 355-356.

Durkee, D. J., D. J. Frye, M. C. Fuhrmann, C. M. Lacke, G. H. Jeong, and L. T. Mote, 2007: Effects of the North Atlantic Oscillation on precipitation-type frequency and distribution in the eastern United States. *Theoretical and Applied Climatology*, **94,** 51-65.

Ebert, E., and Coauthors, 2013: Progress and challenges in forecast verification. *Meteorological Applications*, **20,** 130-139.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, **15,** 51-64.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology*, **239,** 179-202.

Ebert, E. E., U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGNE Assessment of Short-term Quantitative Precipitation Forecasts. *Bulletin of the American Meteorological Society*, **84,** 481-492.

Fritsch, J. M., and Coauthors, 1998: Quantitative Precipitation Forecasting: Report of the Eighth Prospectus Development Team, U.S. Weather Research Program. *Bulletin of the American Meteorological Society,* **79,** 285-299.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. O'Bannon, 1998: The WSR-88D Rainfall Algorithm. *Weather and Forecasting*, **13,** 377-395.

Gershunov, A., and T. P. Barnett, 1998: ENSO Influence on Intraseasonal Extreme Rainfall and Temperature Frequencies in the Contiguous United States: Observations and Model Results. *Journal of Climate*, **11,** 1575-1586.

Grimit, E. P., and C. F. Mass, 2002: Initial Results of a Mesoscale Short-Range Ensemble Forecasting System over the Pacific Northwest. *Weather and Forecasting*, **17,** 192-205.

Hall, T., H. E. Brooks, and C. A. Doswell, 1999: Precipitation Forecasting Using a Neural Network. *Weather and Forecasting*, **14,** 338-345.

Hamill, T. M., and Coauthors, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bulletin of the American Meteorological Society*, **94,** 1553-1565.

Hansen, J. W., A. W. Hodges, and J. W. Jones, 1998: ENSO Influences on Agriculture in the Southeastern United States*. *Journal of Climate*, **11,** 404-411.

Hayhoe, K., and Coauthors, 2006: Past and future changes in climate and hydrological indicators in the US Northeast. *Climate Dynamics*, **28,** 381-407.

Herlihy, A. T., J. L. Stoddard, and C. B. Johnson, 1998: The Relationship between Stream Chemistry and Watershed Land Cover Data in the Mid-Atlantic Region, U.S. *Biogeochemical Investigations at Watershed, Landscape, and Regional Scales: Refereed papers from BIOGEOMON, The Third International Symposium on Ecosystem Behavior; Co-Sponsored by Villanova University and the Czech Geological Survey; held at Villanova University, Villanova Pennsylvania, USA, June 21–25, 1997*, R. K. Wieder, M. Novák, and J. Černý, Eds., Springer Netherlands, 377-386.

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and Temporal Characteristics of Heavy Hourly Rainfall in the United States. *Monthly Weather Review*, **141,** 4564-4575.

Huntington, T. G., G. A. Hodgkins, B. D. Keim, and R. W. Dudley, 2004: Changes in the Proportion of Precipitation Occurring as Snow in New England (1949–2000). *Journal of Climate*, **17,** 2626-2636.

Hwang, Y., A. J. Clark, V. Lakshmanan, and S. E. Koch, 2015: Improved Nowcasts by Blending Extrapolation and Model Forecasts. *Weather and Forecasting*, **30,** 1201-1217.

Jolliffe, I. T., and D. B. Stephenson, 2012: Forecast Verification: a Practioner's Guide in Atmospheric Science. *John Wiley & Sons*.

Jones, J. W., J. W. Hansen, F. S. Royce, and C. D. Messina, 2000: Potential benefits of climate forecasting to agriculture. *Agriculture, Ecosystems & Environment*, **82,** 169-184.

Jones, K. B., and Coauthors, 1997: An ecological assessment of the United States mid-Atlantic region: a landscape atlas.

Kim, G., and A. P. Barros, 2001: Quantitative flood forecasting using multisensor data and neural networks. *Journal of Hydrology*, **246,** 45-62.

Knebl, M. R., Z. L. Yang, K. Hutchison, and D. R. Maidment, 2005: Regional scale flood modeling using NEXRAD rainfall, GIS, and HEC-HMS/RAS: a case study for the San Antonio River Basin Summer 2002 storm event. *Journal of Environmental Management*, **75,** 325-336.

Kolb, L. L., and R. R. Rapp, 1962: The Utility of Weather Forecasts to the Raisin Industry. *Journal of Applied Meteorology*, **1,** 8-12.

Kunkel, K. E., and Coauthors, 2013: Monitoring and Understanding Trends in Extreme Storms: State of Knowledge. *Bulletin of the American Meteorological Society*, **94,** 499-514.

Mason, S. J., and A. P. Weigel, 2009: A Generic Forecast Verification Framework for Administrative Purposes. *Monthly Weather Review*, **137,** 331-349.

Melillo, J. M., T. T. Richmond, and G. Yohe, 2014: Climate change impacts in the United States. *Third National Climate Assessment*.

Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2014: Heteroscedastic Extended Logistic Regression for Postprocessing of Ensemble Guidance. *Monthly Weather Review*, **142,** 448-456.

Montz, B. E., and E. Gruntfest, 2002: Flash flood mitigation: recommendations for research and applications. *Global Environmental Change Part B: Environmental Hazards*, **4,** 15-22.

Moore, B. J., K. M. Mahoney, E. M. Sukovich, R. Cifelli, and T. M. Hamill, 2015: Climatology and Environmental Characteristics of Extreme Precipitation Events in the Southeastern United States. *Monthly Weather Review*, **143,** 718-741.

Murphy, A. H., 1993: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting*, **8,** 281-293.

Murphy, A. H., and R. L. Winkler, 1987: A General Framework for Forecast Verification. *Monthly Weather Review*, **115,** 1330-1338.

Nakamura, J., U. Lall, Y. Kushnir, A. W. Robertson, and R. Seager, 2013: Dynamical Structure of Extreme Floods in the U.S. Midwest and the United Kingdom. *Journal of Hydrometeorology*, **14,** 485-504.

Nam, W.-H., and G. A. Baigorria, 2015: Analysing changes to the spatial structures of precipitation and temperature under different ENSO phases in the Southeast and Midwest United States. *Meteorological Applications*, **22,** 797-805.

Neff, R., H. Chang, C. G. Knight, R. G. Najjar, B. Yarnal, and H. A. Walker, 2000: Impact of climate variation and change on Mid-Atlantic Region hydrology and water resources. *Climate Research*, **14,** 207-218.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and Temperature Forecast Performance at the Weather Prediction Center. *Weather and Forecasting*, **29,** 489-504.

O'Donnell, G. M., K. P. Czajkowski, R. O. Dubayah, and D. P. Lettenmaier, 2000: Macroscale hydrological modeling using remotely sensed inputs: Application to the Ohio River basin. *Journal of Geophysical Research: Atmospheres*, **105,** 12499-12516.

Pagano, T. C., H. C. Hartmann, and S. Sorooshian, 2001: Using Climate forecasts for water management: Arizona and the 1997-1998 EL NIÑO1. *JAWRA Journal of the American Water Resources Association*, **37,** 1139-1153.

Polsky, C., J. Allard, N. Currit, R. Crane, and B. Yarnal, 2000: The Mid-Atlantic Region and its climate: past, present, and future. *Climate Research*, **14,** 161-173.

Prat, O. P., and B. R. Nelson, 2014: Characteristics of annual, seasonal, and diurnal precipitation in the Southeastern United States derived from long-term remotely sensed data. *Atmospheric Research*, **144,** 4-20.

Prat, O. P., and B. R. Nelson, 2015: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). *Hydrol. Earth Syst. Sci.*, **19,** 2037-2056.

Ralph, F. M., E. Sukovich, D. Reynolds, M. Dettinger, S. Weagle, W. Clark, and P. J. Neiman, 2010: Assessment of Extreme Quantitative Precipitation Forecasts and Development of Regional Extreme Event Thresholds Using Data from HMT-2006 and COOP Observers. *Journal of Hydrometeorology*, **11,** 1286-1304.

Roberts, N., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorological Applications*, **15,** 163-169.

Ropelewski, C. F., and M. S. Halpert, 1987: Global and Regional Scale Precipitation Patterns Associated with the El Niño/Southern Oscillation. *Monthly Weather Review*, **115,** 1606-1626.

Röpnack, A., A. Hense, C. Gebhardt, and D. Majewski, 2013: Bayesian Model Verification of NWP Ensemble Forecasts. *Monthly Weather Review*, **141,** 375-387.

Rossa, A., P. Nurmi, and E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. *Precipitation: Advances in Measurement, Estimation and Prediction*, S. Michaelides, Ed., Springer Berlin Heidelberg, 419-452.

Schaake, J., and Coauthors, 2007: Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrology and Earth System Sciences Discussions*, **4,** 655-717.

Scheuerer, M., and T. M. Hamill, 2015: Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions. *Monthly Weather Review*, **143,** 4578-4596.

Sene, K., 2016: River Flooding. *Hydrometeorology*, Springer International Publishing, 237-272.

Shrestha, D. L., D. E. Robertson, J. C. Bennett, and Q. J. Wang, 2015: Improving Precipitation Forecasts by Generating Ensembles through Postprocessing. *Monthly Weather Review*, **143,** 3642-3663.

Siddique, R., A. Mejia, J. Brown, S. Reed, and P. Ahnert, 2015: Verification of precipitation forecasts from two numerical weather prediction models in the Middle Atlantic Region of the USA: A precursory analysis to hydrologic forecasting. *Journal of Hydrology*, **529, Part 3,** 1390-1406.

Spierre, S. G., and C. Wake, 2010: Trends in Extreme Precipitation Events for the Northeastern United States 1948-2007.

Stensrud, D. J., and N. Yussouf, 2007: Reliable Probabilistic Quantitative Precipitation Forecasts from a Short-Range Ensemble Forecasting System. *Weather and Forecasting*, **22,** 3-17.

Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme Quantitative Precipitation Forecast Performance at the Weather Prediction Center from 2001 to 2011. *Weather and Forecasting*, **29,** 894-911.

United States Census Bureau, 2015: U.S. Department of Commerce.

Verlinden, K. L., & Bright, D. R., 2016: An Investigation of Reforecasting Applications for NGGPS Aviation Weather Prediction: An Initial Study of Cloud and Visibility Prediction.

Vislocky, R. L., and J. M. Fritsch, 1995: Generalized Additive Models versus Linear Regression in Generating Probabilistic MOS Forecasts of Aviation Weather Parameters. *Weather and Forecasting*, **10,** 669-680.

Voisin, N., A. W. Wood, and D. P. Lettenmaier, 2008: Evaluation of Precipitation Products for Global Hydrological Prediction. *Journal of Hydrometeorology*, **9,** 388-407.

Voisin, N., F. Pappenberger, D. P. Lettenmaier, R. Buizza, and J. C. Schaake, 2011: Application of a Medium-Range Global Hydrologic Probabilistic Forecast Scheme to the Ohio River Basin. *Weather and Forecasting*, **26,** 425-446.

Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic Verification: A Call for Action and Collaboration. *Bulletin of the American Meteorological Society*, **88,** 503-511.

WPC, 2016: About WPC's PQPF and Percentile QPF Products, accessed on 19 Feb. 2016.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts. *Monthly Weather Review*, **136,** 4470-4487.

White, D., K. Johnston, and M. Miller, 2005: Ohio river basin. *Rivers of North America***,** 375-424.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences.* Vol. 100, Academic press

Yilmaz, K. K., T. S. Hogue, K.-l. Hsu, S. Sorooshian, H. V. Gupta, and T. Wagener, 2005: Intercomparison of Rain Gauge, Radar, and Satellite-Based Precipitation Estimates with Emphasis on Hydrologic Forecasting. *Journal of Hydrometeorology*, **6,** 497-517.

Yu, W., E. Nakakita, S. Kim, and K. Yamaguchi, 2016: Impact Assessment of Uncertainty Propagation of Ensemble NWP Rainfall to Flood Forecasting with Catchment Scale. *Advances in Meteorology*, **2016,** 17.

Zhu, Y., and Y. Luo, 2015: Precipitation Calibration Based on the Frequency-Matching Method. *Weather and Forecasting*, **30,** 1109-1124.

Table 1. Summary and main characteristics of the datasets used in the study.

Dataset	Horizontal Resolution (km^2)	Number of models	Number of ensemble members	Lead time (hours)	Period of analysis (years)
GEFSRv2	~55 x 55 (0.5^0 x 0.5^0)	1	11	1-192	2004-2013
	~73 x 73 (0.67^0 x 0.67^0)	1	11	193-384	2004-2013
SREF	~32 x 32	3/4	21	1-87	2012
	~16 x 16	3/4	21	1-87	2012-2013
WPC-PQPF	~4 x 4	-	7	1-72	2012-2013
MPEs	~4 x 4	-	-	-	2004-2013

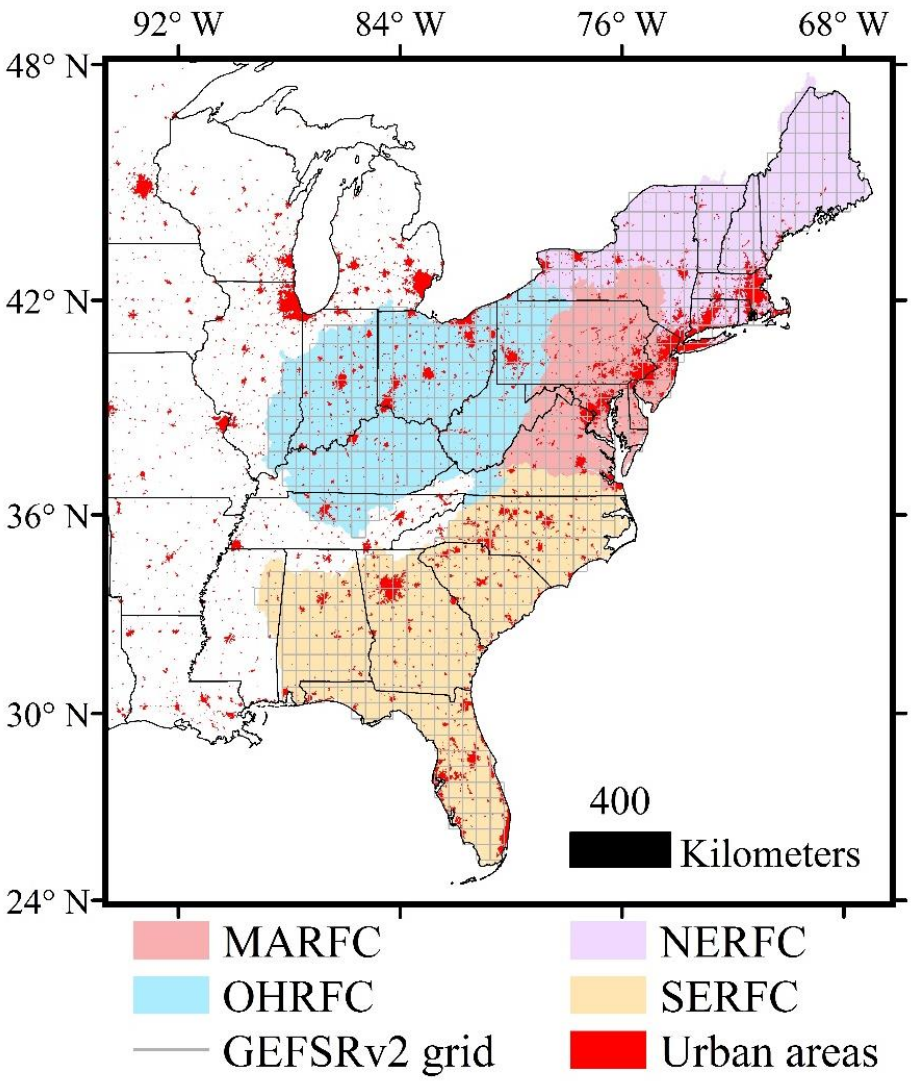


Fig. 1. Map illustrating the spatial extent of the different River Forecasts Centers in the eastern U.S., including the MARFC, NERFC, OHRFC, and SERFC. The map also shows the GEFSRv2 grid over each RFC and urban areas across the eastern U.S.

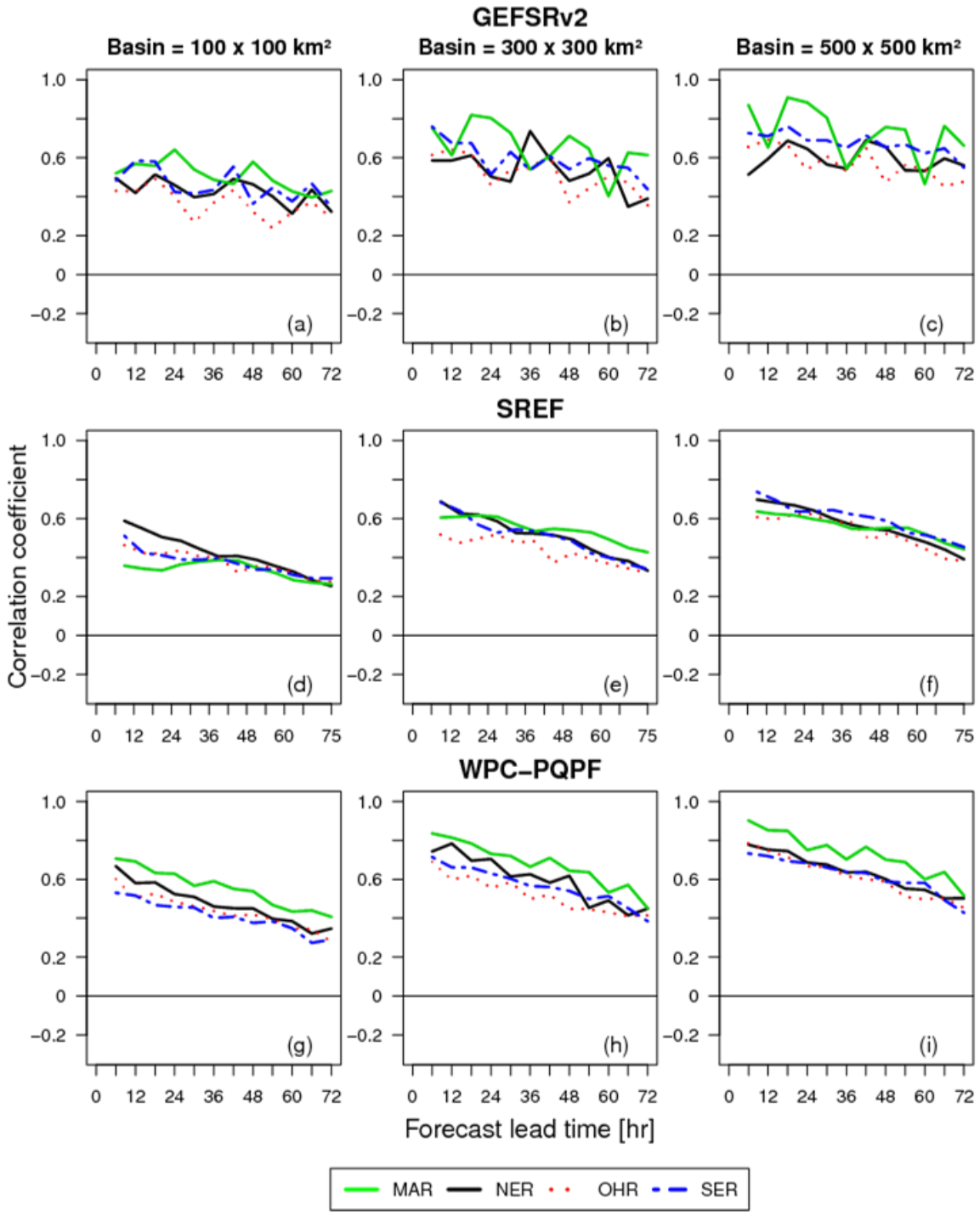


Fig. 2. Correlation coefficient between the mean ensemble forecast and the corresponding observed precipitation values as a function of the forecast lead time for the eastern regions. The correlation coefficient plots are for the (a-c) GEFSRv2, (d-f) SREF, and (g-i) WPC-PQPF, based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and different basin sizes.

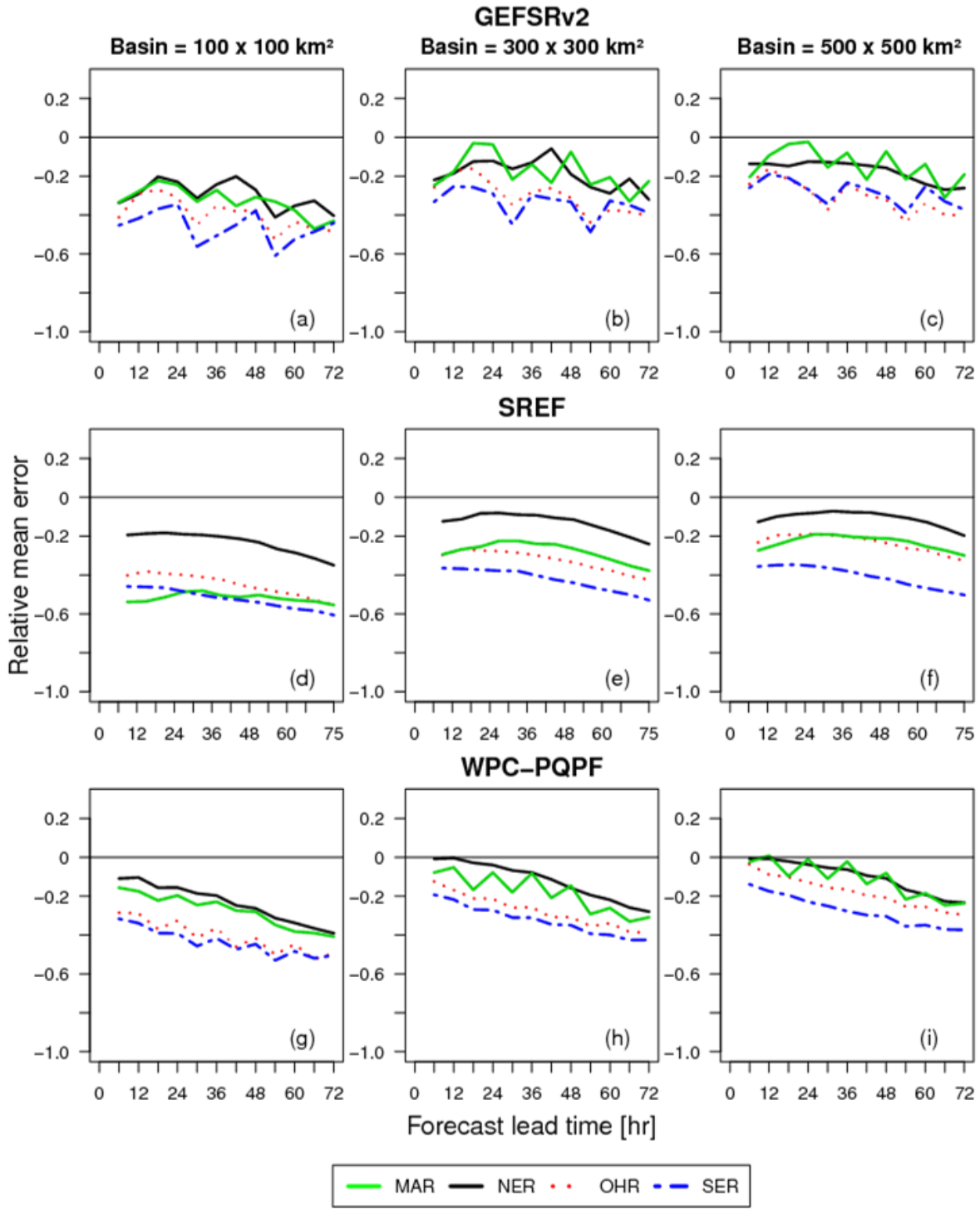


Fig. 3. RME of the mean ensemble forecast versus the forecast lead time for the eastern regions. The RME plots are for the (a-c) GEFSRv2, (d-f) SREF, and (g-i) WPC-PQPF, based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and different basin sizes.

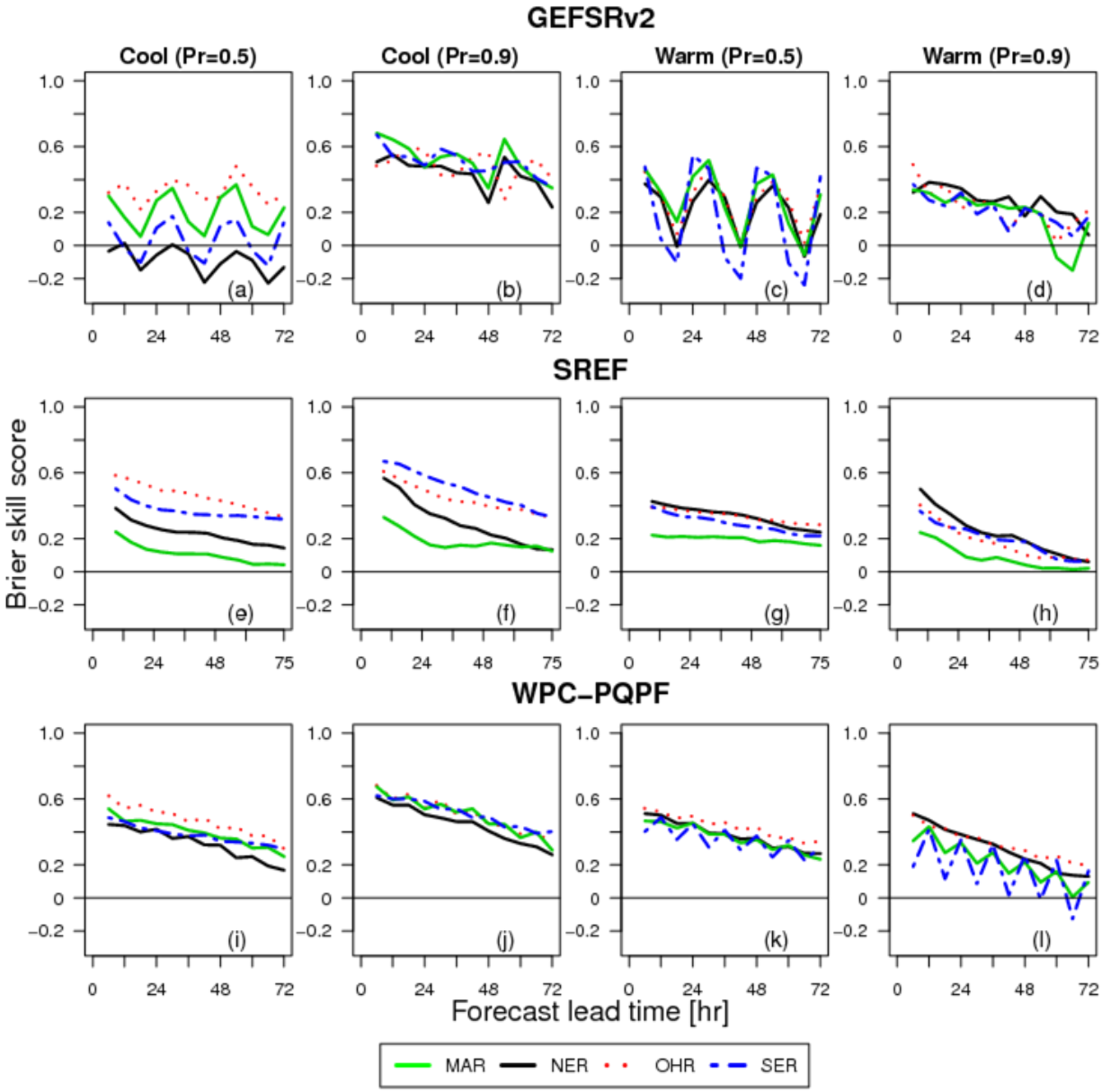


Fig. 4. BSS versus the forecast lead time for the eastern regions. The BSS plots are for the (a-d) GEFSRv2, (e-h) SREF, and (i-l) WPC-PQPF for both the cool and warm season. The BSS plots are based on 6 hourly precipitation accumulations, a basin size of 500 x 500 km^2, and both light to moderate precipitation events (Pr=0.5) as well as moderate to heavy precipitation events (Pr=0.9).

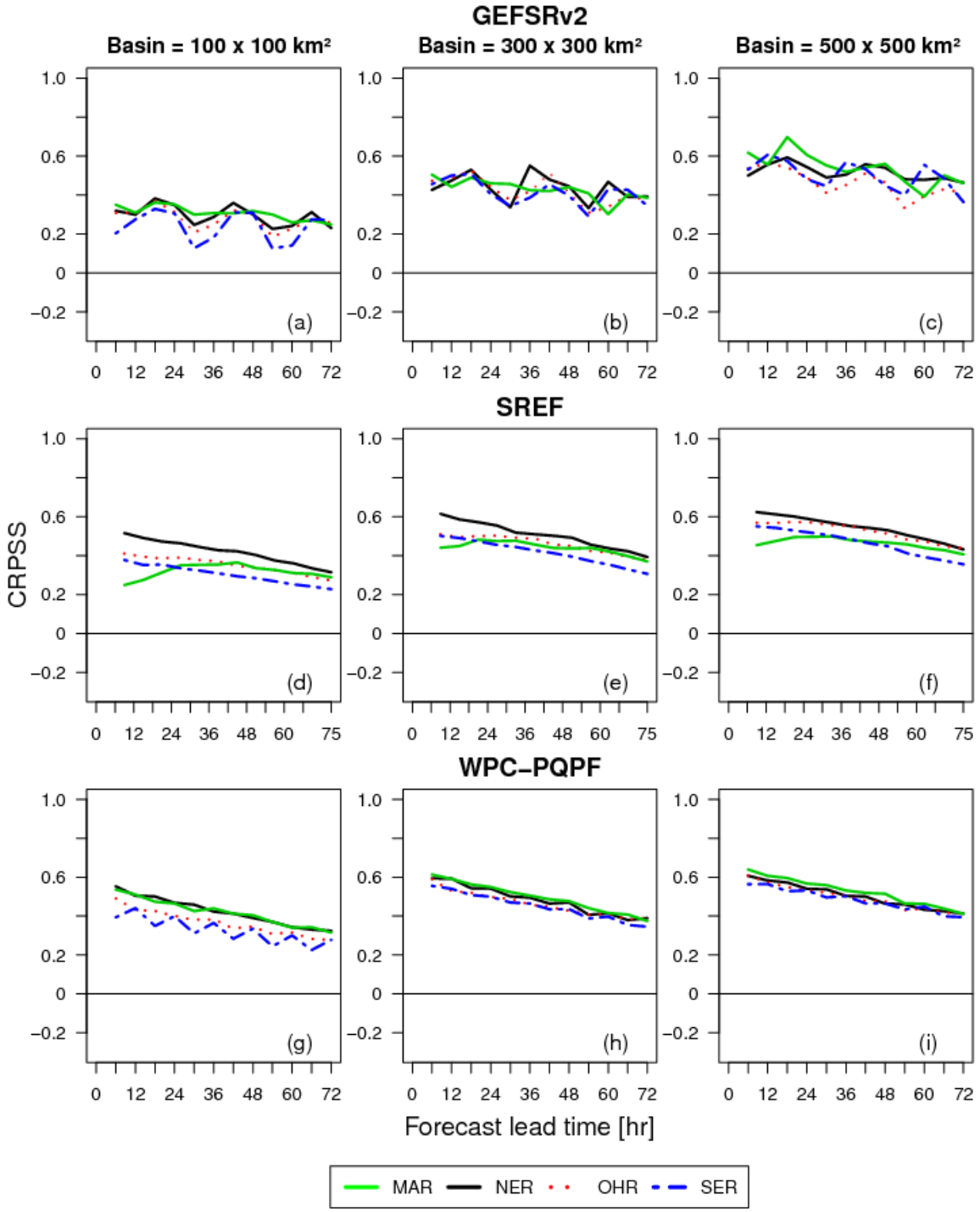


Fig. 5. Mean CRPSS versus the forecast lead time for the eastern regions. The CRPSS plots are for the (a-c) GEFSRv2, (d-f) SREF, and (g-i) WPC-PQPF. The CRPSS plots are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and different basin sizes.

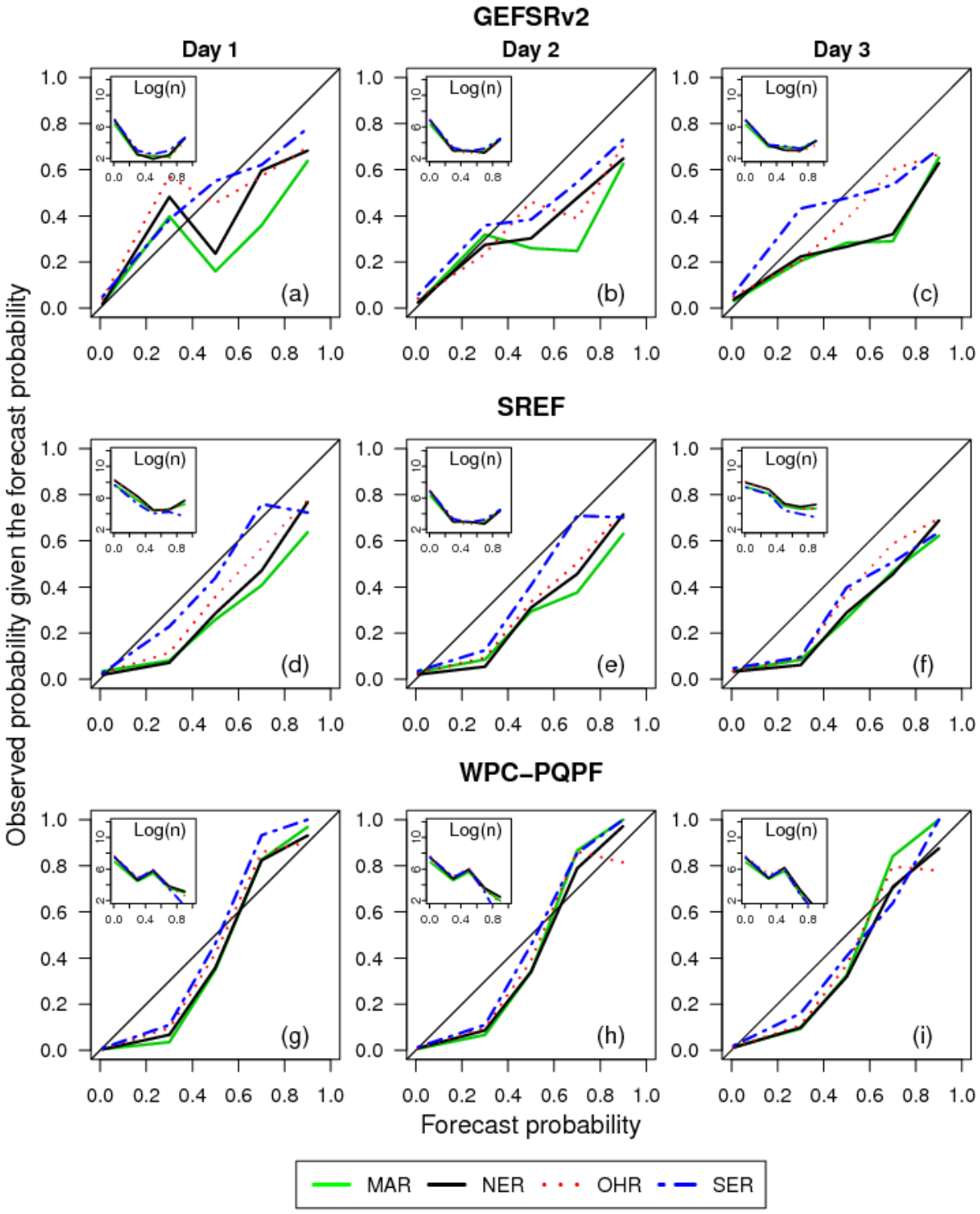


Fig. 6. Reliability diagrams for precipitation forecasts from the (a-c) GEFSRv2, (d-f) SREF, and (g-i) WPC-PQPF for forecast lead times of 1 (19-24 h), 2 (43-48 h) and 3 (67-72 h) days for the eastern regions. The reliability diagrams are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and 500 x 500 km^2 basin sizes. The insets show the sample size in logarithmic scale of the different forecast probability bins.

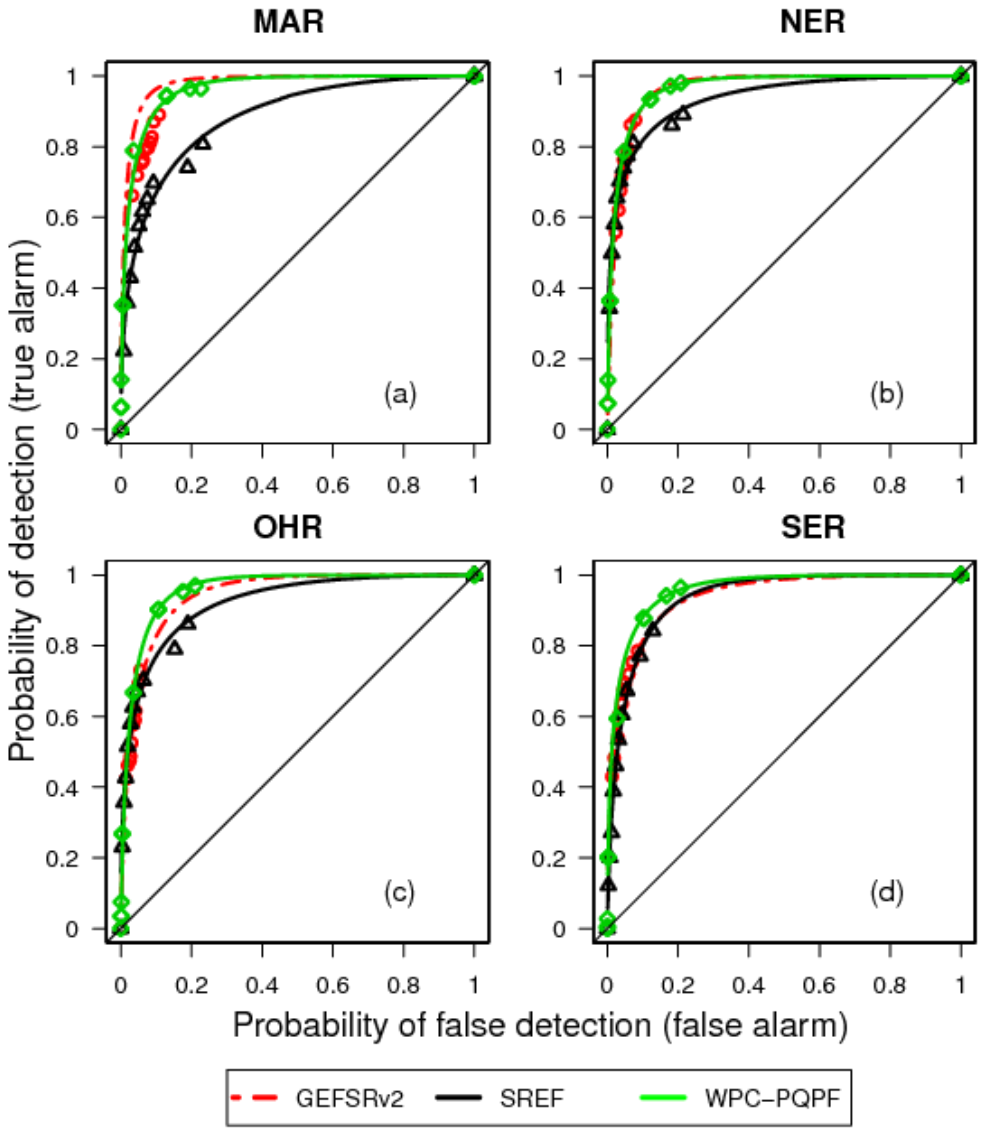


Fig. 7. ROC curves for the GEFSRv2, SREF, and WPC-PQPF for the (a) MAR, (b) NER, (c) OHR, and (d) SER at a lead time of 1 (19-24h) day. The symbols represent the sample values of the probability of detection and probability of false detection, and the curves represent the values fitted using the binomial distribution. All the ROC curves are based on 6 hourly precipitation accumulations and 500 x 500 km^2 basin sizes. The diagonal line represents the ROC curve associated with sampled climatology.

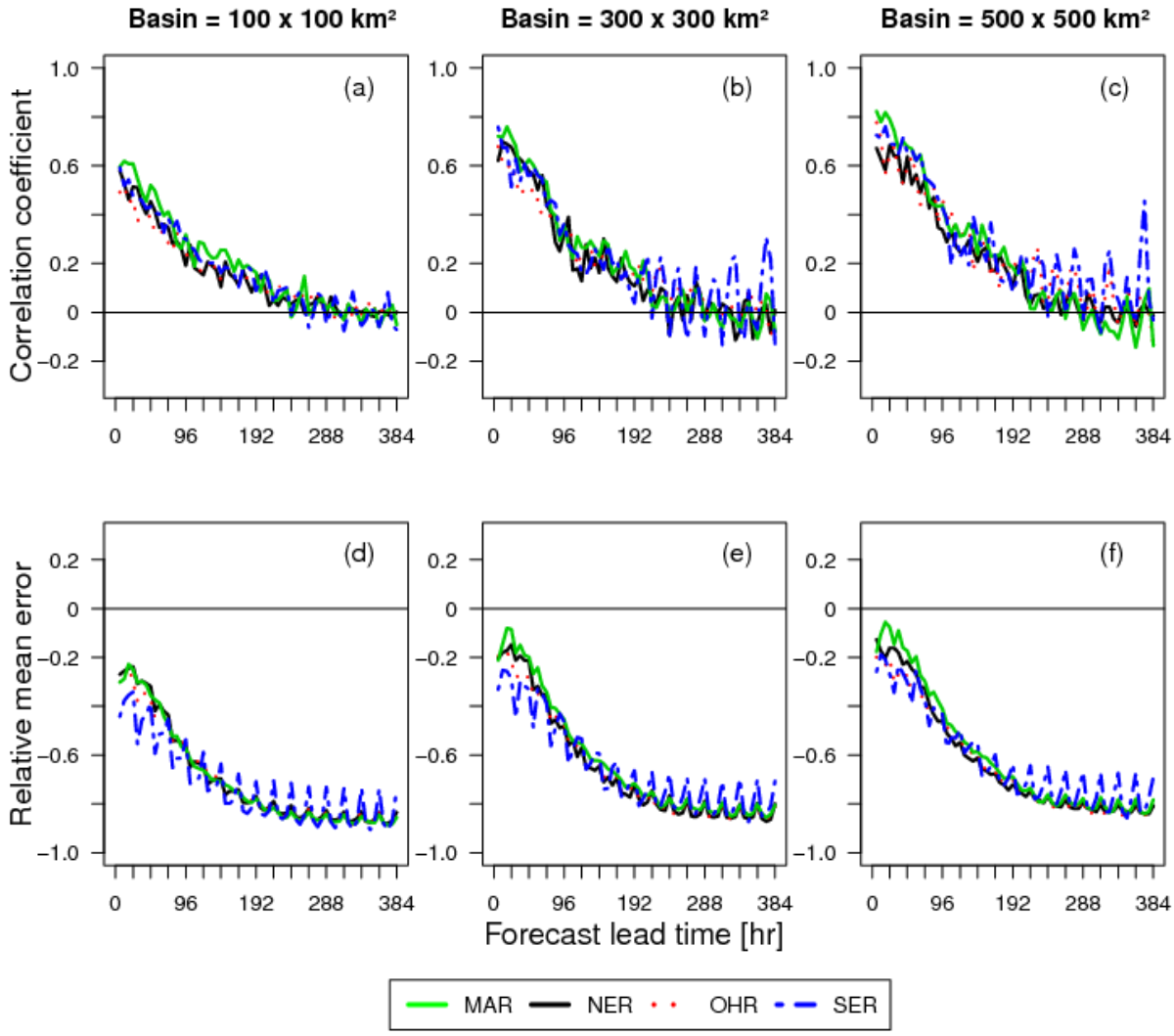


Fig. 8. (a)-(c) Correlation coefficient and (d)-(f) RME between the GEFSRv2 mean ensemble forecast and the corresponding observed precipitation values as a function of the forecast lead time for the eastern regions. The plots are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and basin sizes of (a) 100 x 100, (b) 300 x 300, and (c) 500 x 500 km^2.

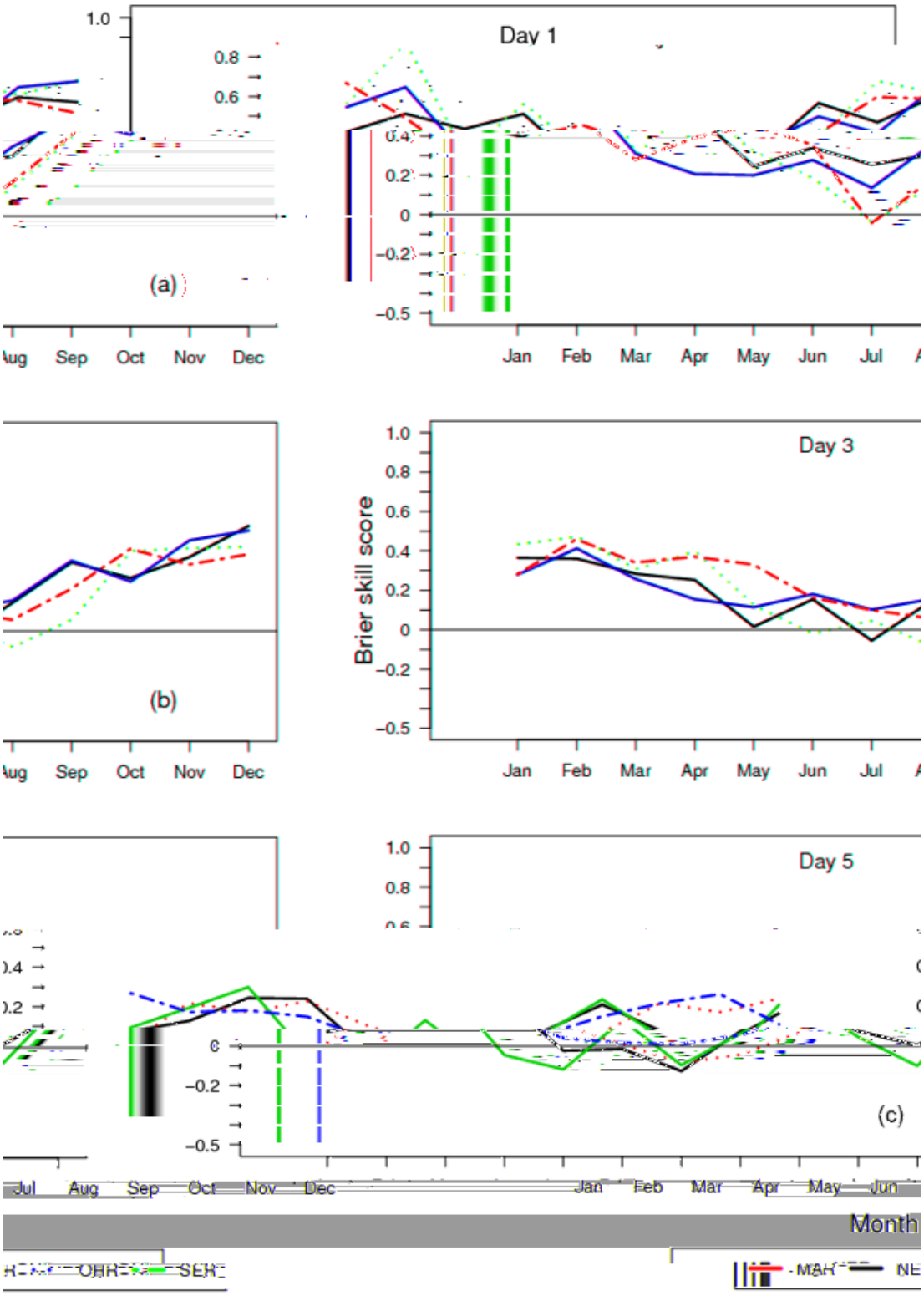


Fig. 9. Monthly BSS for the GEFSRv2 precipitation forecasts versus the calendar month for the eastern regions. The plots are for lead times of (a) 1 (19-24h), (b) 3 (67-72hr), and 5 (115-120h) days, and are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and 500 x 500 km^2 basin sizes.

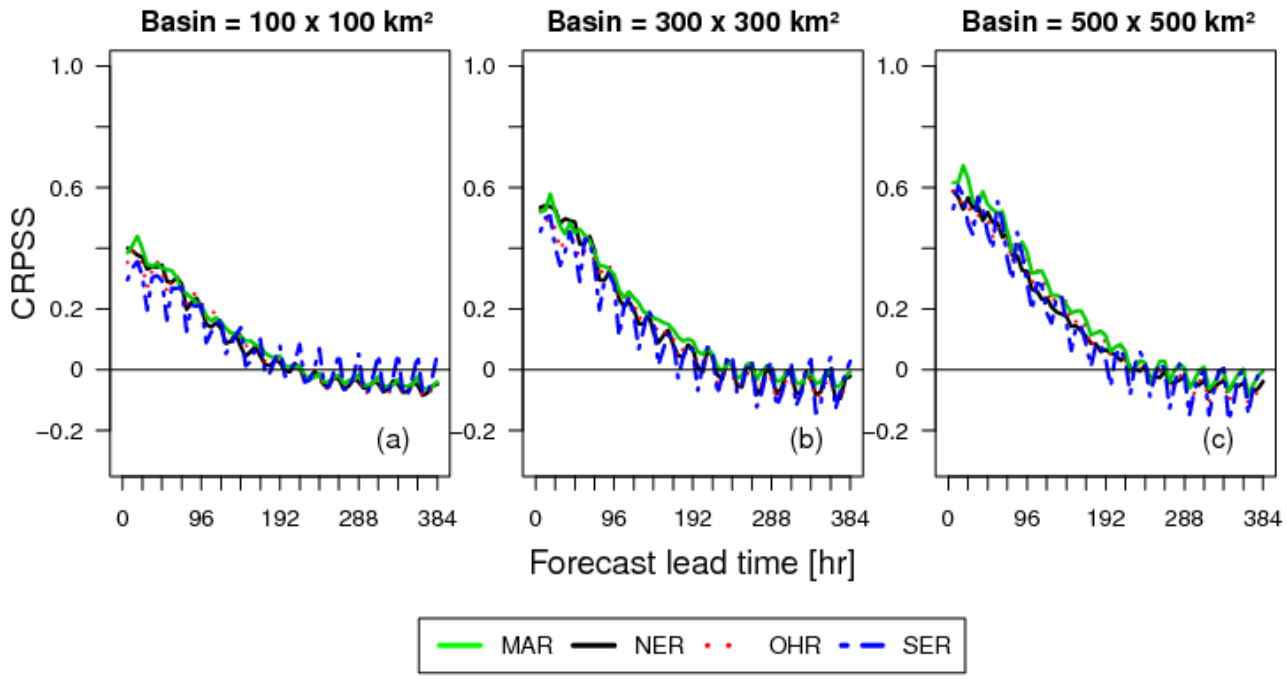


Fig. 10. Mean CRPSS for the GEFSRv2 precipitation forecasts versus the forecast lead time for the eastern regions. The plots are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and basin sizes of (a) 100 x 100, (b) 300 x 300, and (c) 500 x 500 km^2.

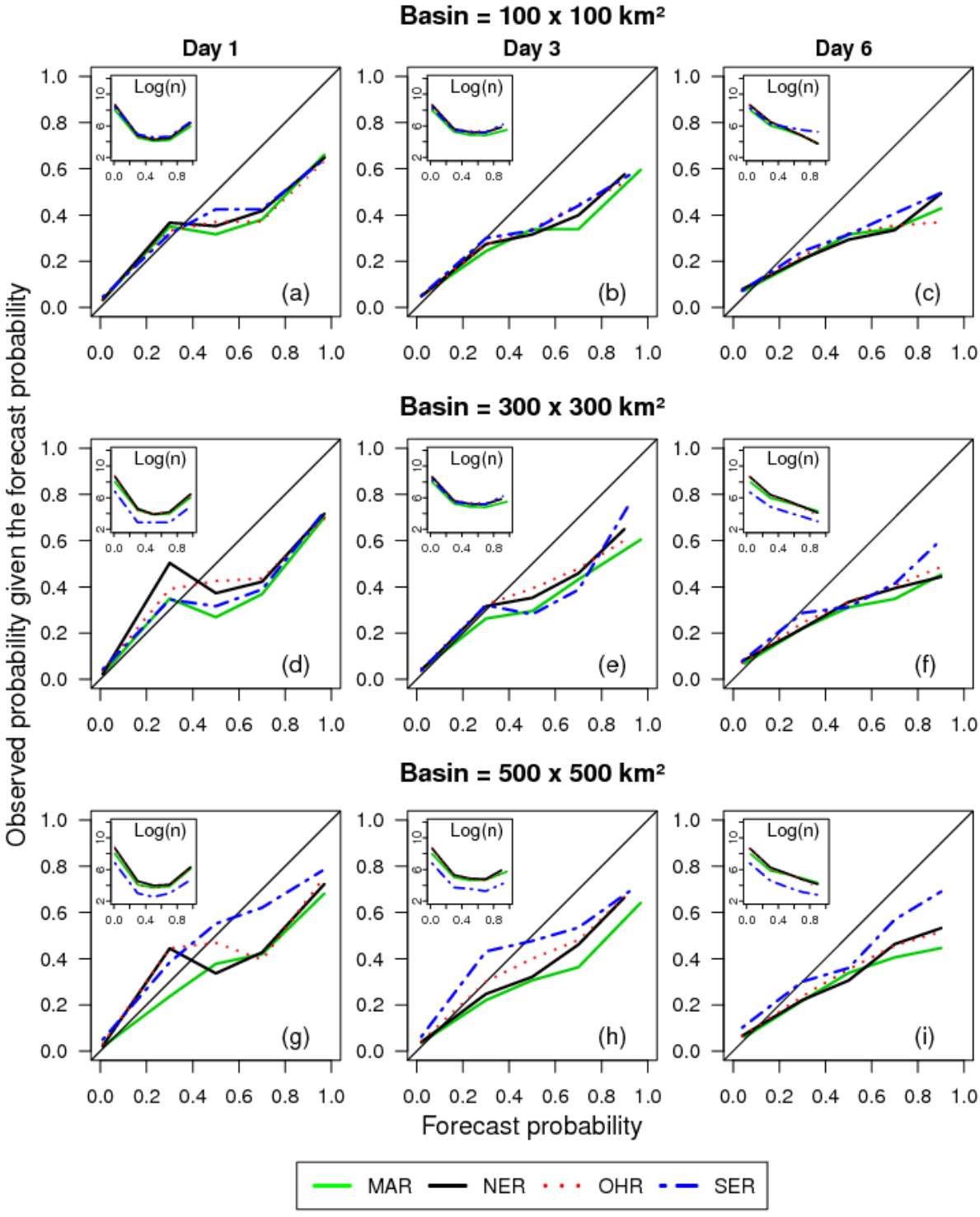


Fig. 11. Reliability diagrams for precipitation forecasts from the GEFSRv2 for basin sizes of (a-c) 100 x 100, (d-f) 300 x 300, and (g-i) 500 x 500 km^2 for the eastern regions. The reliability diagrams are for forecast lead times of 1 (19-24h), 3 (67-72h) and 6 (139-144h) days, and are based on 6 hourly precipitation accumulations and moderate to heavy precipitation events

(Pr=0.9). The insets show the sample size in logarithmic scale of the different forecast probability bins.

Chapter 4: Postprocessing of precipitation ensembles using Bayesian model averaging and heteroscedastic censored logistic regression

ABSTRACT

The potential of Bayesian model averaging (BMA) and heteroscedastic censored logistic regression (HCLR) to postprocess precipitation ensembles is investigated. For this, we use outputs from the National Oceanic and Atmospheric Administration's (NOAA's) National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2) dataset. The GEFSRv2 dataset is based on a single model and single physics ensembles. To implement BMA, we select two different modeling scenarios: exchangeable and non-exchangeable members. We term the BMA postprocessing with exchangeable members BMAx. As part of our experimental setting, to compare the postprocessors, we use 24-h precipitation accumulations and forecast lead times of 24- to 120-h. For the study area, we select the middle Atlantic region (MAR) of the U.S. In contrast with previous postprocessing studies, we consider here a wider range of forecasting conditions (e.g., the effect of spatial pooling, training length, lead time, precipitation threshold, and seasonality) when evaluating BMA and HCLR. Additionally, BMA and HCLR have not been compared against each other yet, under a common and consistent experimental setting.

To implement BMA and BMAx, we use a sliding window of 25 days and train each GEFSRv2 cell separately, as opposed to using spatial pooling. These training conditions were selected by carefully examining the skill of forecasts associated with different window lengths and number of cells. To compare and verify the postprocessors, we use different metrics (e.g., skills scores and reliability diagrams) conditioned upon the forecast lead time, precipitation threshold, and season. Overall, we find that HCLR tends to outperform BMA and BMAx but the differences among the postprocessors are not as significant. Also, BMA and BMAx behave similarly across lead times and seasons, thereby indicating that the GEFSRv2 members remain indistinguishable across lead times. The improved performance of HCLR over that of BMA seems related to the ability of HCLR to include the ensemble variance as a predictor. In the future, an alternative approach could be to combine HCLR with BMA to take advantage of their relative strengths.

1. Introduction

Numerical weather prediction (NWP) models are used, as part of an ensemble prediction system (EPS), to generate ensemble forecasts of a future weather variable or quantity (Tracton and Kalnay 1993; Toth et al. 2003; Buizza et al. 2005). The ensemble forecasts, in turn, can be used to determine the probability and uncertainty of the weather variable. In the case of precipitation forecasts, however, the magnitude and dispersion of the ensemble forecasts are normally characterized by the presence of biases (Sloughter et al. 2007; Wilks 2009), which makes the determination of forecast probabilities from such ensembles unreliable. To correct the biases and improve the reliability of ensemble forecasts, a number of techniques have been developed and implemented (e.g., Raftery et al. 2005; Wilks 2006b; Bröcker and Smith 2008). These techniques are collectively known as statistical weather postprocessing or calibration.

Postprocessing for ensemble prediction systems has several goals: correct systematic forecast errors or biases, correct (calibrate) ensemble spread so that it is a useful estimate of

forecast uncertainty, and (optionally) weight ensemble members according to past performance. Some of the available techniques for postprocessing weather forecasts are: regression-based methods (Bjørnar Bremnes 2004; Clark and Hay 2004; Hamill et al. 2004; Friederichs and Hense 2007; Wilks 2009; Roulin and Vannitsem 2011; Messner et al. 2014a,b), Gaussian ensemble dressing (Roulston and Smith 2003; Wang and Bishop 2005), non-parametric methods (Brown and Seo 2010), and Bayesian model averaging (BMA) (Raftery et al. 2005; Sloughter et al. 2007; Schmeits and Kok 2010), among others (e.g., Wu et al. 2011). Many of these techniques share in common the model output statistics (MOS) approach (Glahn and Lowry 1972; Wilks 2006b) since, as part of their methodology, they require the derivation of statistical forecast equations as a function of one or more outputs (i.e. predictors) from the NWP model. Additionally, some of the proposed techniques allow the complete characterization of the predictive probability density function (pdf) of precipitation forecasts (Sloughter et al. 2007; Wilks 2009; Messner et al. 2014b).

Some of the postprocessing techniques mentioned have been evaluated for the case of ensemble precipitation forecasts (Sloughter et al. 2007; Wilks and Hamill 2007; Wilks 2009; Brown and Seo 2010; Schmeits and Kok 2010; Messner et al. 2014a,b; Zhu et al. 2015). For instance, Sloughter et al. (2007) extended the BMA approach introduced by Raftery et al. (2005) to the case of ensemble precipitation forecasts. As a statistical weather postprocessor, BMA generates bias-corrected predictive pdfs from the ensemble forecasts (Sloughter et al. 2007; Fraley et al. 2010). Bremnes (2004) employed quantile regression to estimate the conditional quantiles of future precipitation using the forecast precipitation amounts, alongside other weather-related variables such as mean relative humidity and wind flow, as predictors. Wilks (2009) proposed and implemented the extended logistic regression (ELR) approach to include the threshold quantiles of the precipitation forecast as predictor variables, as opposed to relying on the precipitation amounts alone. Messner et al. (2014a) complemented the ELR approach by including the precipitation ensemble spread as a predictor. They termed this approach heteroscedastic extended logistic regression (HELR). They also proposed two additional logistic regression-based approaches for postprocessing precipitation: heteroscedastic ordered logistic regression (HOLR) and heteroscedastic censored logistic regression (HCLR) (Messner et al. 2014b). It is useful to note that HCLR fits the same model as HELR, with the only difference being that the HCLR parameters optimize the continuous predictive pdf, as opposed to the quantile thresholds (Messner et al. 2014b).

A few precipitation postprocessing studies have compared the performance of different postprocessing techniques under a common set of experimental conditions, e.g., by using the same geographic region, dataset, and training period to evaluate the postprocessors (Wilks 2006a; Sloughter et al. 2007; Schmeits and Kok 2010; Mendoza et al. 2014; Messner et al. 2014b). The general findings from these studies indicate that the performance of the postprocessors, both relative to sampled climatological conditions and to each other, vary depending on the training strategy (Greybush et al. 2008; Zhu et al. 2015), verification metric considered (Mendoza et al. 2014), forecast lead time (Schmeits and Kok 2010), and bias-correction type (Schmeits and Kok 2010; Erickson et al. 2012), among other factors.

In this study, our primary goal is to assess and verify the potential of BMA and HCLR to postprocess precipitation ensemble reforecasts from the National Oceanic and Atmospheric Administration's (NOAA's) National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2). We employ GEFSRv2 since its reforecasts, based on a consistent model run, are available over a long time period. This

is relevant because forecasts produced by a model whose structure changes in time will produce less statistically consistent forecasts. Although this situation may be unavoidable in operational forecasting, it should be avoided when interest lies in assessing the performance of different postprocessors. We use multisensor precipitation estimates (MPEs) as the observed precipitation when training the postprocessors and verifying the raw and postprocessed ensemble precipitation forecasts. Additionally, we highlight that our evaluation here of BMA is more comprehensive than previous ones since we account for the effect of training period length, spatial pooling strategy, lead time, seasonality, and BMA weight interpretation (i.e. exchangeable versus non-exchangeable) on the BMA postprocessed precipitation forecasts. Moreover, BMA and HCLR have not been compared against each other yet.

We select BMA and HCLR for this study for various reasons. BMA is desirable because it provides an integrated approach for combing ensemble members from a single or multiple NWP models. At the same time, techniques based on logistic regression have been shown to perform as well as or slightly better than BMA in several applications (Sloughter et al. 2007; Schmeits and Kok 2010), while being less computationally demanding. The latter becomes particularly relevant when working with long reforecast datasets. Furthermore, HCLR has recently been shown to outperform and overcome key shortcomings of other logistic regression-based techniques, such as allowing the determination of the full predictive pdf of precipitation forecasts (Messner et al. 2014a).

Key questions that we seek to address with this study are: How does the BMA and HCLR postprocessed forecasts compare against the raw precipitation ensembles? What is the dependence between the performance of the postprocessors and the forecast lead time, training period length, spatial pooling, seasonality, and precipitation threshold? Does assuming exchangeable versus non-exchangeable weights affect the performance of BMA across lead times? Which postprocessing method is more reliable for the MAR? The remainder of the paper is organized as follows. In sections 2 and 3, we describe the study area and datasets employed, respectively. In section 4, we review the postprocessing techniques. Section 5 outlines the verification strategy. The main results and their implications are examined in sections 6 and 7. Lastly, section 8 summarizes the key findings.

2. Study area

We use the Middle Atlantic Region (MAR) of the U.S. as our study area. The geographic location and boundary of the MAR is illustrated in Figure 1. The MAR is comprised by the state of Delaware and the District of Columbia, along with parts of the states of Maryland, New York, New Jersey, Pennsylvania, Virginia, and West Virginia (Polsky et al. 2000; Greene et al. 2005). It only occupies approximately 5% of the total land mass of the U.S. but it contains approximately 10% of its population (~41 million people) (Siddique et al. 2015). Some of the largest metropolitan areas in the U.S. are located in the MAR, e.g., Baltimore, Philadelphia, and Washington D.C. Additionally, the MAR comprises several major U.S. river basins including the Delaware, Susquehanna, Potomac, and James River. The climate in the MAR is relatively humid. The average annual temperature is approximately 11 °C and the mean annual precipitation is approximately 900-1200 mm (Polsky et al. 2000). Rainfall is distributed evenly throughout the year with the mean annual rainfall total being approximately 1009 mm; annual rainfall has ranged from approximately 647 to 1288 mm over the historical record (Neff et al. 2000).

The MAR has a high frequency of heavy precipitation events relative to other regions in the Continental U.S. (CONUS), particularly in the summer months, as indicated by

climatological analysis of heavy rainfall events (i.e. hourly accumulations of at least 25 mm at a 4x4 km^2 grid cell) at 1-3-h durations across CONUS (Hitchens et al. 2013). In relation to the patterns of large-scale heavy precipitation events, Grumm and Holmes (2007) classified events over the MAR, using both station and reanalysis data, to find that the dominat event types are Maddox synoptic and Maddox frontal (Maddox et al. 1979). They also identified a sub-type of the Maddox synoptic characterized by the interactions between synoptic events with the remnants of tropical and subtropical systems. Indeed, they highlight that these sub-type events produce the heaviest rainfall events over the MAR. Generally, the magnitudes, patterns, and anomalies associated with wind components, precitable water, and 850 hPa specific humidity are useful signatures for predicting heavy precipitation over the MAR (Grumm and Holmes 2007).

3. Data and methodology

a. GEFSRv2

For the precipitation ensemble forecasts, we use outputs from the GEFSRv2 dataset. GEFSRv2 are the retrospective forecasts produced using the 2012 operational version (version 9.0.1) of the NCEP's Global Ensemble Forecast System (Hamill et al. 2013). The model runs for the GEFSRv2 were initiated once a day at 00 Coordinated Universal Time (UTC) (Hamill et al. 2013). Initial conditions were perturbed using the ensemble transform technique with rescaling (Wei et al. 2008). The forecast lead times extend from 1 to 16 days and each forecast cycle consists of forecasts valid for 3 hourly accumulations from day 1 to day 3 and 6 hourly accumulations from day 4 to day 16. We use here for the evaluation of the postprocessors 24-hr accumulations from day 1 to 5. The native resolution of the reforecasts is ~0.5-degree on a Gaussian grid for forecasts in the first week and ~0.67-degree for forecasts in the second week. The GEFSRv2 data is also available at the ~1-degree resolution for the entire range of lead times (days 1 to 16). We use here the 1-degree resolution dataset. Further details about the GEFSRv2 dataset or information on how to access it are provided elsewhere (Hamill et al. 2013; Hamill et al. 2015).

a. MPEs

We use MPEs to train the postprocessors and verify the raw and postprocessed ensemble precipitation forecasts. The MPEs were provided by the NOAA's Middle Atlantic River Forecast Center (MARFC) (Lawrence et al. 2003). This dataset is similar to the NCEP stage-IV MPEs (Prat and Nelson 2015). As with the NCEP stage-IV dataset, the MPEs provided by the MARFC represent a continuous time series of hourly, high-resolution gridded precipitation observations at 4x4 km^2 cells, over the MAR. We aggregated the MPEs to the temporal (24-hr) and spatial scale (1-degree) of the GEFSRv2 data over the period 2002-2007. Note that MPEs are subject to errors, such as radar artifacts, but they are also one of the best high-resolution gridded precipitation datasets available (Prat and Nelson 2015) and therefore appropriate for this study.

b. Postprocessing techniques

1) *BMA*

We provide here a brief overview of the BMA technique as used for the postprocessing of ensemble precipitation forecasts since a detailed description is provided elsewhere (Sloughter et al. 2007). As a statistical weather postprocessor, BMA generates bias-corrected predictive pdfs from the ensemble forecasts (Sloughter et al. 2007; Fraley et al. 2010). Specifically, the BMA predictive pdf is a weighted average of pdfs centered on the individual bias-corrected

precipitation forecasts. The weights reflect the predictive skill of the individual ensemble members over a selected training prediod.

The BMA predictive pdf, $p(y| f_1,\ldots, f_k)$, for the cube root of precipitation accumulation y, given the forecast members $f_1,\ldots, f_k$ at a particular lead time, is given by:

$$p(y \mid f_1,..., f_k) = \sum_{k=1}^{K} w_k \{P(y=0 \mid f_k) I[y=0] + P(y>0 \mid f_k) g_k(y \mid f_k) I[y>0]\}. \quad (1)$$

The weight w_k is the posterior probability of ensemble member k being the best one, provided that $\sum_{k=1}^{K} w_k = 1$. K is the total number of ensemble members; K=11 for the GEFSRv2 data. The weights are specified according to the relative performance of each ensemble member during the training period employed for parameter estimation. $P(y=0|f_k)$ is the probability of the cube root of precipitation being equal to zero given the forecast member f_k and assuming that f_k is the best forecast member. I[.] is the indicator function which is equal to 1 if the term inside the brackets holds true and 0 otherwise. $P(y>0|f_k)$ is the probability of the cube root of precipitation being greater than 0 given the forecast member f_k and assuming that f_k is the best forecast member. The cube root of precipitation is used since this transformation has been found to improve the modeling of $P(y>0|f_k)$ (Sloughter et al. 2007), which is normally represented by a gamma pdf as further explained in the next paragraphs.

The term $P(y=0|f_k)$ is determined as

$$\text{logit}P(y=0 \mid f_k) \equiv \log \frac{P(y=0 \mid f_k)}{P(y>0 \mid f_k)} = a_{0,k} + a_{1,k} f_k^{1/3} + a_{2,k}\delta_k. \quad (2)$$

Equation (2) is a logistic regression with parameters $a_{i,k}$ (i=1,2,3) that need to be estimated for each ensemble member k. The predictor δ_k is equal to 1 if f_k=0 and 0 otherwise. The parameters in equation (2) are determined directly from the ensemble forecast and observed data, using logistic regression with precipitation/no precipitation as the dependent variable, and $f_k^{1/3}$ and δ_k as the two predictor variables.

The term $P(y>0|f_k)$ is equal to 1-$P(y=0|f_k)$ while $g(y|f_k)$ is defined as

$$g_k(y \mid f_k) = \frac{1}{\beta_k^{\alpha_k}\Gamma(\alpha_k)} y^{\alpha_k - 1} \exp(-y/\beta_k) \quad (3)$$

for y>0, and $g(y)$=0 for y=0. Equation (3) is a gamma pdf with shape parameter $\alpha_k = \mu_k^2/\sigma_k^2$ and scale parameter $\beta_k = \sigma_k^2/\mu_k$. The mean, μ_k, and variance, σ_k^2, of this distribution depend on f_k as follow

$$\mu_k = b_{0,k} + b_{1,k} f_k^{1/3} \quad (4)$$

and

$$\sigma_k^2 = c_0 + c_1 f_k. \quad (5)$$

The parameters $b_{i,k}$ (i=0, 1) in equation (4) are member specific. They are determined separately for each ensemble member using linear regression with the cube root of the observed precipitation amount as the dependent variable and $f_k^{1/3}$ as the predictor variable.

Lastly, using the training data, the parameters c_o and c_1 in equation (5), as well as the w_k's (k=1,…, K) in equation (1) are estimated by maximum likelihood, as in Sloughter et al. (2007). The approach of Sloughter et al. (2007) maximizes the log-likelihood function numerically using the expectation-maximization algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997).

To implement the BMA postprocessor, we use 24-h precipitation accumulations from the GEFSRv2 for lead times from 24- to 120-h. To train the BMA, we use the sliding time window approach of Sloughter et al. (2007). In this approach, a sliding time window comprised of the L training days preceding the forecast day is used. The window moves with the forecast day (i.e. the day the forecast is issued) and, typically, it is comprised of the preceding 20 to 40 days prior to the forecast day. We use this same approach here with one important modification. We select training days from the 4 years preceding the forecast day using the same calendar days in each year, as opposed to just using training days from a single year. For example, for a GEFSRv2 reforecast issue on March 31, 2005, we select as the training data the days from March 1 to 30 (assuming a 30-day training window) in the years 2002 to 2005, thus we use in this example a total of 120 training days, i.e. (30 days)x(4 years). We select the size of the training window empirically by testing different window sizes.

Additionally, when training the BMA algorithm, it is common to rely on spatial pooling to increase the sample size of the training dataset. However, the effect of spatial pooling on the performance of BMA is rarely assessed. Thus, we evaluate this here by varying the number of GEFSRv2 cells that are used for training. In this study, we select a total of 20 GEFSRv2 cells since they cover the majority of the MAR. To test different training scenarios, we use 1, 5, 10, and 20 cells to train the BMA algorithm. The 1 cell scenario means that each cell is trained individually without pooling data from the other cells. In contrast, the 5 cells scenario means that the 20 GEFSRv2 cells that encompass the MAR are divided into 4 groups of neighboring cells with 5 cells in each group. Each group is then trained separately by pooling the data from its 5 cells. For example, for the case of 5 cells and a 30-day sliding window, we use 6600 reforecasts to train the BMA algorithm at a given forecast day, i.e. (30 days)x(5 cells)x(4 years)x(11 members).

2) *BMA with exchangeable members*

Our previous description of BMA assumes that the ensemble members are individually distinguishable where distinct weights may have a physical interpretation. In our BMA postprocessing experiment, however, all the ensemble members come from the same NWP model, which means that the members lack individually distinguishable physical features. In this situation, the ensemble members are exchangeable, which means that the BMA weights can be assumed to be equal (Fraley et al. 2010; Schmeits and Kok 2010). We use the term BMAx to indicate the implementation of BMA using equal weights, i.e. w_k in equation (1) is equal to $1/K$. Additionally, the exchangeability condition makes other parameter constraints possible. Specifically, the parameters $a_{i,k}$ (i=1,2,3) in equation (2) and $b_{i,k}$ (i=0,1) in equation (4) are the same for all the exchangeable members that come from the same NWP model so that $a_{i,k}=a_i$ (i=1,2,3) and $b_{i,k}=b_i$ (i=0,1) (Fraley et al. 2010; Schmeits and Kok 2010).

Nonetheless, we note that there might still be significant differences, particularly at longer lead times, among ensemble members from the same model that could make the non-exchangeable approach meaningful and useful. Thus, we evaluate both approaches in this study, BMA and BMAx.

3) *HCLR*

HCLR is based on the logistic regression model initially proposed by Hamill et al. (2004) to postprocess precipitation ensembles. In essence, HCLR fits a logistic distribution to the transformed, in this case the square root of the ensemble mean, and bias-corrected precipitation

ensembles (Messner et al. 2014b). Additionally, HCLR uses the ensemble spread as a predictor, which allows HCLR to consider uncertainty information contained in the ensembles. We describe next the HCLR postprocessor as it evolved from the logistic regression model of Hamill et al. (2004) and the extended version of Wilks (2009).

The logistic regression model of Hamill et al. (2004) is given by

$$p(y \le q \mid \mathbf{x}) = \frac{\exp[f(\mathbf{x})]}{1+\exp[f(\mathbf{x})]} = \Lambda[f(\mathbf{x})], \tag{6}$$

where y is the transformed precipitation, q is a specified threshold, $\mathbf{x}$ is a vector of predictor variables, and $f(\mathbf{x})$ is a linear function of the predictor variables $\mathbf{x}$. Further, Messner et al. (2014a) noted that equation (6) has the same form as the cumulative distribution function (cdf) of the standard logistic distribution $\Lambda(.)$.

One limitation with equation (6) is that separate logistic regressions with different linear functions $f(\mathbf{x})$ need to be fitted to each threshold of interest (Wilks 2009). This results in logistic regressions that can cross each other which in turn implies the occurrence of nonsense negative probabilities. To overcome this limitation, Wilks (2009) extended the logistic regression model by adding another predictor variable for the threshold q such that

$$p(y \le q \mid \mathbf{x}) = \Lambda[g(q) - f(\mathbf{x})], \tag{7}$$

where the transformation $g(.)$ is a monotone nondecreasing function. In addition to avoiding negative probabilities, equation (7) has the advantage that fewer parameters need to be estimated; instead of having a linear function $f(\mathbf{x})$ for each threshold, $f(\mathbf{x})$ is now the same for all the thresholds. This can be particularly relevant when dealing with small training datasets.

Furthermore, to appropriately utilize the uncertainty information in the ensemble spread, Messner et al. (2014a) proposed the use of an additional predictor vector $\boldsymbol{\varphi}$ to control the dispersion of the logistic predictive distribution,

$$P(y \le q \mid \mathbf{x}) = \Lambda\left\{\frac{g(q) - f(\mathbf{x})}{\exp[h(\boldsymbol{\varphi})]}\right\}, \tag{8}$$

where $h(.)$ is another linear function that has to be estimated. The exponential function in the denominator of equation (8) is used as a simple method to ensure positive values (Messner et al. 2014a). Messner et al. (2014a) termed HELR the approach based on equation (8).

In HELR, the function $f(\mathbf{x})$ is defined as

$$f(\mathbf{x}) = d_0 + d_1\overline{\sqrt{x_{\text{ens}}}}, \tag{9}$$

where d_0 and d_1 are parameters that need to be estimated, and the predictor variable $\overline{\sqrt{x_{\text{ens}}}}$ is the mean of the transformed, via the square root, ensemble forecasts. $h(\boldsymbol{\varphi})$ is defined as

$$h(\boldsymbol{\varphi}) = e_0 + e_1\varphi, \tag{10}$$

where e_0 and e_1 are parameters that need to be estimated, and φ is the standard deviation of the square root transformed, precipitation ensemble forecasts.

To determine the parameters associated with equation (8), maximum likelihood estimation with the log-likelihood function is used (Messner et al. 2014a, b). For this, one needs to determine the predicted probability π_i of the ith observed outcome. When determining π_i, one should account for the fact that $y \ge 0$. One variation of the HELR model that can easily accommodate nonnegative variables that are continuous for positive values and have a natural threshold at zero, such as precipitation amounts, is censored regression or, as termed by Messner et al. (2014b), HCLR. For HCLR, π_i can be expressed as (Messner et al. 2014b)

$$\pi_i = \begin{cases} \Lambda\left[\dfrac{g(0) - f(\mathbf{x})}{\exp[h(\boldsymbol{\varphi})]}\right] & y_i = 0 \\ \lambda\left[\dfrac{g(y_i) - f(\mathbf{x})}{\exp[h(\boldsymbol{\varphi})]}\right] & y_i > 0, \end{cases} \tag{11}$$

where λ[.] denotes the likelihood function of the standard logistic function. In essence, HCLR fits a logistic error distribution with point mass at zero to the transformed predictand. Such an error distribution appears reasonable for dealing with the square root transformed precipitation amounts (Schefzik et al. 2013; Scheuerer 2014).

As was the case with BMA, to implement the HCLR postprocessor, we use 24-h precipitation accumulations from the GEFSRv2 for lead times from 24- to 120-h. To train the HCLR, we use a modified version of the sliding window approach of BMA. We use a stationary training period for each season and year to be forecasted, comprised of the seasonal data from the previous four years. Thus, for example, we use the 90 days of summer data available from the previous four years to train the HCLR algorithm and forecast days within the summer season of the current year. In this case, the total number of forecasts used for training is (90 days)x(4 years)x(11 members). Note that the training window is moved forward one entire year after all the forecast days in that year have been forecasted.

d. Verification strategy

To verify the raw and postprocessed ensemble precipitation forecasts, we use the Ensemble Verification System (EVS) (Brown et al. 2010). We use for the verification analysis different metrics, including the Brier skill score (BSS), continuous ranked probability skill score (CRPSS), and reliability diagram. We also examine the decomposed components of the CRPS. The definition of each of these metrics is provided in the Appendix. Additional details about the verification metrics can be found elsewhere (e.g., Wilks 2010; Jolliffe and Stephenson 2012).

For the verification analysis, we use two years of data, 2006 and 2007, the remaining years, 2002-2005, are used to train the postprocessors. The verification is done conditionally upon the season, lead time, and precipitation threshold. We focus our verification on moderate precipitation amounts. For this, we select precipitation amounts greater than that implied by a non-exceedance probability, in the sampled climatological probability distribution, of approximately 0.9 (~10 mm). To account for the effect of spatial scale on postprocessing, we assess the influence of spatially pooling data to train the postprocessors.

4. Results and discussion

a. Selection of the training length for BMA/BMAx

An initial step in implementing the BMA/BMAx postprocessor is to determine the appropriate training length for the sliding time window approach of BMA/BMAx (Fraley et al. 2010; Sloughter et al. 2007). If the length of the training window is too short or too long, the performance of BMA/BMAx can become suboptimal or less skillful. To assess the effect of the training length on the performance of BMAx, we plot the BSS against the training length for moderate precipitation events (>10 mm) in the summer (Figs. 2a and 2b) and fall (Figs. 2c and 2d). We find that the BSS tends to peak or reach a maximum value at a training length of ~25 days (Fig. 2). For the most part, after 25 days the value of BSS declines (Fig. 2). This is the case for both, forecast lead times of 1 (Figs. 2a and 2c for the summer and fall, respectively) and 5 days (Figs. 2b and 2d for the summer and fall, respectively). The results are similar

independently of the number of GEFSRv2 cells used to train the BMAx algorithm (Fig. 2), i.e. the optimum value of the training length still tends to be ~25 days. For example, in Fig. 2a, when using 20 cells or training each cell separately (1 cell), both curves reach a maximum at 25 days.

Fig. 3 shows the same information as Fig. 2 but plots instead the CRPSS against the training length. In Fig. 3, the general tendency is as in Fig. 2, the skill of the BMAx postprocessed forecasts tends to reach a maximum at ~25 days. We thus select for this study 25 days as the training length for all of our BMA-based experiments. We note that similar results (not shown) were obtained for both BMA and BMAx. Additionally, our findings regarding the training length are in agreement with previous results obtained by Sloughter et al. (2007), althouh we consider longer lead times here than Slougheter et al. (2007). Specifically, for forecast lead times of 2 days, Sloughter et al. (2007) found the optimal training length of BMA to be ~30 days.

b. Effect of spatial pooling on the performance of the postprocessors

To assess the effect of spatial pooling on the performance of the postprocessors, we plot the BSS against the number of cells used to train the BMA, BMAx, and HCLR postprocessors (Figs. 4a and 4b for the summer and fall, respectively). To train the BMA and BMAx postprocessors, we use a training window of 25 days. The forecasts from all three postprocessors show notable gains in skill relative to the raw ensembles for the summer (Fig. 4a) but the gains seem largely insignificant for the fall (Fig. 4b). The general tendency in Fig. 4, nevertheless, is for the BSS to decline somewhat as the number of cells used for training are increased. Additionally, the HCLR seems to perform slightly better than both BMA and BMAx.

We also show the CRPSS as a function of the number cells used to train the BMA, BMAx, and HCLR postprocessors for the summer (Fig. 5a) and fall (Fig. 5b). For the summer (Fig. 5a), all of the postprocessors seem to significantly improve upon the raw ensembles, and the skill declines slightly as additional cells are used to train the postprocessors, as was the case with the BSS (Fig. 4a). For the fall (Fig. 5b), only HCRL seems able to improve upon the raw ensembles; however, overall the differences in skill among the postprocessors appear not as significant (Fig. 5b).

According to the results in Figs. 4 and 5, for the remainder of our analysis, we train the postprocessors separately at each GEFSRv2 cell since this approach seems to perform somewhat better than when cells are spatially pooled. We note that this is different from the way BMA and BMAx are normally implemented (Sloughter et al. 2007; Fraley et al. 2010). Spatial pooling is normally required by BMA to increase the sample size used for training because the typical training window length of 25 to 30 days is small. We are less constrained here by the length of the training window since we sample data from the previous four years when training the postprocessors. This is feasible in this case because we are working with reforecasts but it may not be as feasible when dealing with outputs from an actual forecasting system.

c. Verification of the raw and postprocessed precipitation ensembles

1) BSS

The BSS indicates that generally the skill of the postprocessed ensemble precipitation forecasts is improved relative to the raw ensembles (Fig. 6). The relative improvements in skill are generally greater in the summer (Figs. 6a and 6b) than fall (Figs. 6c and 6d). Additionally, the improvements tend to be greater when considering all the precipitation events (Fig. 6c) than when focusing on moderate precipitation events (Fig. 6d). Overall, the skills gains from

postprocessing decline with increasing lead time. For example, for moderate precipitation events in the fall (Fig. 6d), the BSS associated with the different postprocessors is slightly better than the BSS of the raw ensembles at a forecast lead time of 1 day; however, the BSS of the postprocessed ensembles becomes slightly less at a lead time of 5 days. Contrasting the postprocessors against each other, it appears that the general tendency is for the postprocessors to perform similarly (Fig. 6). The differences between BMA and BMAx seem insignificant while HCLR tends to show a slight skill gain over both BMA and BMAx across lead times, precipitation thresholds, and seasons (Fig. 6).

The plot of the BSS against the non-exceedance probability associated with different precipitation thresholds (Fig. 7) further confirms the findings from Fig. 6. It demonstrates that for the most part the postprocessors behave similarly with respect to each other. Additionally, the trend in the BSS for the postprocessed forecasts tends to mimic the behavior of the raw ensembles. For example, the BSS values, for both the raw and postprocessed forecasts, tend to increase with the precipitation threshold in Fig. 7c while they remain relatively stable in Fig. 7b. Also, as was the case in Fig. 6, the gains in skill from postprocessing are somewhat more noticeable in the summer (Figs. 7a and 7b) than fall (Figs. 7c and 7d) and generally the gains in skill are reduced for the longer forecast lead times (e.g., day 2 in Fig. 7c and day 5 in Fig. 7d). Indeed, at a lead time of 5 days in the fall (Fig. 7d), all of the postprocessed ensembles outperform the raw ensembles for probability thresholds less than 0.9; at a probability threshold of 0.9, the raw ensembles exhibit a slightly better skill than the postprocessed ensembles. Thus, the performance of the postprocessors varies with the precipitation threshold.

2) CRPSS

The CRPSS shows that the postprocessed precipitation ensembles are overall more skillful than the raw ensembles across lead times and seasons (Fig. 8). As was the case with the BSS (Figs. 6 and 7), the relative gains in skill from postprocessing are greater in the summer (Fig. 8a) than in the fall (Fig. 8b), but the overall skill of the raw as well as postprocessed ensembles is significantly better in the fall than the summer. Furthermore, contrasting the postprocessors against each other, HCLR tends to slightly outperform BMA and BMAx. Indeed, for the fall, HCLR is the only postprocessor that shows improvements upon the raw ensembles at a forecast lead time of 5 days. The close similarities between the performance of BMA and BMAx (Figs. 6-8) indicate that the GEFSRv2 ensembles members remain indistinguishable, even at the longer lead times considered.

The CRPS can be decomposed into a reliability ($CRPS_{rel}$) and potential ($CRPS_{pot}$) component (Hersbach 2000). The $CRPS_{rel}$ measures the ability of the precipitation ensembles to generate cumulative distributions that have, on average, the correct or desired statistical properties. While the $CRPS_{pot}$ measures the CRPS that one would obtain for a perfect reliable system. The decomposition of the CRPS shows that the gains in skill from postprocessing are mainly related to improvements in $CRPS_{rel}$ (Fig. 9a). Note that the CRPS, $CRPS_{rel}$, and $CRPS_{pot}$ have a negative orientation (i.e. negative values are better). The CRPS decomposition reveals that the gains are considerably greater in the summer (Fig. 9a) than fall (Fig. 9b). It also shows that HCLR tends to have similar (Fig. 9a) or even larger (Fig. 9a) $CRPS_{rel}$ than BMA and BMAx but a smaller $CRPS_{pot}$. The reduction in $CRPS_{pot}$ is the main source of improvement for HCLR over BMA and BMAx. This means, in relation to the sampled climatology, that the resolution associated with HCLR is likely better than that of BMA and BMAx. This may be due to the fact

that HCLR uses the ensemble spread as a predictor of the dispersion of the predictive pdf (Messner et al. 2014a) and the $CRPS_{pot}$ is sensitive to the spread (Hersbach 2000). The CRPS decomposition also illustrates the fact that BMA and BMAx can improve the reliability of the forecasts relative to the raw ensembles while at the same time reducing the overall skill of the forecasts. This is observed in Fig. 9b at a forecast lead time of 5 days where BMA and BMAx have slightly lower $CRPS_{rel}$ than the raw ensembles but much higher $CRPS_{pot}$.

3) Reliability

According to the CRPS decomposition (Fig. 9), the postprocessed ensemble precipitation forecasts tend to be more reliable than the raw ensembles. This is further confirmed using reliability diagrams under various forecasting conditions (Fig. 10). In Fig. 10, the reliability of the postprocessed forecasts from BMA, BMAx, and HCLR is improved relative to the raw ensembles across forecast probabilities, lead times, and seasons. There is, however, a tendency to underforecast the small forecast probabilities in the summer (Fig. 10b) and fall (Fig. 10d), i.e. the postprocessed forecasts tend to be somewhat underconfident. This tendency is significantly more apparent in the raw ensembles than in the postprocessed ones (Fig. 10a). For the larger forecast probabilities, the raw ensembles tend to overforecast the forecast probabilities, i.e. the forecasts are overconfident, while the postprocessed ones seem, for the most part, to fix this overforecasting bias (Fig. 10c).

Contrasting the postprocessors against each other, all three postprocessors show similar reliability and sharpness (assessed by examining the insets in Fig. 10). The reliability of the postprocessors does not seem to vary greatly with the season (Figs. 10a and 10c) or forecast lead time (Figs. 10a and 10b). It does vary, however, with the precipitation threshold. The reliability curves associated with each of the postprocessors show more variability for moderate precipitation events (Fig. 11) than when considering all the precipitation events (Fig. 10). For moderate precipitation events, the raw ensembles are strongly overconfident; they overforecast the larger forecast probabilities (Figs. 11a and 11c). The overforecasting is stronger in the summer (Figs. 11a and 11b) than the fall (Figs. 11c and 11d). Nonetheless, the reliability of the different postprocessors is overall similar for moderate precipitation events. In general, forecast BMA and BMAx seem more reliable than forecasts from HCLR in some cases (Figs. 10c and 11a) while in other cases HCLR is more reliable (Figs. 10a, 10b, and 11c). Overall, the three postprocessors are able to improve the biases in the raw ensembles to make them more reliable.

5. Summary and conclusions

Ensemble forecasts can be used to determine the probability and uncertainty of a weather variable. In the case of ensemble precipitation forecasts, the determination of forecast probabilities from ensembles is generally unreliable, because the magnitude and dispersion of the ensemble forecasts are often characterized by the presence of biases (Messner 2014a,b; Sloughter et al. 2007; Wilks 2009). Statistical postprocessing is, thus, needed to correct the biases and improve the reliability of ensemble precipitation forecasts. In this study, we assessed the potential of BMA (Sloughter et al. 2007), BMAx (Fraley et al. 2010), and HCLR (Messner et al. 2014b) to postprocess precipitation ensembles from the 11-member GEFSRv2 dataset (Hamill et al. 2013). As part of our experimental setting, we employed 24-h precipitation accumulations for lead times of 24- to 120-h over the U.S. MAR. We used MPEs as the observed precipitation.

To implement BMA and BMAx, we first selected the length of the sliding time window and the number of cells needed to train the postprocessors. Using the BSS and CRPSS to assess

the skill associated with different window lengths, we found that generally the optimum value tended to be ~25 days across lead times and seasons. Similar results have been reported by others (Fraley et al. 2010; Sloughter et al. 2007). We note that the sensitivity of the skill scores to the training window length was not large. Furthermore, since we used training data from different years to train the BMA and BMAx, the effective training length is greater than 25 days. In terms of the number of cells, we found that training each cell in the GEFSRv2 separately yielded slightly more skillful forecasts than when spatially pooling data from several cells. This may be partly the case here since we sampled data from the previous four years when training the postprocessors, which potentially makes spatial pooling less effective. But relying on past forecasts to train the postprocessors may not always be feasible, particularly when dealing with operational forecasting systems.

We used the BSS, CRPSS, and reliability diagrams, conditioned upon the lead time, precipitation threshold, and season, to compare against the raw ensembles and each other the BMA, BMAx, and HCLR postprocessors. From this comparison, we found that overall there is a slight tendency for HCLR to outperform BMA and BMAx but the differences appear to be not as significant. They become more apparent at the longer forecast lead times (e.g., 5 days) during both the summer and fall. In terms of the forecast skill, the postprocessors show significant gains relative to the raw ensembles in the summer across lead times while gains are less significant in the fall. But overall the raw and postprocessed ensembles are more skillful in the fall than summer. The reliability diagrams showed that the postprocessors are able to correct biases in the raw ensembles that ultimately make the postprocessed ensembles be more reliable than the raw ones across lead times, precipitation thresholds, and seasons. The three postprocessors result in forecasts with similar reliability. Additionally, we found that differences between BMA and BMAx are small, thereby indicating that the GEFSRv2 ensemble members are indistinguishable. This is the case here since the GEFSRv2 dataset is based on a single model and single physics ensembles. In the case of multiple models or multi-physics ensembles, the relative performance of BMA might show greater improvements than the ones observed in this study.

By decomposing the CRPS into a reliability ($CRPS_{rel}$) and potential ($CRPS_{pot}$) component, we were able to examine more carefully the differences between BMA/BMAx and HCLR. From this, we observed that the improved performance of HCLR over that of BMA/BMAx is due to having a lower $CRPS_{pot}$. Indeed, the $CRPS_{rel}$ component tends to be slightly lower (better) for BMA/BMAx than HCLR. Thus, we attributed the better performance of HCLR to the fact that it uses the ensemble spread as a predictor of the dispersion of the predictive pdf since the $CRPS_{pot}$ is sensitive to the spread. We also note that, based on the decomposition of the CRPS, HCLR is the only postprocessor to consistently improve upon the raw ensembles across lead times and seasons.

In summary, based on our analysis and comparison, we found that generally the postprocessors perform similarly. An important advantage of BMA/BMAx, which we were not able to evaluate here, is to allow in a consistent manner the incorporation of ensembles members from different forecasting systems. A future alternative could be to combine the strengths of both BMA and HCLR, e.g., by using HCLR to determine the predictive pdf of each forecasting system and BMA to weight the pdfs. However, this may come at a considerable computational cost, particularly when considering a range of lead times and multiyear reforecasts datasets.

APPENDIX

Verification metrics

a. Brier Skill Score (BSS)

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

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}\left[F_{xi}(q) - F_{yi}(q)\right]^2, \tag{1}$$

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

$$F_{X_i}(q) = \mathrm{P_r}[X_i > q], \tag{2}$$

n is again the total number of forecast-observation pairs, and

$$F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In order 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):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS_{main}}}{\mathrm{BS_{reference}}}, \tag{4}$$

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

b. Reliability diagram

As suggested by Murphy (1973), the BS can be further decomposed into a reliability, resolution, and uncertainty component. In this study, instead of using the decomposed BS to quantify the reliability and resolution of the forecasts, we use the so-called reliability diagram. The reliability diagram shows the full joint distribution of forecasts and observations to reveal the reliability of the probability forecasts. For the forecast values portioned into bin B_k and defined by the exceedance of threshold q, the average forecast probability can be expressed as

$$\bar{F}_{X_k}(q) = \frac{1}{|I_k|}\sum_{I_k} F_{X_i}(q), \text{ where } I_k = \{i : \mathrm{X}_i \in B_k\}, \tag{5}$$

where I_k is the collection of all indices i for which X_i falls into bin B_k, and $|I_k|$ denotes the number of elements in I_k. The corresponding fraction of observations that fall in the K^{th} bin is given by

$$\bar{F}_{Y_k}(q) = \frac{1}{|I_k|}\sum_{I_k} F_{Y_i}(q), \text{ where } F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The reliability diagram plots $\bar{F}_{X_k}(q)$ against $\bar{F}_{Y_k}(q)$.

c. Mean Continuous Ranked Probability Skill Score (CRPSS)

The 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

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left[F_x(q) - F_y(q) \right]^2 dq. \quad (7)$$

To evaluate the skill of the main forecasting system relative to the reference forecast system, the associated skill score, the Mean Continuous Ranked Probability Skill Score (CRPSS), is defined as:

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\mathrm{main}}}{\mathrm{CRPS}_{\mathrm{reference}}}, \quad (8)$$

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

Additionally, to further explore the effect of postprocessing on forecast skill, we separate the $\mathrm{CRPS}_{\mathrm{main}}$ into different components according to the procedure developed by Hersbach (2000). Specifically, we consider the CRPS reliability ($\mathrm{CRPS}_{\mathrm{rel}}$) and potential ($\mathrm{CRPS}_{\mathrm{pot}}$) such that

$$\mathrm{CRPS}_{\mathrm{main}} = \mathrm{CRPS}_{\mathrm{rel}} + \mathrm{CRPS}_{\mathrm{pot}}. \quad (9)$$

The $\mathrm{CRPS}_{\mathrm{rel}}$ measures the ability of the precipitation ensembles to generate cumulative distributions that have, on average, the correct or desired statistical properties. The reliability is closely connected to the rank histogram, which shows whether the frequency that the verifying analysis was found in a given bin is equal for all bins (Hersbach 2000). The $\mathrm{CRPS}_{\mathrm{pot}}$ measures the CRPS that one would obtain for a perfect reliable system. It is sensitive to the average spread of the ensemble and outliers. For instance, the narrower the spread of the ensemble is, the smaller the $\mathrm{CRPS}_{\mathrm{pot}}$ becomes. As indicated by Hersbach (2000), provided a certain degree of unpredictability, a balance between the ensemble spread and the statistics of outliers will result in the optimal value of the $\mathrm{CRPS}_{\mathrm{pot}}$.

References

Bjørnar Bremnes, J., 2004: Probabilistic Forecasts of Precipitation in Terms of Quantiles Using NWP Model Output. *Monthly Weather Review*, **132,** 338-347.

Brown, J. D., and D.-J. Seo, 2010: A Nonparametric Postprocessor for Bias Correction of Hydrometeorological and Hydrologic Ensemble Forecasts. *Journal of Hydrometeorology*, **11,** 642-665.

Brown, J. D., J. Demargne, D.-J. Seo, and Y. Liu, 2010: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environmental Modelling & Software*, **25,** 854-872.

Bröcker, J., and L. A. Smith, 2008: From ensemble forecasts to predictive distribution functions. *Tellus A*, **60**, 663-678.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Monthly Weather Review*, **133,** 1076-1097.

Clark, M. P., and L. E. Hay, 2004: Use of Medium-Range Numerical Weather Prediction Model Output to Produce Forecasts of Streamflow. *Journal of Hydrometeorology*, **5,** 15-32.

Erickson, M. J., B. A. Colle, and J. J. Charney, 2012: Impact of Bias-Correction Type and Conditional Training on Bayesian Model Averaging over the Northeast United States. *Weather and Forecasting*, **27,** 1449-1469.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review*, **138,** 190-202.

Friederichs, P., and A. Hense, 2007: Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. *Monthly Weather Review*, **135,** 2365-2378.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, **11,** 1203-1211.

Greene, E., A., A. E. LaMotte, and K.-A. Cullinan, 2005: Ground-Water Vulnerability to Nitrate Contamination at Multiple Thresholds in the Mid-Atlantic Region Using Spatial Probability Models. Scientific Investigations Report 2004-5118, US Department of the Interior, USGS, 24.

Greybush, S. J., S. E. Haupt, and G. S. Young, 2008: The Regime Dependence of Optimally Weighted Ensemble Model Consensus Forecasts of Surface Temperature. *Weather and Forecasting*, **23,** 1146-1161.

Grumm, R., and R. Holmes, 2007: Patterns of heavy rainfall in the mid-Atlantic region. Preprints, *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction*, Park City, UT, Amer. Meteor. Soc., 5A.2.

Hamill, T., G. Bates, J. Whitaker, D. Murray, M. Fiorino, and T. Galarneau, 2015: A description of the 2nd-generation NOAA global ensemble reforecast data set. NOAA Earth System Research Lab.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, **132,** 1434-1447.

Hamill, T. M., and Coauthors, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, **94,** 1553-1565.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15,** 559-570.

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and temporal characteristics of heavy hourly rainfall in the United States. *Monthly Weather Review*, **141,** 4564-4575.

Lawrence, B. A., M. I. Shebsovich, M. J. Glaudemans, and P. S. Tilles, 2003: Enhancing precipitation estimation capabilities at National Weather Service field offices using multi-sensor precipitation data mosaics. Preprints, *19th Conf. on Interactive Information Processing Sys-tems for Meteorology, Oceanography, and Hydrology,* Long Beach, CA, Amer. Meteor. Soc., 15.1.

Maddox, R. A., C. F. Chappell, and L. R. Hoxit, 1979: Synoptic and meso-α scale aspects of flash flood events. *Bulletin of the American Meteorological Society*, **60,** 115-123.

Mendoza, P. A., B. Rajagopalan, M. P. Clark, K. Ikeda, and R. M. Rasmussen, 2014: Statistical Postprocessing of High-Resolution Regional Climate Model Output. *Monthly Weather Review*, **143,** 1533-1553.

Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2014a: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review*, **142,** 448-456.

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014b: Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored. *Monthly Weather Review*, **142,** 3003-3014.

Neff, R., H. Chang, C. G. Knight, R. G. Najjar, B. Yarnal, and H. A. Walker, 2000: Impact of climate variation and change on Mid-Atlantic Region hydrology and water resources. *Climate Research*, **14,** 207-218.

Polsky, C., J. Allard, N. Currit, R. Crane, and B. Yarnal, 2000: The Mid-Atlantic Region and its climate: past, present, and future. *Climate Research*, **14,** 161-173.

Prat, O., and B. Nelson, 2015: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge datasets (2002-2012). *Hydrology and Earth System Sciences Discussions*, **19,** 2037-2056.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133,** 1155-1174.

Roulin, E., and S. Vannitsem, 2011: Postprocessing of Ensemble Precipitation Predictions with Extended Logistic Regression Based on Hindcasts. *Monthly Weather Review*, **140,** 874-888.

Roulston, M., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus A*, **55,** 16-30.

Schmeits, M. J., and K. J. Kok, 2010: A comparison between raw ensemble output,(modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Monthly Weather Review*, **138,** 4199-4211.

Siddique, R., A. Mejia, J. Brown, S. Reed, and P. Ahnert, 2015: Verification of precipitation forecasts from two numerical weather prediction models in the Middle Atlantic Region of the USA: A precursory analysis to hydrologic forecasting. *Journal of Hydrology*, **529, Part 3,** 1390-1406.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, **135,** 3209-3220.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. Wiley, 137-163.

Tracton, M. S., and E. Kalnay, 1993: Operational Ensemble Prediction at the National Meteorological Center: Practical Aspects. *Weather and Forecasting*, **8,** 379-398.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, **131,** 965-986.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A*, **60,** 62-79.

Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, **13,** 243-256.

——, 2006b: *Statistical Methods in the Atmospheric Sciences.* Elsevier Academic Press, 627 pp.

——, 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, **16,** 361-368.

——, 2010: Use of stochastic weathergenerators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, **1,** 898-907.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Review*, **135,** 2379-2390.

Wu, L., D.-J. Seo, J. Demargne, J. D. Brown, S. Cong, and J. Schaake, 2011: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *Journal of Hydrology*, **399,** 281-298.

Zhu, J., F. Kong, L. Ran, and H. Lei, 2015: Bayesian Model Averaging with Stratified Sampling for Probabilistic Quantitative Precipitation Forecasting in Northern China during Summer 2010. *Monthly Weather Review*, **143,** 3628-3641.

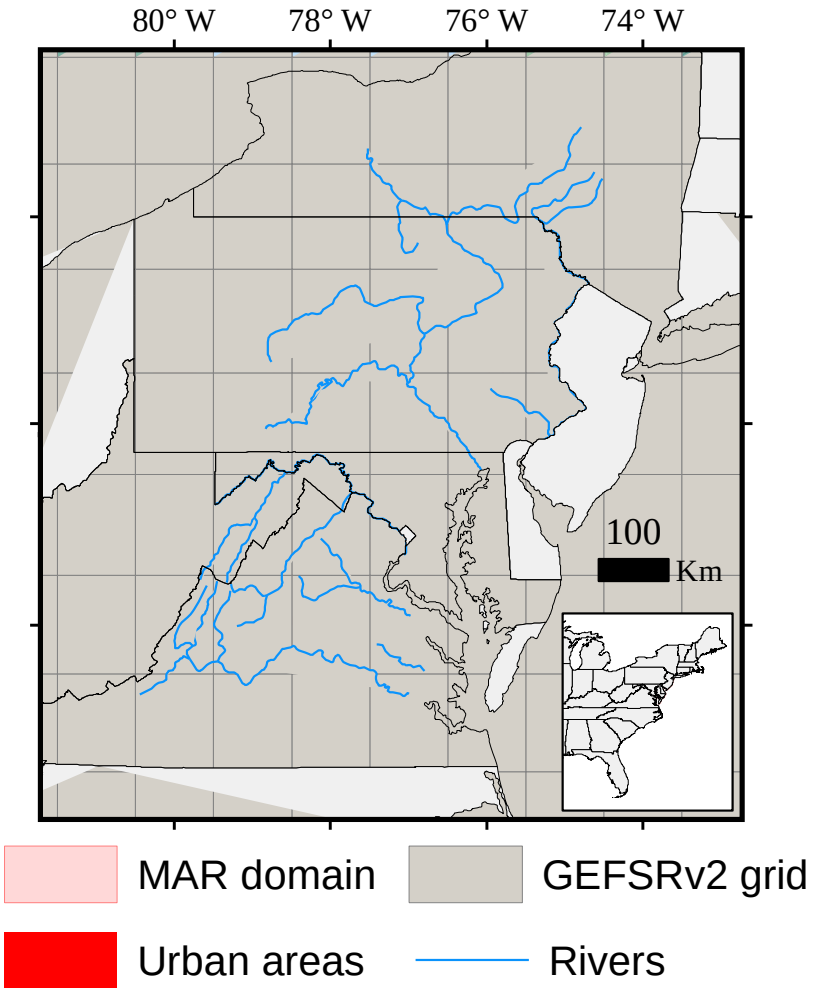


Figure 1. Map illustrating the geographic domain of the MAR in the U.S. The map also shows the major rivers, urban areas, and the GEFSRv2 grid. The inset illustrates the location of the MAR within the eastern portion of the U.S.

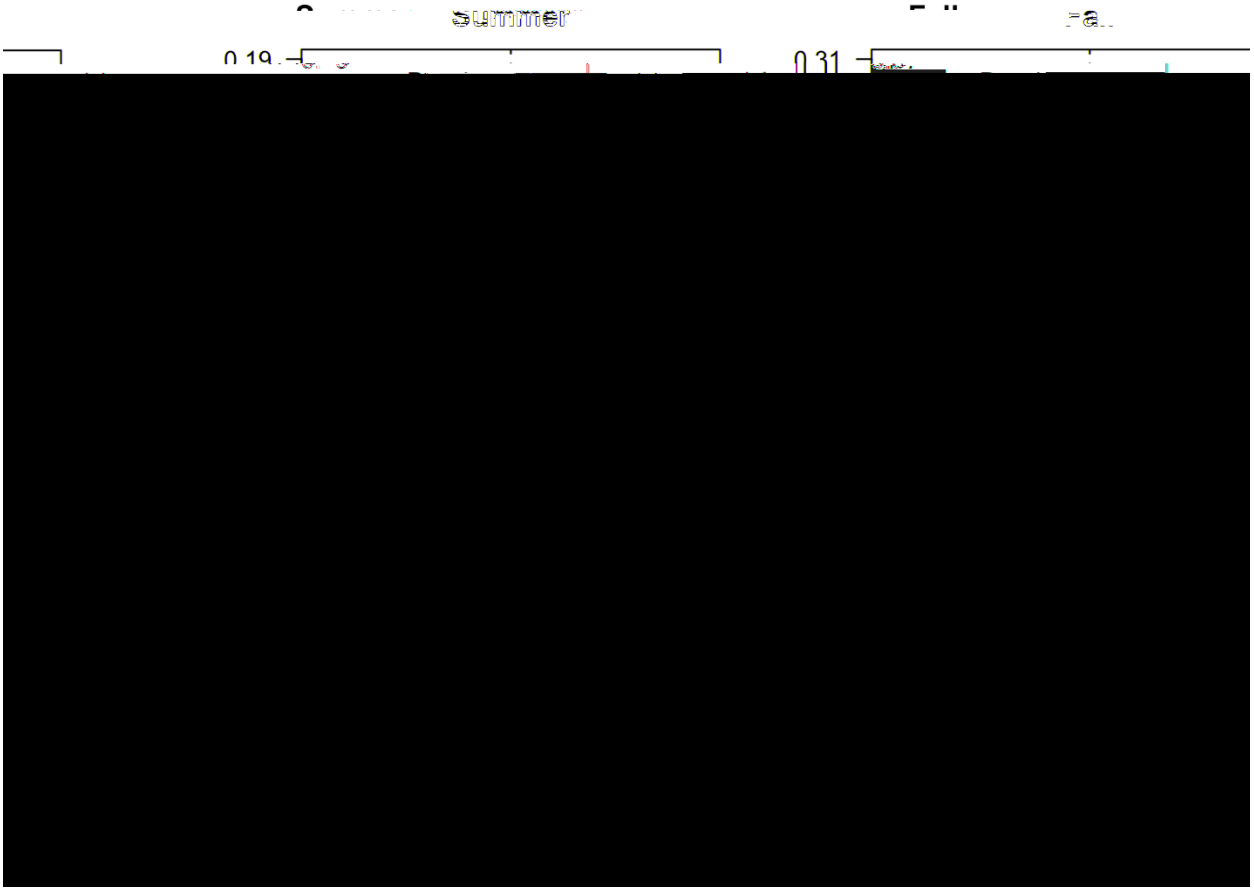


Figure 2. BSS for all the precipitation events (>0 mm) versus the BMAx training length for forecast lead times of (a) 1 and (b) 5 days during the summer and lead times of (c) 1 and (d) 5 days during the fall. The different BSS curves represent the number of cells used to train the BMAx.

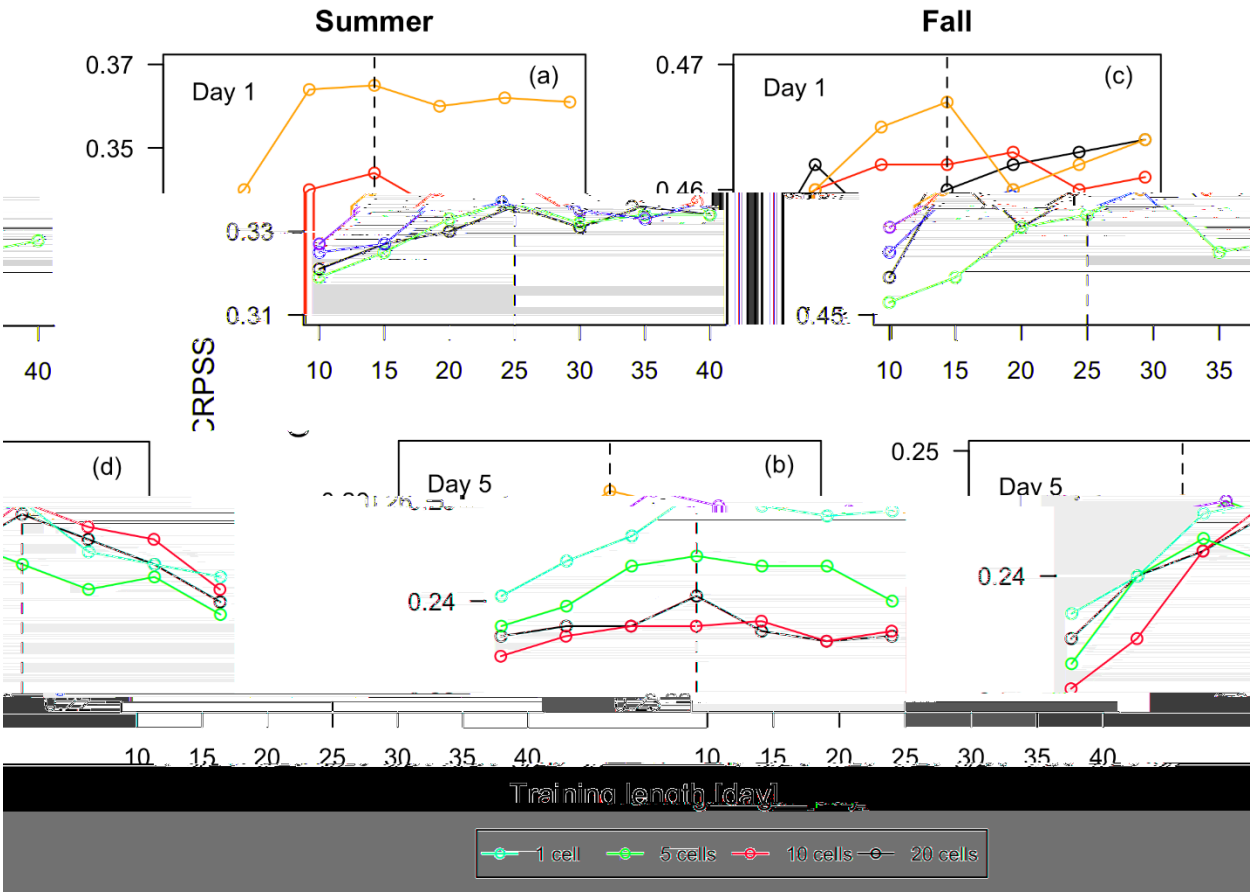


Figure 3. CRPSS versus the BMAx training length for forecast lead times of (a) 1 and (b) 5 days during the summer and lead times of (c) 1 and (d) 5 days during the fall. The different CRPSS curves represent the number of cells used to train the BMAx.

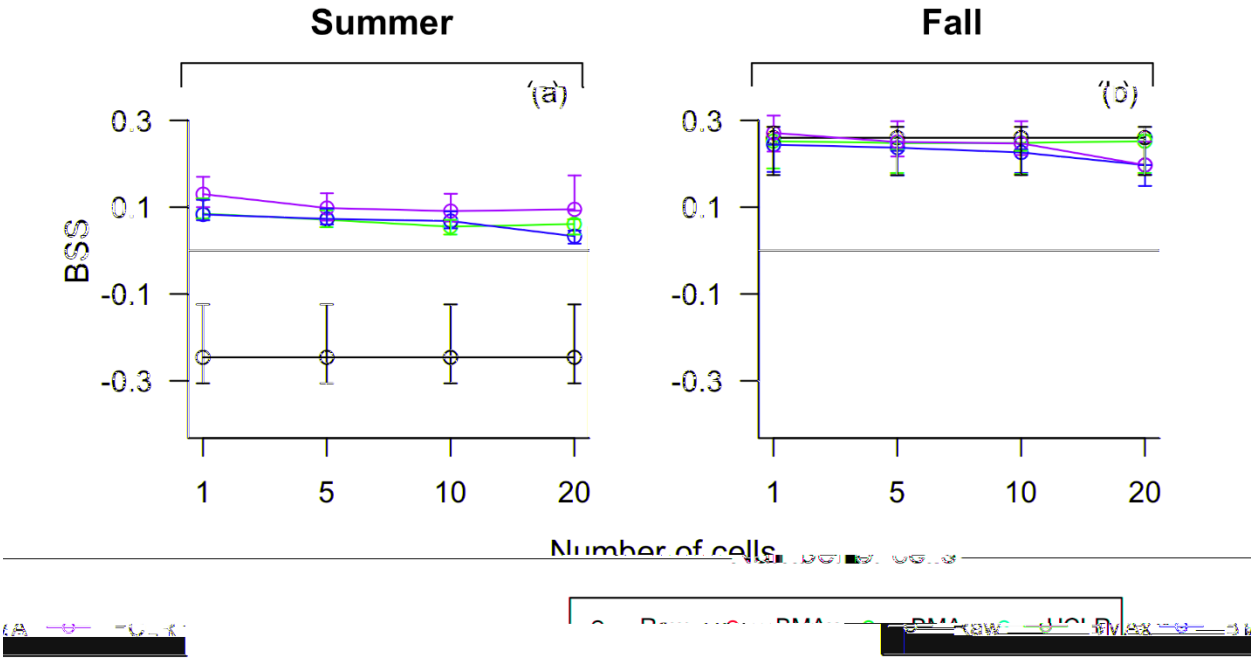


Figure 4. BSS for moderate precipitation events (>10 mm) versus the number of cells used to train the postprocessors during the (a) summer and (b) fall. The different BSS curves represent the raw and postprocessed precipitation ensembles. The figure is for a forecast lead time of 4 days.

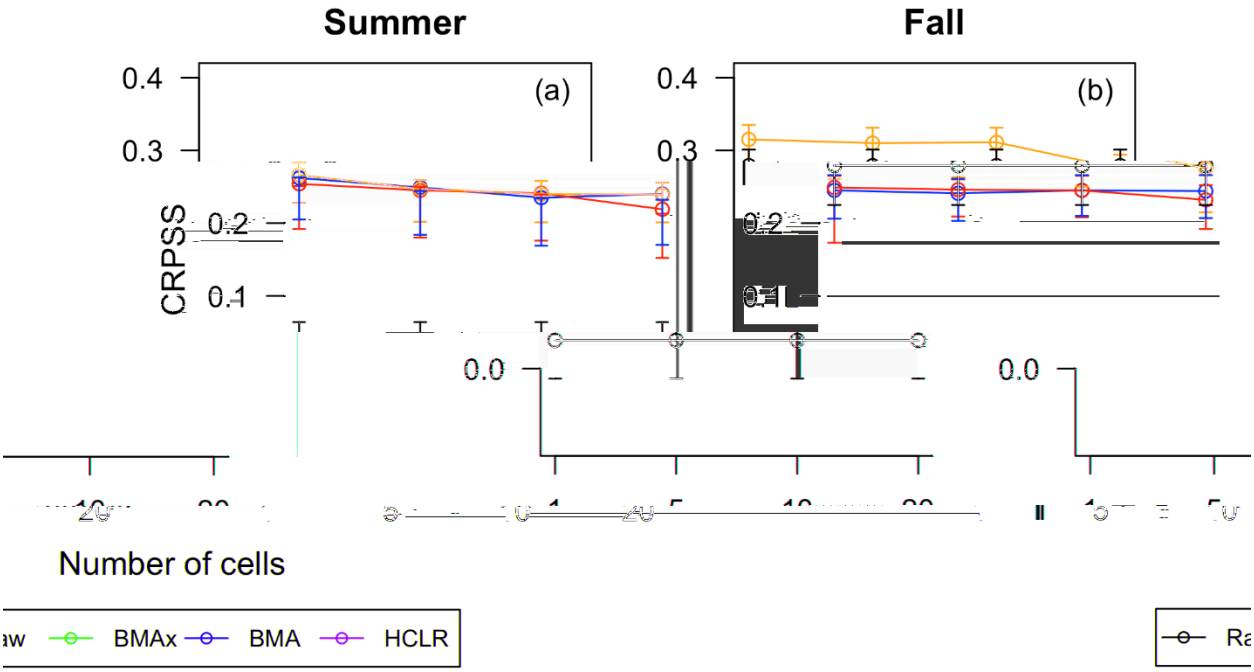


Figure 5. CRPSS versus the number of cells used to train the postprocessors during the (a) summer and (b) fall. The different CRPSS curves represent the raw and postprocessed precipitation ensembles. The figure is for a forecast lead time of 5 days.

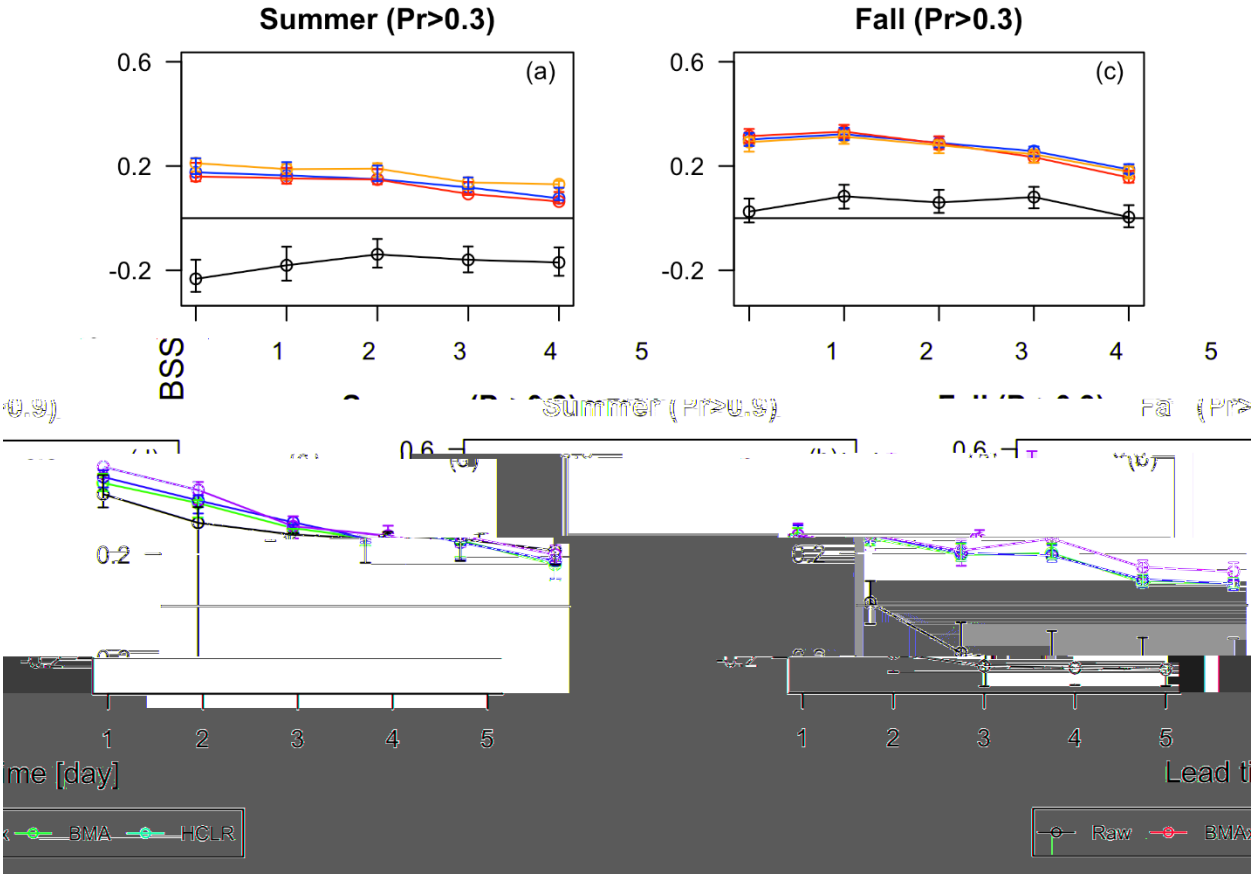


Figure 6. BSS for (a) all (>0 mm) and (b) moderate (>10 mm) precipitation events during the summer versus the forecast lead time. BSS for (c) all (>0 mm) and (d) moderate (>10 mm) precipitation events during the fall versus the forecast lead time. The different BSS curves represent the raw and postprocessed precipitation ensembles.

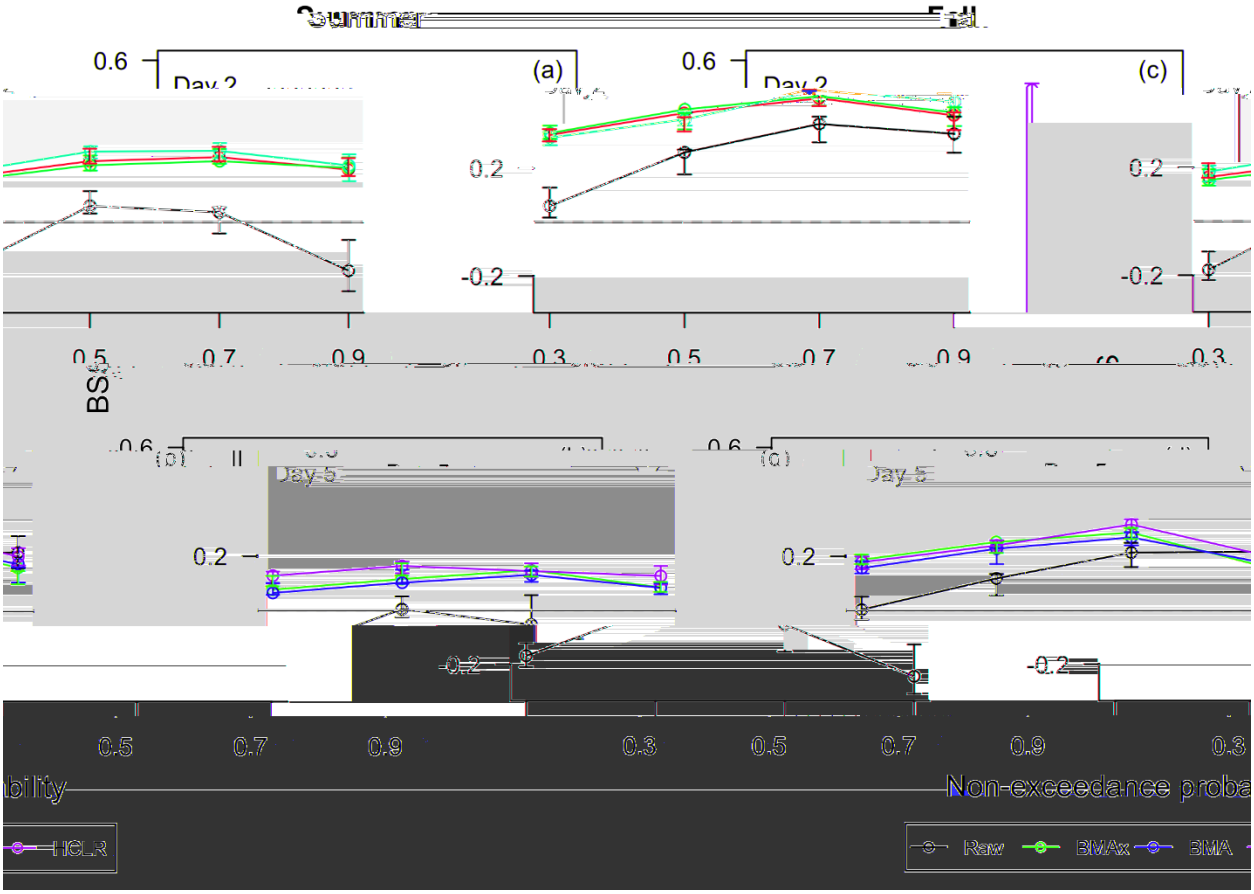


Figure 7. BSS versus the precipitation threshold for forecast lead times of (a) 2 and (b) 5 days during the summer and forecast lead times of (c) 2 and (d) 5 days during the fall. The different BSS curves represent the raw and postprocessed precipitation ensembles.

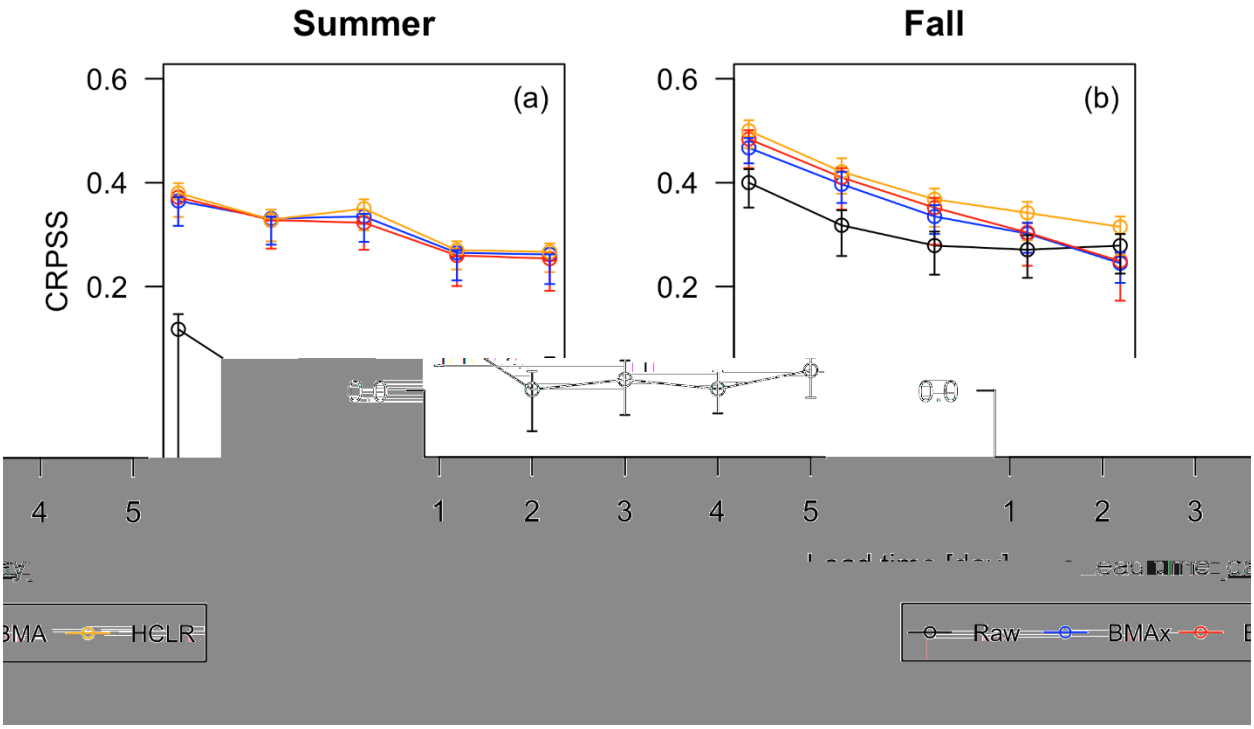


Figure 8. CRPSS for the ensemble precipitation forecasts versus the forecast lead time during the (a) summer and (b) fall. The different CRPSS curves represent the raw and postprocessed precipitation ensembles.

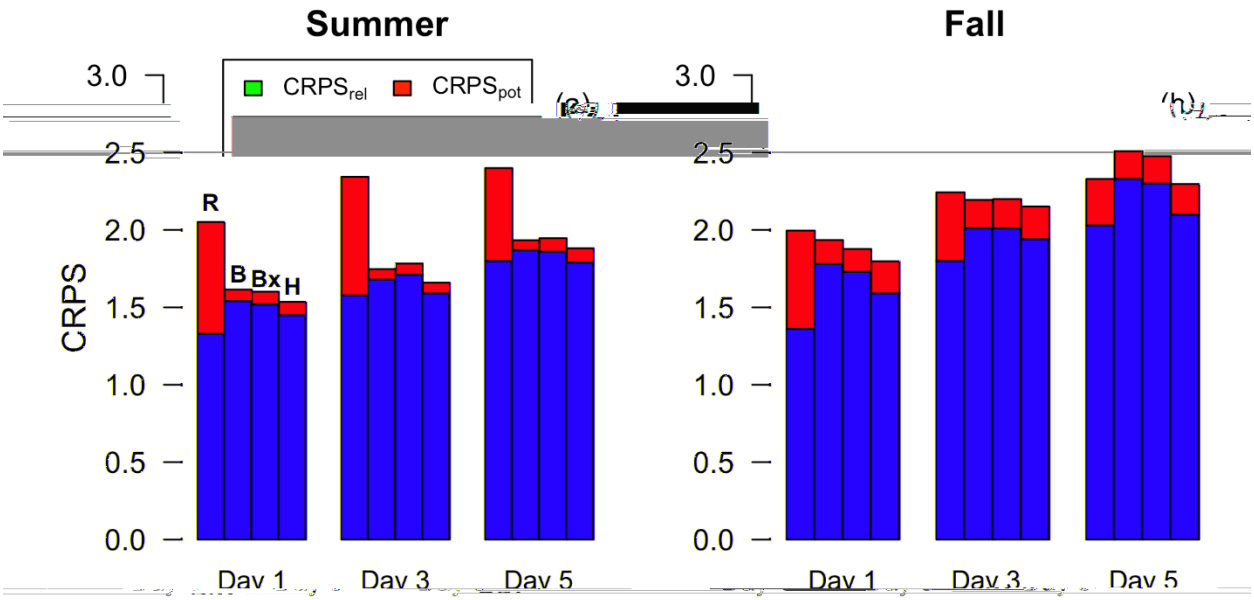


Figure 9. Decomposition of the CRPS into CRPS reliability ($CRPS_{rel}$) and CRPS potential ($CRPS_{pot}$) for forecasts lead times of 1, 3, and 5 days during the (a) summer and (b) fall. The four columns associated with each forecast lead time represent, from left to right, the raw (R), BMA postprocessed (B), BMAx postprocessed (Bx), and HCLR postprocessed (H) precipitation ensembles.

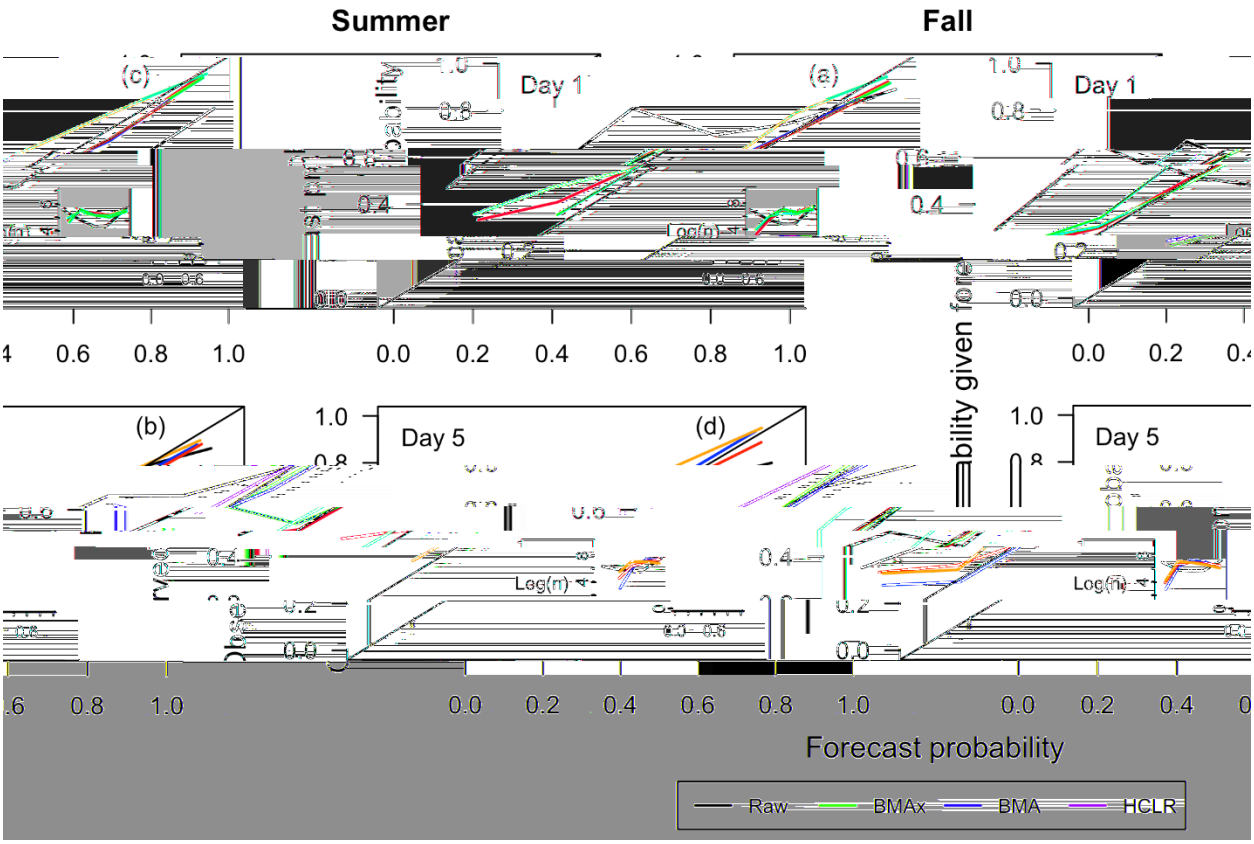


Figure 10. Reliability diagrams for all the summer precipitation events and forecast lead times of (a) 1 and (b) 5 days. Reliability diagrams for all the fall precipitation events and forecast lead times of (c) 1 and (d) 5 days. The different reliability curves represent the raw and postprocessed precipitation ensembles. The insets show the sample size in logarithmic scale of the different forecast probability bins.

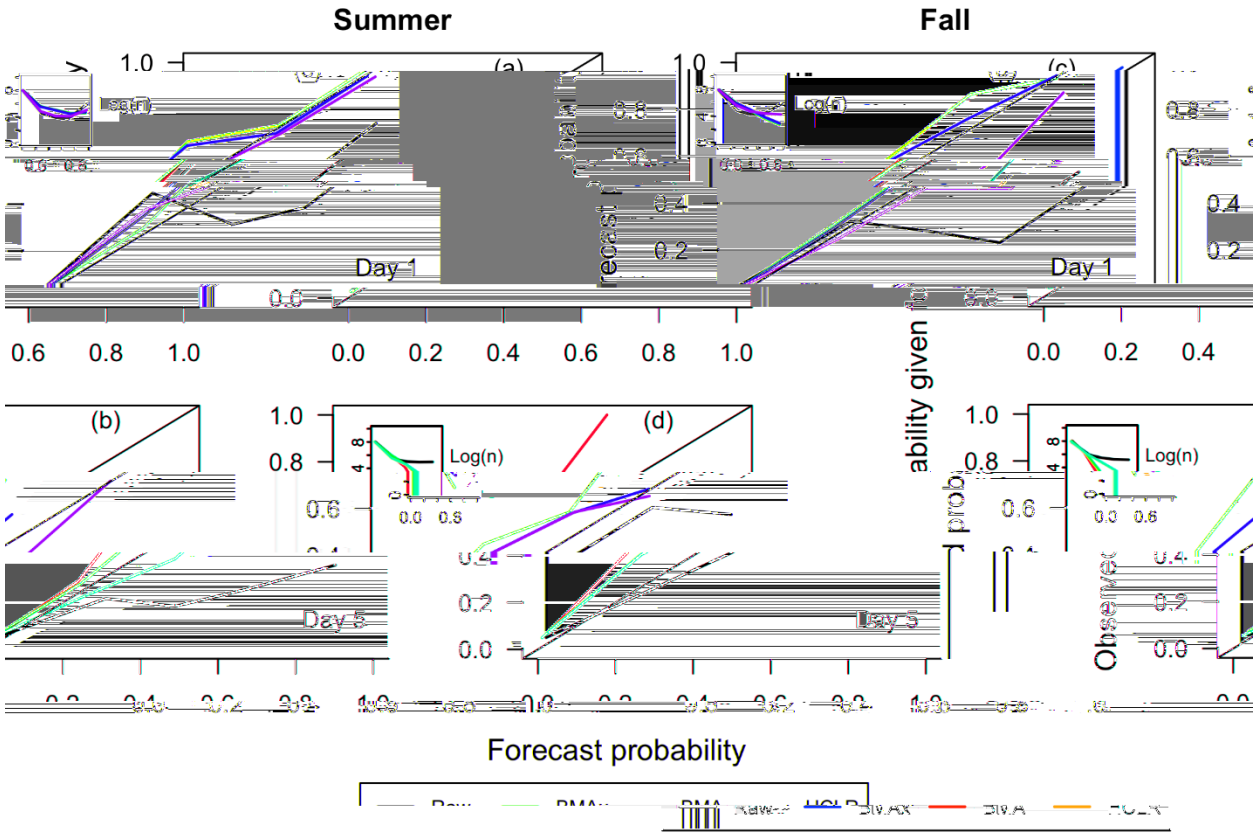


Figure 11. Reliability diagrams for moderate precipitation events (>10 mm) during the summer and forecast lead times of (a) 1 and (b) 5 days. Reliability diagrams for moderate precipitation events (>10 mm) during the fall and forecast lead times of (c) 1 and (d) 5 days. The different reliability curves represent the raw and postprocessed precipitation ensembles. The insets show the sample size in logarithmic scale of the different forecast probability bins.

Chapter 5: Ensemble streamflow forecasting across the U.S. middle Atlantic region with a distributed hydrological model forced by GEFS reforecasts

ABSTRACT

The quality of ensemble streamflow forecasts in the U.S. middle Atlantic region (MAR) is investigated for short- to medium-range forecast lead times (6-168 h). To this end, a regional hydrologic ensemble prediction system (RHEPS) is assembled and implemented. The RHEPS is comprised, in this case, by 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's (NOAA's) 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 auto-regressive model as the postprocessor. To verify streamflow forecasts from the RHEPS, 8 river basins in the MAR are selected, ranging in drainage area from ~262 to 29,965 km^2 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; Vörösmarty 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 hydrologic 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. 2014a; Khan et al. 2014; 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. 2010; Anghileri et al. 2016; Bartholmes et al. 2009; Bennett et al. 2014; Boucher et al. 2011; Brown et al. 2014; Franz et al. 2008; Georgakakos et al. 2014; Harshburger et al. 2012; Schellekens et al. 2011; Van Cooten et al. 2011; Verbunt et al. 2007; Wood et al. 2015).

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; Schaake et al. 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- (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 U.S., the National Oceanic Atmospheric Administration's 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, particularly for medium-range streamflow forecasts.

A key component of a 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 semi-distributed hydrological models which do not account, or only in a limited fashion, for the spatial variability of inputs (e.g., meteorological, topographical, pedological, land-cover, etc.), parameters, and variables (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; ene Michaud and Sorooshian 1994; Krajewski et al. 1991; Smith et al. 2012; Spies et al. 2014). 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 few applications have been developed and comprehensively evaluated that apply a distributed hydrological model within a HEPS (Alfieri et al. 2014; Anghileri et al. 2016; Fan et al. 2014b; Georgakakos et al. 2014; 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.

Furthermore, streamflow forecasts generated from meteorological ensembles tend to exhibit systematic biases which makes the determination of forecast probabilities from such streamflow data unreliable (Roulin 2006; 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 a HEPS that must be considered when verifying the quality of 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 comprised 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 (hereafter the RHEPS) to improve short- to medium-range streamflow forecasts in the U.S. middle Atlantic region (MAR). The objective is also to quantify the relative importance of different sources of uncertainty (meteorological and hydrological) in streamflow forecasts. To meet these objectives, we assemble and implement the RHEPS, which is comprised here by the NOAA's Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM), forced by ensemble precipitation and near surface temperature outputs from the NCEP GEFSRv2 (Hamill et al. 2013; Siddique et al. 2015). Specifically, we use the RHEPS in this study to produce and verify ensemble streamflow forecasts for lead times from 6 to 168 hours across eight river basins of varying spatial scales in the MAR. 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 and discussed in section 4. Lastly, in section 5, we outline the key conclusions.

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 U.S., 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, 8 river basins are selected (Fig. 1), ranging in drainage area from ~262 to 29,965 km^2 and covering the major rivers in the MAR, including the Delaware, James, Potomac, and Susquehanna River. 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 of 17,574 km^2 while its nested subbasin is only 860 km^2 (Table 1). All of the selected basins are gauged by the United States 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 reforecasts (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 rests are perturbed members generated with perturbed initial conditions. For days +1 to days +8, it uses the T254L42 model resolution which runs at a spatial grid of 0.5^0 or ~55 km. From days +9 to +16, the resolution is changed to T190L42 which runs at a 0.67^0 resolution or ~73 km. Each day the model is initiated at 00 UTC to produce reforecasts for the next 16 days of lead time. For days +1 to days +3, 3 hourly forecast accumulations are available, after that forecasts are saved every 6 hours, providing 6 hourly accumulations of forecasts for days +4 to +16. In total, more than 29 years of reforecast data is archived (1984- present) for a large number of selected meteorological variables. Further details about the GEFSRv2 are discussed by Hamill et al. (2013).

2) Observations

We use multi-sensor 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 x 4 km^2 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 the 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. 2016; 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 x 4 km^2 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, ten years (2004-2013) of streamflow observations were used.

3. Methods

In this study, the RHEPS is comprised by the following four main components: (i) meteorological forecasts (precipitation and near surface temperature ensembles), (ii) distributed hydrological model, (iii) statistical postprocessor, and (iv) verification strategy. This subsection describes the latter 3 components since the meteorological ensembles were described in the previous section.

a. Distributed hydrological model

The 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. 2014; Thorstensen et al. 2015; Wood et al. 2015) 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. 2007). Here we

run HL-RDHM in a fully distributed manner at a spatial resolution of 2 x 2 km^2, using kinematic wave routing to route, across the river network, the runoff generated at each grid cell by SAC-HT and SNOW-17 (Koren et al. 2004; Smith et al. 2012).

To run HL-RHDM, we work with both uncalibrated and calibrated parameter runs (Smith et al. 2012). The uncalibrated model runs are based on a-priori parameter estimates from available datasets (Anderson et al. 2006; Koren et al. 2000; Reed et al. 2004). For the calibrated model runs, we select for calibration 10 out of the 17 SAC-HT parameters based upon prior experience and preliminary parameter sensitivity tests. To calibrate the selected HL-RDHM parameters, we first adjust manually the a-priori parameter fields; 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). We use SLS since this method is readily available within HL-RDHM and have been shown to provide reliable parameter estimates (Kuzmin et al. 2008; Kuzmin 2009). We use 3 and 1 year of streamflow data to calibrate the small and large basins, respectively. We use a short calibration period to ameliorate computational demand. To assess the model performance during calibration, we use the correlation coefficient (R), modified correlation coefficient (R_m), percent bias (PB), and Nash-Sutcliffe efficiency (NSE) (see appendix for their mathematical definition).

b. Statistical postprocessor

To statistically postprocess the ensemble streamflow forecasts (i.e. to quantify the streamflow uncertainty and adjust forecast biases), we implement the so-called hydrological model output statistics (HMOS) approach (Regonda et al. 2013). Similar approaches are common and widely used in weather forecasting (Glahn and Lowry 1972; Hamill et al. 2004; Wilks 2015). 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:

$$Z_{k+1}^0 = (1-b_{k+1})Z_k^0 + b_{k+1}Z_{k+1}^f + E_{k+1}, \quad (1)$$

where Z_k^0 and Z_{k+1}^0 denote the normalized observed flow at times k and $k+1$, Z_{k+1}^f 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^0, E_{k+1} can be calculated as follows:

$$E_{k+1} = \frac{\sigma_{E_{k+1}}}{\sigma_{E_k}} \rho(E_{k+1}, E_k) E_k + W_{k+1}, \quad (2)$$

where σ_{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 $¥(0, \sigma_{W_{k+1}}^2)$. To estimate the parameter $\sigma_{W_{k+1}}^2$, we use the following:

$$\sigma_{W_{k+1}}^2 = [1-\rho^2(E_{k+1}, E_k)]\sigma_{E_{k+1}}^2 \quad (3)$$

The step-by-step procedure for implementing the postprocessor (eqs. 1-3) 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 (NQT) (Krzysztofowicz 1997). In this study, eight years (2004-2011) of forecast and observation data are used as the training period.

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

iii) For each b_{k+1}, $\sigma^2_{W_{k+1}}$ is calculated from eq. (3) using the training data to estimate the parameters in eq. (3).

iv) W_{k+1} is generated from $¥(0, \sigma^2_{W_{k+1}})$ and E_{k+1} is calculated from eq. (2).

v) A trace of Z^0_{k+1} from eq. (1) is generated and transformed back to real space using the inverse NQT.

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), we use 1 day-ahead predictions as the observed streamflows. For the cases where Z^0_{k+1} falls beyond the historical maxima or minima, we use extrapolation to model the tails of the forecast distribution. For the upper tail (high flows), we use a hyperbolic distribution (Journel and Huijbregts 1978) while linear extrapolations is 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. 2016; Siddique et al. 2015). We use for the verification 6-hourly streamflow forecasts and daily observed streamflows at the outlet of each of the 8 selected basins. The verification is done conditioned upon the lead time, streamflow threshold, and season. We use the relative mean error (RME), Brier Skill Score (BSS), and Continuous Ranked Probability Skill Score (CRPSS) 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, we use ten years of data (2004-2013). To verify the postprocessed ensemble forecasts, we use two years of data (2012-2013) while the remaining years (2004-2011) are used to train the postprocessor. 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, contain only 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.

4. Results and discussion

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-moderate and high flows to verify the performance of the RHEPS under these two different flow conditions. The low-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).

a. Verification of the simulated streamflows

The main results associated with the performance of the simulated streamflows for the entire period of analysis (2004-2013) and the 8 selected basins are summarized in Table 2. Further, the results in Table 2 are based on the performance metrics used in calibration (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-moderate and high flow conditions. Note that the simulated, as opposed to forecasted, streamflows are obtained by forcing the HL-RDHM model with observed precipitation and near surface temperature data.

1) Low-moderate flows

The correlation coefficient R between the simulated and observed low-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 8 basins, between the uncalibrated and calibrated simulation runs is ~7%, but it can be as high as 40% in the case of SHBN6. The modified correlation coefficient R_m is also computed since it is better than R in 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 of SHBN6, the upper zone free water storage (UZFWM) and the percolation exponent (REXP) were found to be among the most influential of all the parameters, suggesting that the performance of SHBN6 is in this case particularly dependent on interflow conditions as might be expected for a karst basin. 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 a-priori model parameters were nearly optimum for this basin. In one of the 8 basins, WVYN6, the NSE decreases slightly after calibration for the low-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-moderate flows in WVYN6 is likely due to trade-offs in some of the parameter values. Indeed, for the 8 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, PB, 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-moderate flows, the high flows tend to outperform the low-moderate flows, but in a few basins (WALN6, SHBN6, and RMDV2 in Table 2) the low-moderate flows perform better. To further understand this, we examined the simulation runs (hydrographs) and noticed that some of the high flow events, mostly during the winter months (Nov-April), are somewhat underestimated. Thus, incorporating additional data when implementing the SNOW-17 mode 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-moderate flows. Further, as was the case with the low-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). Moreover, the 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 model performance statistics shown in the past (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 model with the GEFS precipitation and near surface temperature reforecasts. To verify the raw ensemble streamflow forecasts, we use the RME, CRPSS, and BSS (see appendix) as the verification metrics for the period 2004-2013. For each of the three verification metrics used, we compute two different version of the metric, one using observed flows as the reference and another one using simulated flows. The former captures the influence of the total (meteorological and hydrological) uncertainty on the streamflow forecasts and the latter emphasizes the influence of meteorological uncertainty alone. The difference between the two versions of the same metric is used to assess the relative influence of hydrological uncertainty on the streamflow forecasts.

1) Low-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) which 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-moderate flow forecasts is for the most part small as compared to the effect of the meteorological uncertainty.

To measure the skill of the raw ensemble flow forecasts, we use the CRPSS (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-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 is low for DAWM2 because this is a small basin and meteorological forecast skill tends to decrease considerably with decreasing spatial scale or basin size (Li et al. 2009; Siddique et al. 2015). Further, by comparing the two versions of the CRPSS metric (i.e. the solid line against the dashed line in Fig. 3), we find that the 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). 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 towards each other at the 7 day lead time (see, e.g., Figs. 3g and 3h).

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 8 selected basins, under low-moderate flow conditions (Fig. 4). The BSS is computed from the Brier Score (BS) which is analogous to the mean squared error of the forecasts (see appendix). A BSS score of one implies perfect skill and a BSS score of zero 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 km^2, out of the 8 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), in some basins is much smaller; 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-moderate flows. For instance, in Fig. 5f, the high flows are almost unbiased at lead times of 1-2 days when accounting for meteorological uncertainty alone (i.e. RME relative to simulated flows), but the negative bias jumps to ~18-20% when the total uncertainty (i.e. RME relative to observed flows) is considered. Thus, in this case, 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-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-moderate flows are in this case dominated by subsurface processes and only indirectly by the precipitation events.

As expected, the high flow forecasts are less skillful as the forecast lead time increases (Fig. 6). As was the case with the low-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 0.4-0.6 at a lead time of 1 day and between 0.1-0.3 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 basins. 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) 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 and 6f), 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 latter is particularly relevant in the Susquehanna River basin due to its complex geological conditions, which complicates the estimation of reliable subsurface parameters.

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 more skillful than in the dry one, except for SHBN6 and PYAV2 (Figs. 7b and 7d, respectively) 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, in this case, to be greater in the dry season than in the wet one across basin sizes in the MAR region. A similar phenomenon was reported by Li et al. (2009) for seasonal hydrological forecasts in the eastern U.S. They suggested that this is due to having a larger uncertainty in the model initial conditions (e.g., soil moisture states) during the dry season than the wet one.

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 as the lead times increases. This is the case for both low-moderate and high flows (Fig. 8). For instance, the CRPSS for both the low-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 day 1 lead time but rises by ~20% at the day 7 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 deterministic forecasts, from the day 1 to the day 7 lead time, is ~10-20% and ~15-40% for the low-moderate and high flows, respectively. It is also interesting 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-moderate flows

To verify the postprocessed ensemble forecasts for the low-moderate flow conditions, the RME is plotted for both the postprocessed and raw (without postprocessing) ensemble mean (Fig 9). For the low-moderate flows, the RME indicates that postprocessing tends to reduce the forecast bias in some of the selected basins. For example, SHBN6 (Fig. 9b), PYAV2 (Fig. 9d), BRKM2 (Fig. 9e), and RMDV2 (Fig. 9h) demonstrate improved RME values relative to the raw forecast values. The most noticeable improvements are seen in SHBN6 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), WVYN6 (Fig. 9f), and TREN4 (Fig. 9g). 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, e.g., the basins of the Delaware River, WALN6 and TREN4. In these basins, it may be necessary to account for low-flow regulations to ultimately 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-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 35% (BRKM2 at a lead time of 1 day, Fig. 10e).

2) High flows

Postprocessing is more effective for the high flows than the low-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 and 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 drops to 20% 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 exception of DAWM2 (Fig. 10a), which reveals little to no gains from postprocessing. Postprocessing demonstrates, overall, significant potential for improving flow forecasts.

5. Summary and conclusions

In this study, we generated, using the RHEPS, 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-year period (2004-2013) 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 ensemble forecast mean mostly underestimates the observed and simulated mean across lead times of 1 to 7 days, under both low-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 (day 1) 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 (>4 days), hydrological uncertainty becomes less pronounced and meteorological uncertainty dominates. This trend is apparent in both the low-moderate and high flows.
- The raw ensemble streamflow forecasts exhibit seasonal behavior across all the basins in the MAR, with forecasts having slightly better skill in the wet season compared to the dry one. Overall, hydrological uncertainty seems to have a greater impact on the streamflow 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 ensembles. After postprocessing, the skill of the streamflow forecasts for high flow conditions are improved across the entire 7-day forecast cycle. The improvements in low-moderate forecasts are mainly seen across the short-range lead times (<4 days).

APPENDIX

Correlation coefficient (R):

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

$$\mathrm{R} = \frac{N\sum_{i=1}^{N} S_i Y_i - \sum_{i=1}^{N} S_i \sum_{i=1}^{N} Y_i}{\left[N\sum_{i=1}^{N} S_i^2 - \left(\sum_{i=1}^{N} S_i\right)^2\right]^{1/2}\left[N\sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2\right]^{1/2}}, \quad (4)$$

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.

Modified correlation coefficient (R_m):

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 flow is introduced to refine the conventional correlation coefficient R. The modified correlation coefficient $\mathrm{R_m}$ is defined as:

$$\mathrm{R_m} = \mathrm{R}\frac{\min\{\sigma_{sim}, \sigma_{obs}\}}{\max\{\sigma_{sim}, \sigma_{obs}\}}, \quad (5)$$

where σ_{sim} and σ_{obs} denote the standard deviation of the simulated and observed flows, respectively.

Percent bias (PB):

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

$$\mathrm{PB} = \frac{\sum_{i=1}^{N}(Y_i - S_i)}{\sum_{i=1}^{N} Y_i} \times 100, \quad (6)$$

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

Nash-Sutcliffe efficiency (NSE):

The 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 NSE can vary between negative infinity to 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:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(S_i - Y_i)^2}{\sum_{i=1}^{N}(Y_i - \overline{Y_i})^2} \tag{7}$$

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

Relative mean error (RME):

RME quantifies the average error between the ensemble mean forecast and their corresponding observation 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:

$$\mathrm{RME} = \frac{\sum_{i=1}^{n}(\overline{X_i} - Y_i)}{\sum_{i=1}^{n} Y_i}, \tag{8}$$

where $\overline{X_i} = 1/m\sum_{k=1}^{m} X_{i,k}$, 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.

Brier Skill Score (BSS):

The Brier score (BS) is analogous to the mean squared 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

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}[F_{X_i}(q) - F_{Y_i}(q)]^2, \tag{9}$$

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

$$F_{X_i}(q) = \mathrm{P_r}[X_i > q], \tag{10}$$

n is again the total number of forecast-observation pairs, and

$$F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

In order 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):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS_{main}}}{\mathrm{BS_{reference}}}, \tag{12}$$

where $\mathrm{BS_{main}}$ and $\mathrm{BS_{reference}}$ 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.

Mean Continuous Ranked Probability Skill Score (CRPSS):

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

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left[F_x(q) - F_y(q) \right]^2 dq. \tag{13}$$

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:

$$\mathrm{CRPSS} = 1 - \frac{\overline{\mathrm{CRPS}}_{\mathrm{main}}}{\overline{\mathrm{CRPS}}_{\mathrm{reference}}}, \tag{14}$$

where the CRPS is averaged across *n* pairs of forecasts and observations to calculate the mean CRPS of the main forecast system ($\overline{\mathrm{CRPS}}_{\mathrm{main}}$) and reference forecast system ($\overline{\mathrm{CRPS}}_{\mathrm{reference}}$). The CRPSS ranges from -∞ 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

Addor, N., S. Jaun, F. Fundel, and M. Zappa, 2011: An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.*, **15,** 2327-2347.

Alemu, E., R. Palmer, A. Polebitski, and B. Meaker, 2010: Decision Support System for Optimizing Reservoir Operations Using Ensemble Streamflow Predictions. *Journal of Water Resources Planning and Management*, **137,** 72-82.

Alfieri, L., F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, and P. Salamon, 2014: Evaluation of ensemble streamflow predictions in Europe. *Journal of Hydrology*, **517,** 913-922.

Anderson, R. M., V. I. Koren, and S. M. Reed, 2006: Using SSURGO data to improve Sacramento Model a priori parameter estimates. *Journal of Hydrology*, **320,** 103-116.

Anghileri, D., N. Voisin, A. Castelletti, F. Pianosi, B. Nijssen, and D. P. Lettenmaier, 2016: Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. *Water Resources Research*, **52,** 4209-4225.

Bartholmes, J. C., J. Thielen, M. H. Ramos, and S. Gentilini, 2009: The european flood alert system EFAS–Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrology and Earth System Sciences*, **13,** 141-153.

Bennett, J. C., D. E. Robertson, D. L. Shrestha, Q. J. Wang, D. Enever, P. Hapuarachchi, and N. K. Tuteja, 2014: A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9 days. *Journal of Hydrology*, **519, Part D,** 2832-2846.

Boucher, M. A., F. Anctil, L. Perreault, and D. Tremblay, 2011: A comparison between ensemble and deterministic hydrological forecasts in an operational context. *Adv. Geosci.*, **29,** 85-94.

Boyle, D. P., H. V. Gupta, S. Sorooshian, V. Koren, Z. Zhang, and M. Smith, 2001: Toward improved streamflow forecasts: value of semidistributed modeling. *Water Resources Research*, **37,** 2749-2759.

Brown, J., 2014: Verification of temperature, precipitation and streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service: an evolution of the medium-range forecasts with forcing inputs from NCEP's Global Ensemble Forecast System (GEFS) and a comparison to the frozen version of NCEP's Global Forecast System (GFS). *Hydrologic Solutions Limited, Subcontract Agreement 2013-09 with LEN Technologies Inc*.

Brown, J., J. Demargne, D.-J. Seo, and Y. Liu, 2010: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations *Environmental Modeling and Software*, **25,** 854-872.

Brown, J. D., and D.-J. Seo, 2010: A Nonparametric Postprocessor for Bias Correction of Hydrometeorological and Hydrologic Ensemble Forecasts. *Journal of Hydrometeorology*, **11,** 642-665.

Brown, J. D., M. He, S. Regonda, L. Wu, H. Lee, and D.-J. Seo, 2014: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *Journal of Hydrology*, **519, Part D,** 2847-2868.

Burnash, R., and V. Singh, 1995: The NWS river forecast system-Catchment modeling. *Computer models of watershed hydrology***.,** 311-366.

Burnash, R. J., R. L. Ferral, and R. A. McGuire, 1973: A generalized streamflow simulation system, conceptual modeling for digital computers.

Carpenter, T. M., and K. P. Georgakakos, 2006: Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales. *Journal of Hydrology*, **329,** 174-185.

Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *Journal of Hydrology*, **375,** 613-626.

Day, G., 1985: Extended Streamflow Forecasting Using NWSRFS. *Journal of Water Resources Planning and Management*, **111,** 157-170.

Demargne, J., and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society*, **95,** 79-98.

Demeritt, D., S. Nobert, H. Cloke, and F. Pappenberger, 2010: Challenges in communicating and using ensembles in operational flood forecasting. *Meteorological Applications*, **17,** 209-222.

ene Michaud, J., and S. Sorooshian, 1994: Comparison of simple versus complex distributed runoff models on a midsized semiarid watershed. *Water Resources Research*, **30,** 593-605.

Famiglietti, J. S., and M. Rodell, 2013: Water in the Balance. *Science*, **340,** 1300-1301.

Fan, F. M., W. Collischonn, A. Meller, and L. C. M. Botelho, 2014a: Ensemble streamflow forecasting experiments in a tropical basin: The São Francisco river case study. *Journal of Hydrology*, **519, Part D,** 2906-2919.

——, 2014b: Ensemble streamflow forecasting experiments in a tropical basin: The São Francisco river case study. *Journal of Hydrology*.

Franz, K. J., T. S. Hogue, and S. Sorooshian, 2008: Snow Model Verification Using Ensemble Prediction and Operational Benchmarks. *Journal of Hydrometeorology*, **9,** 1402-1415.

Georgakakos, A. P., H. Yao, and K. P. Georgakakos, 2014: Ensemble streamflow prediction adjustment for upstream water use and regulation. *Journal of Hydrology*, **519, Part D,** 2952-2966.

Glahn, H. R., and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, **11,** 1203-1211.

Habib, E., A. T. Haile, Y. Tian, and R. J. Joyce, 2012: Evaluation of the High-Resolution CMORPH Satellite Rainfall Product Using Dense Rain Gauge Observations and Radar-Based Estimates. *Journal of Hydrometeorology*, **13,** 1784-1798.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Monthly Weather Review*, **132,** 1434-1447.

Hamill, T. M., and Coauthors, 2013: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bulletin of the American Meteorological Society*, **94,** 1553-1565.

Harshburger, B. J., V. P. Walden, K. S. Humes, B. C. Moore, T. R. Blandford, and A. Rango, 2012: Generation of Ensemble Streamflow Forecasts Using an Enhanced Version of the Snowmelt Runoff Model. *JAWRA Journal of the American Water Resources Association*, **48,** 643-655.

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci.*, **11,** 939-950.

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and Temporal Characteristics of Heavy Hourly Rainfall in the United States. *Monthly Weather Review*, **141,** 4564-4575.

Jones, K. B., and Coauthors, 1997: An ecological assessment of the United States mid-Atlantic region: a landscape atlas.

Journel, A. G., and C. J. Huijbregts, 1978: *Mining geostatistics.* Academic press.

Kelly, P., 2014: What to do when we run out of water. *Nature Clim. Change*, **4,** 314-316.

Khan, M., A. Shamseldin, B. Melville, and M. Shoaib, 2014: Stratification of NWP Forecasts for Medium-Range Ensemble Streamflow Forecasting. *Journal of Hydrologic Engineering*, **20,** 04014076.

Kitzmiller, D., and Coauthors, 2011: Evolving Multisensor Precipitation Estimation Methods: Their Impacts on Flow Prediction Using a Distributed Hydrologic Model. *Journal of Hydrometeorology*, **12,** 1414-1431.

Koren, V., S. Reed, M. Smith, Z. Zhang, and D.-J. Seo, 2004: Hydrology laboratory research modeling system (HL-RMS) of the US national weather service. *Journal of Hydrology*, **291,** 297-318.

Koren, V., F. Moreda, S. Reed, M. Smith, and Z. Zhang, 2007: Evaluation of a grid-based distributed hydrological model over a large area. *Water and Energy Abstracts*, **16,** 13-14.

Koren, V. I., M. Smith, D. Wang, and Z. Zhang, 2000: Use of soil property data in the derivation of conceptual runoff-runoff model parameters. *15th Conference on Hydrology, AMS, January 9-14,2000, Long Beach, CA*.

Krajewski, W. F., V. Lakshmi, K. P. Georgakakos, and S. C. Jain, 1991: A Monte Carlo study of rainfall sampling effect on a distributed catchment model. *Water resources research*, **27,** 119-128.

Krzysztofowicz, R., 1997: Transformation and normalization of variates with specified distributions. *Journal of Hydrology*, **197,** 286-292.

Kuzmin, V., D.-J. Seo, and V. Koren, 2008: Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search. *Journal of Hydrology*, **353,** 109-128.

Kuzmin, V. A., 2009: Algorithms of automatic calibration of multi-parameter models used in operational systems of flash flood forecasting. *Russian Meteorology and Hydrology*, **34,** 473-481.

Lee, H., D.-J. Seo, and V. Koren, 2011: Assimilation of streamflow and in situ soil moisture data into operational distributed hydrologic models: Effects of uncertainties in the data and initial model soil moisture states. *Advances in Water Resources*, **34,** 1597-1615.

Lee, H., Y. Zhang, D.-J. Seo, and P. Xie, 2015: Utilizing satellite precipitation estimates for streamflow forecasting via adjustment of mean field bias in precipitation data and assimilation of streamflow observations. *Journal of Hydrology*, **529, Part 3,** 779-794.

Li, H., L. Luo, E. F. Wood, and J. Schaake, 2009: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *Journal of Geophysical Research: Atmospheres*, **114**.

Lin, Y., and K. E. Mitchell, 2005: 1.2 the NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. Hydrology, American Meteorological Society, San Diego, CA, USA*, Citeseer.

Madadgar, S., H. Moradkhani, and D. Garen, 2014: Towards improved post-processing of hydrologic forecast ensembles. *Hydrological Processes*, **28,** 104-122.

Maupin, M. A., J. F. Kenny, S. S. Hutson, J. K. Lovelace, N. L. Barber, and K. S. Linsey, 2014: Estimated use of water in the United States in 2010: U.S. Geological Survey Circular 1405. 56.

McCuen, R. H., and W. M. Snyder, 1975: A proposed index for comparing hydrographs. *Water Resources Research*, **11,** 1021-1024.

Mejia, A. I., and S. M. Reed, 2011: Evaluating the effects of parameterized cross section shapes and simplified routing with a coupled distributed hydrologic and hydraulic model. *Journal of Hydrology*, **409,** 512-524.

Mekonnen, M. M., and A. Y. Hoekstra, 2016: Four billion people facing severe water scarcity. *Science Advances*, **2**.

Moore, B. J., K. M. Mahoney, E. M. Sukovich, R. Cifelli, and T. M. Hamill, 2015: Climatology and environmental characteristics of extreme precipitation events in the Southeastern United States. *Monthly Weather Review*, **143,** 718-741.

NOAA, 2014: http://www.nws.noaa.gov/ohd/hrl/general/indexdoc.htm.

Olsson, J., and G. Lindström, 2008: Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. *Journal of Hydrology*, **350,** 14-24.

Pagano, T. C., D. L. Shrestha, Q. J. Wang, D. Robertson, and P. Hapuarachchi, 2013: Ensemble dressing for hydrological applications. *Hydrological Processes*, **27,** 106-116.

Prat, O. P., and B. R. Nelson, 2015: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). *Hydrology and Earth System Sciences*, **19,** 2037-2056.

Rafieeinasab, A., D.-J. Seo, H. Lee, and S. Kim, 2014: Comparative evaluation of maximum likelihood ensemble filter and ensemble Kalman filter for real-time assimilation of streamflow data into operational hydrologic models. *Journal of Hydrology*, **519, Part D,** 2663-2675.

Rafieeinasab, A., A. Norouzi, D.-J. Seo, and B. Nelson, 2015a: Improving high-resolution quantitative precipitation estimation via fusion of multiple radar-based precipitation products. *Journal of Hydrology*, **531, Part 2,** 320-336.

Rafieeinasab, A., and Coauthors, 2015b: Toward high-resolution flash flood prediction in large urban areas – Analysis of sensitivity to spatiotemporal resolution of rainfall input and hydrologic modeling. *Journal of Hydrology*, **531, Part 2,** 370-388.

Ramos, M. H., S. J. Van Andel, and F. Pappenberger, 2013: Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, **17,** 2219-2232.

Reed, P. M., and Coauthors, 2006: Bridging river basin scales and processes to assess human-climate impacts and the terrestrial hydrologic system. *Water Resources Research*, **42**.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, D.-J. Seo, and D. Participants, 2004: Overall distributed model intercomparison project results. *Journal of Hydrology*, **298,** 27-60.

Regonda, S. K., D.-J. Seo, B. Lawrence, J. D. Brown, and J. Demargne, 2013: Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts – A Hydrologic Model Output Statistics (HMOS) approach. *Journal of Hydrology*, **497,** 80-96.

Roulin, E., 2006: Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrology and Earth System Sciences Discussions*, **3,** 1369-1406.

Roulin, E., and S. Vannitsem, 2015: Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors. *Hydrological Processes*, **29,** 1434-1449.

Schaake, J., K. Franz, A. Bradley, and R. Buizza, 2006: The Hydrologic Ensemble Prediction EXperiment (HEPEX). *Hydrology and Earth System Sciences Discussions*, **3,** 3321-3332.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX: The Hydrological Ensemble Prediction Experiment. *Bulletin of the American Meteorological Society*, **88,** 1541-1547.

Schellekens, J., A. H. Weerts, R. J. Moore, C. E. Pierce, and S. Hildon, 2011: The use of MOGREPS ensemble rainfall forecasts in operational flood forecasting systems across England and Wales. *Adv. Geosci.*, **29,** 77-84.

Sharma, S., and Coauthors, 2016: Eastern U.S. verification of ensemble precipitation forecasts. *Weather and Forecasting (In review)*.

Siddique, R., A. Mejia, J. Brown, S. Reed, and P. Ahnert, 2015: Verification of precipitation forecasts from two numerical weather prediction models in the Middle Atlantic Region of the USA: A precursory analysis to hydrologic forecasting. *Journal of Hydrology*, **529, Part 3,** 1390-1406.

Smith, M. B., and Coauthors, 2004: The distributed model intercomparison project (DMIP): motivation and experiment design. *Journal of Hydrology*, **298,** 4-26.

——, 2012: The distributed model intercomparison project – Phase 2: Motivation and design of the Oklahoma experiments. *Journal of Hydrology*, **418–419,** 3-16.

Sorooshian, S., and V. K. Gupta, 1983: Automatic calibration of conceptual rainfall- runoff models: The question of parameter observability and uniqueness. *Water Resources Research*, **19,** 260-268.

Spies, R. R., K. J. Franz, T. S. Hogue, and A. L. Bowman, 2014: Distributed Hydrologic Modeling Using Satellite-Derived Potential Evapotranspiration. *Journal of Hydrometeorology*, **16,** 129-146.

Tang, Y., P. Reed, and T. Wagener, 2006: How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration? *Hydrol. Earth Syst. Sci.*, **10,** 289-307.

Tang, Y., P. Reed, K. Van Werkhoven, and T. Wagener, 2007: Advancing the identification and evaluation of distributed rainfall- runoff models using global sensitivity analysis. *Water Resources Research*, **43**.

Thielen, J., J. Bartholmes, M. H. Ramos, and A. de Roo, 2009: The European Flood Alert System – Part 1: Concept and development. *Hydrol. Earth Syst. Sci.*, **13,** 125-140.

Thorstensen, A., P. Nguyen, K. Hsu, and S. Sorooshian, 2015: Using Densely Distributed Soil Moisture Observations for Calibration of a Hydrologic Model. *Journal of Hydrometeorology*, **17,** 571-590.

van Andel, S. J., A. Weerts, J. Schaake, and K. Bogner, 2013: Post-processing hydrological ensemble predictions intercomparison experiment. *Hydrological Processes*, **27,** 158-161.

Van Cooten, S., and Coauthors, 2011: The CI-FLOW project: a system for total water level prediction from the summit to the sea. *Bulletin of the American Meteorological Society*, **92,** 1427.

Van Steenbergen, N., J. Ronsyn, and P. Willems, 2012: A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication. *Environmental Modelling & Software*, **33,** 92-105.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang, 2008: Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resources Research*, **44,** n/a-n/a.
Verbunt, M., A. Walser, J. Gurtz, A. Montani, and C. Schär, 2007: Probabilistic Flood Forecasting with a Limited-Area Ensemble Prediction System: Selected Case Studies. *Journal of Hydrometeorology*, **8,** 897-909.
Vörösmarty, C. J., P. Green, J. Salisbury, and R. B. Lammers, 2000: Global Water Resources: Vulnerability from Climate Change and Population Growth. *Science*, **289,** 284-288.
Wetterhall, F., and Coauthors, 2013: HESS Opinions "Forecaster priorities for improving probabilistic flood forecasts". *Hydrol. Earth Syst. Sci.*, **17,** 4389-4399.
Wilks, D. S., 2015: Multivariate ensemble Model Output Statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, **141,** 945-952.
Wood, A. W., and J. C. Schaake, 2008: Correcting Errors in Streamflow Forecast Ensemble Mean and Spread. *Journal of Hydrometeorology*, **9,** 132-148.
Wood, A. W., T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2015: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill. *Journal of Hydrometeorology*, **17,** 651-668.
Xuan, Y., I. D. Cluckie, and Y. Wang, 2009: Uncertainty analysis of hydrological ensemble forecasts in a distributed model utilising short-range rainfall prediction. *Hydrol. Earth Syst. Sci.*, **13,** 293-303.
Yuan, X., E. F. Wood, and M. Liang, 2014: Integrating weather and climate prediction: Toward seamless hydrologic forecasting. *Geophysical Research Letters*, **41,** 5891-5896.
Zalachori, I., M. H. Ramos, R. Garçon, T. Mathevet, and J. Gailhard, 2012: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Advances in Science & Research*, **8,** p. 135 - p. 141.
Zhang, J., and Coauthors, 2011: NATIONAL MOSAIC AND MULTI-SENSOR QPE (NMQ) SYSTEM: Description, Results, and Future Plans. *Bulletin of the American Meteorological Society*, **92,** 1321-1338.

Table 1. Main characteristics of the eight study basins.

	Delaware		**NB Susquehanna**		**Potomac**		**James**	
Location of outlet	Walton, New York	Trenton, New Jersey	Sherburne New York	Waverly, New York	Dawsonville, Maryland	Little Falls, DC	Palmyra, Virginia	Richmond, Virginia
NWS id	WALN6	TREN4	SHBN6	WVYN6	DAWM2	BRKM2	PYAV2	RMDV2
USGS id	01423000	01463500	01505000	01515000	01645000	01646500	02034000	02037500
Area [km^2]	860	17,574	682	12,372	262	29,965	1719	17,504
Latitude	$42^0$09’58”	$40^0$13’18”	$42^0$40’43”	$41^0$59’05”	$39^0$07’41”	$38^0$56’59”	$37^0$51’28”	$37^0$33’47”
Longitude	$75^0$08’24”	$74^0$46’41”	$75^0$30’38”	$76^0$30’04”	$77^0$20’08”	$77^0$07’39”	$78^0$15’58”	$77^0$32’50”
Minimum daily flow* [m^3/s]	0.62 (0.37)	71.36 (35.11)	0.59 (0.39)	13.08 (6.71)	1.30 (0.05)	11.47 (3.42)	0.68 (0.15)	12.45 (0.28)
Maximum daily flow* [m^3/s]	634.29 (634.29)	6512.87 (7900)	278.63 (278.63)	4417.42 (4417.42)	17.90 (280.34)	5436.83 (12060)	430.42 (1926)	2860.00 (8382)
Mean daily flow* [m^3/s]	21.25 (17.23)	415.87 (338.90)	13.32 (11.38)	275.55 (215.01)	4.28 (3.28)	323.67 (325.72)	17.56 (20.23)	193.77 (196.47)
Climatological flow (Pr=0.50) [m^3/s]	13.36	241.17	8.7	158	3.03	134.65	11.66	118.47
Climatological flow (Pr=0.90) [m^3/s]	42.53	505.31	23.89	434	6.68	404.65	35.88	396.60

*The number in parenthesis is the historical (based on entire available record, as opposed to the period 2004-2011 used in this study) daily minimum, maximum, or mean recorded flow.

Table 2. Performance statistics for the uncalibrated and calibrated simulation runs for the entire period of analysis (2004-2013).

Performance statistic	Model run	Flow condition	Delaware		North Branch Susquehanna		Potomac		James	
			WALN6	TREN4	SHBN6	WVYN6	DAWM2	BRKM2	PYAV2	RMDV2
Correlation coefficient, R	Uncalibrated	Low-moderate	0.76	0.81	0.59	0.82	0.72	0.73	0.75	0.79
		High	0.74	0.88	0.76	0.62	0.55	0.86	0.67	0.58
		Overall	0.86	0.92	0.85	0.82	0.68	0.91	0.82	0.84
	Calibrated	Low-moderate	0.76	0.83	0.87	0.79	0.71	0.76	0.79	0.80
		High	0.81	0.89	0.77	0.72	0.82	0.88	0.66	0.76
		Overall	0.88	0.93	0.89	0.86	0.86	0.92	0.82	0.89
Modified correlation coefficient, R_m	Uncalibrated	Low-moderate	0.74	0.59	0.46	0.66	0.43	0.44	0.50	0.71
		High	0.45	0.84	0.50	0.59	0.29	0.70	0.57	0.56
		Overall	0.60	0.90	0.61	0.79	0.46	0.75	0.80	0.74
	Calibrated	Low-moderate	0.74	0.63	0.80	0.64	0.40	0.60	0.60	0.76
		High	0.67	0.89	0.53	0.69	0.79	0.87	0.66	0.66
		Overall	0.70	0.92	0.74	0.80	0.84	0.90	0.81	0.88
Percent bias, PB	Uncalibrated	Low-moderate	8.94	-5.67	-17.55	3.45	20.01	27.67	27.21	1.88
		High	-23.39	-4.94	17.02	-12.59	-34.65	15.24	1.2	-19.07
		Overall	-3.99	-5.23	-0.74	-4.21	4.82	20.54	16.84	-5.52
	Calibrated	Low-moderate	6.95	-5.49	4.74	4.71	-0.67	2.99	8.04	-0.78
		High	-21.73	-5.57	-8.12	-10.99	-5.15	-5.37	-12.24	-8.03
		Overall	-4.52	-5.54	-1.52	-2.79	0.88	-1.8	-0.04	-3.34
Nash-Sutcliffe efficiency, NSE	Uncalibrated	Low-moderate	0.52	0.32	-0.21	0.49	-0.49	-0.50	-0.14	0.53
		High	0.44	0.73	-0.07	0.16	0.18	0.57	0.42	0.03
		Overall	0.71	0.82	0.42	0.66	0.46	0.71	0.64	0.70
	Calibrated	Low-moderate	0.52	0.43	0.69	0.41	0.67	0.32	0.34	0.57
		High	0.58	0.79	0.57	0.42	0.66	0.76	0.42	0.40
		Overall	0.77	0.86	0.78	0.74	0.71	0.85	0.68	0.79

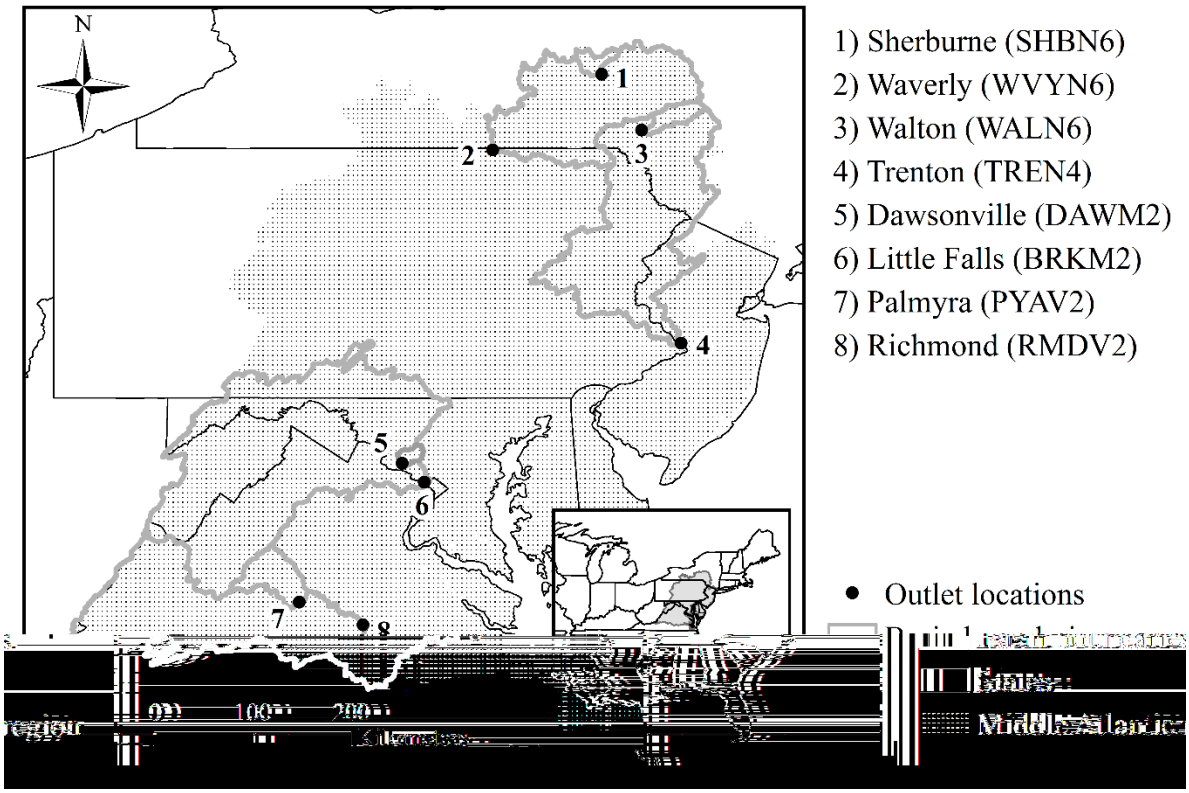


Figure 1. Map illustrating the location of the selected study basins in the MAR.

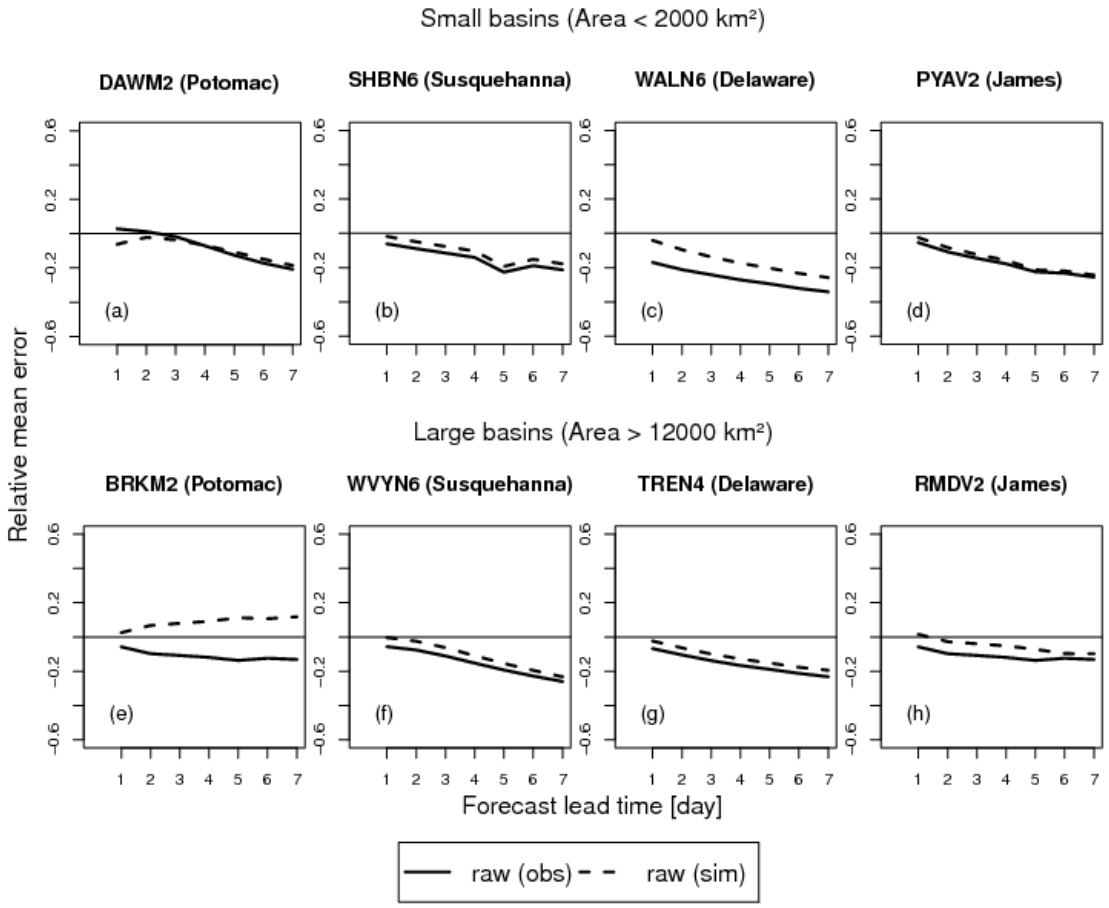


Figure 2. RME of the mean raw, ensemble streamflow forecasts versus the forecast lead time, with reference to both observed (solid line) and simulated (dashed line) flows. Results are shown for small (a-d) and large (e-h) basins, under low-moderate flow conditions (flows with nonexceedance probability of 0.5).

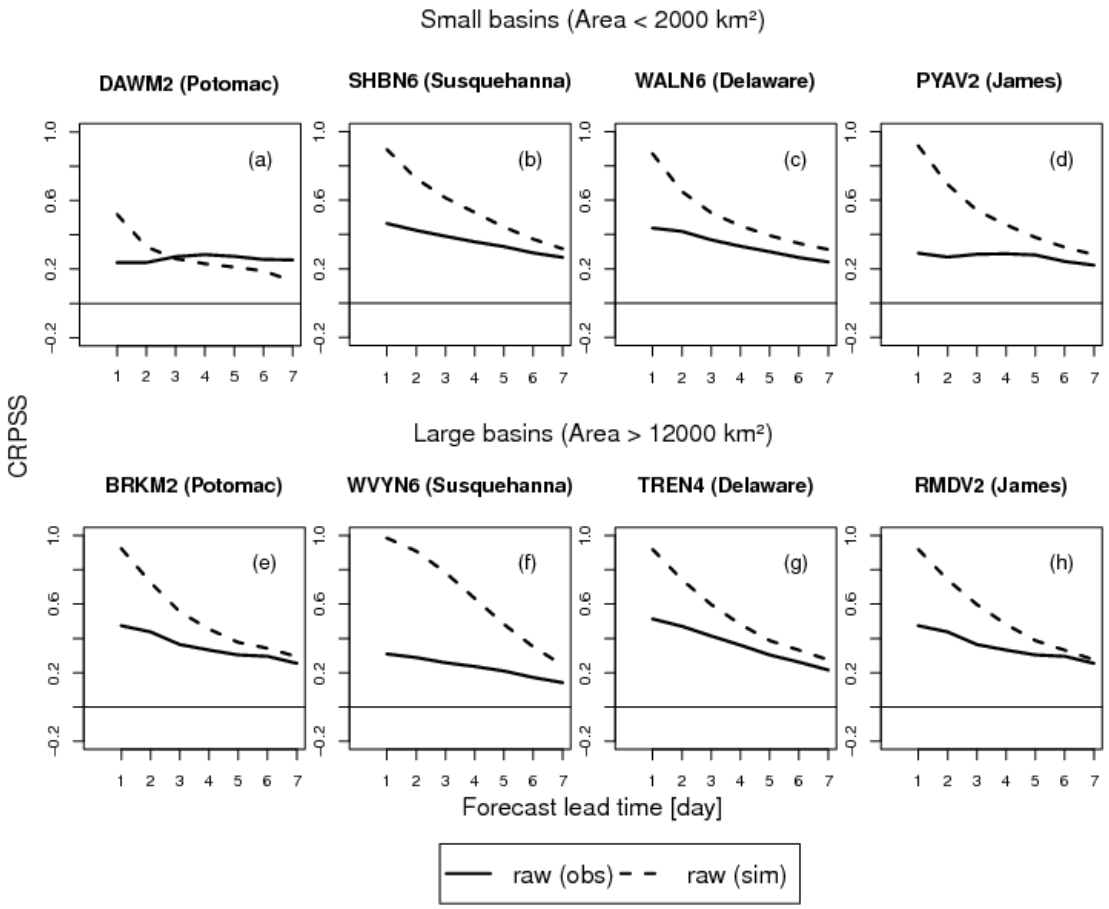


Figure 3. CRPSS of the raw, ensemble streamflow forecasts versus the forecast lead time, with reference to both the observed (solid line) and simulated (dashed line) climatological flows. Results are shown for small (a-d) and large (e-h) basins, under low-moderate flow conditions (flows with nonexceedance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

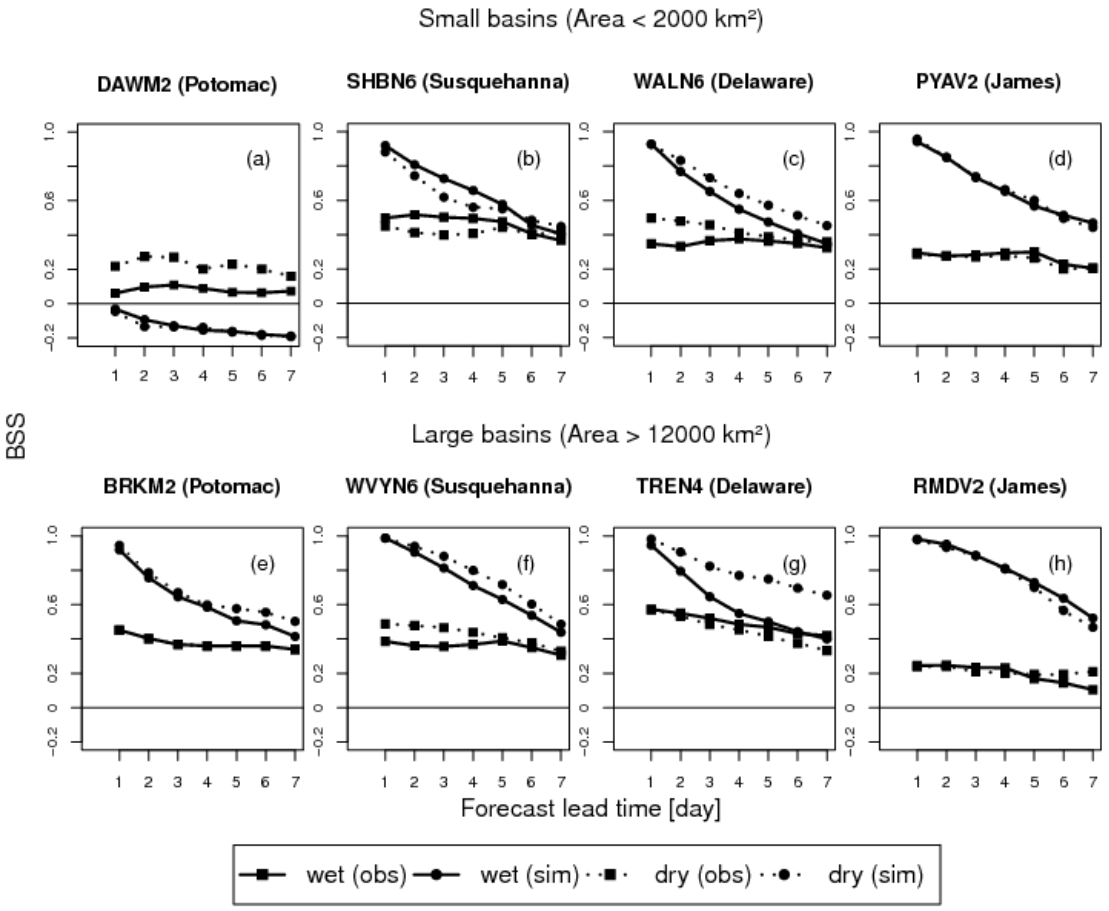


Figure 4. BSS of the raw, ensemble streamflow forecasts versus the forecast lead time, with reference to both the observed (lines with squared symbols) and simulated (lines with dotted symbols) climatological flows. Results are shown for the dry (dotted lines) and wet (solid lines) seasons, for small (a-d) and large (e-h) basins, under low-moderate flow conditions (flows with nonexceedance probability of 0.5). The dry season includes the months of June-November and the wet season the months of December-May.

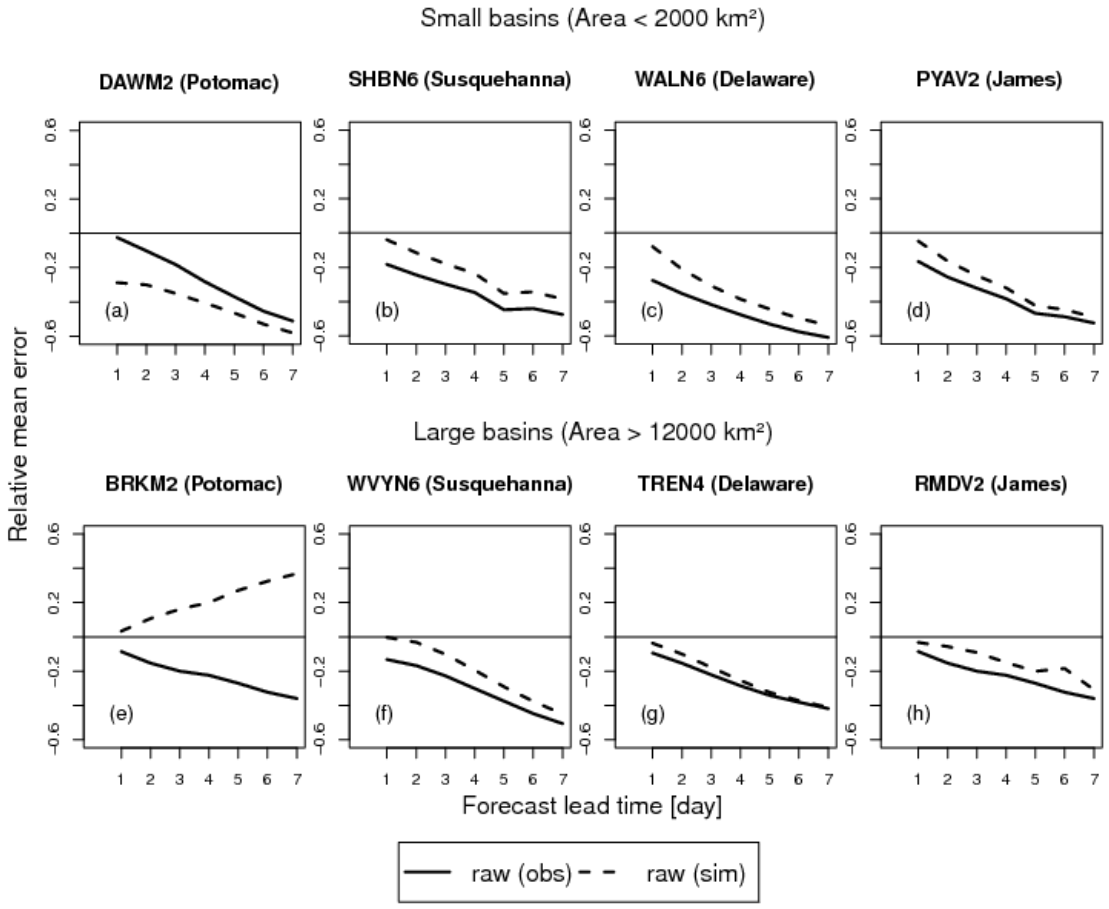


Figure 5. RME of the mean raw, ensemble streamflow forecasts versus the forecast lead time, with reference to both the observed (solid line) and simulated (dashed line) flows. Results are shown for small (a-d) and large (e-h) basins, under high flow conditions (flows with nonexceedance probability of 0.9).

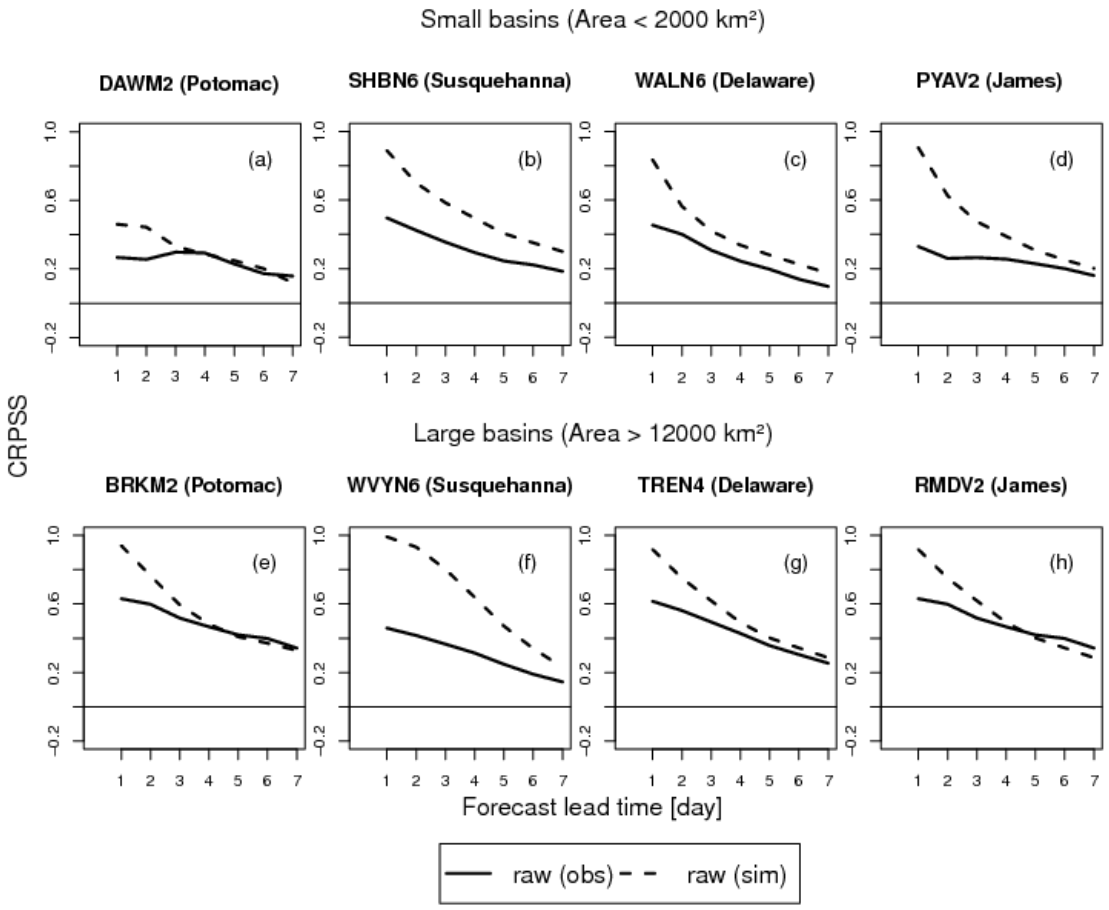


Figure 6. CRPSS of the raw, ensemble streamflow forecasts versus the forecast lead time, with reference to both the observed (solid line) and simulated (dashed line) climatological flows. Results are shown for small (a-d) and large (e-h) basins, under high flow conditions (flows with nonexceedance probability of 0.9).

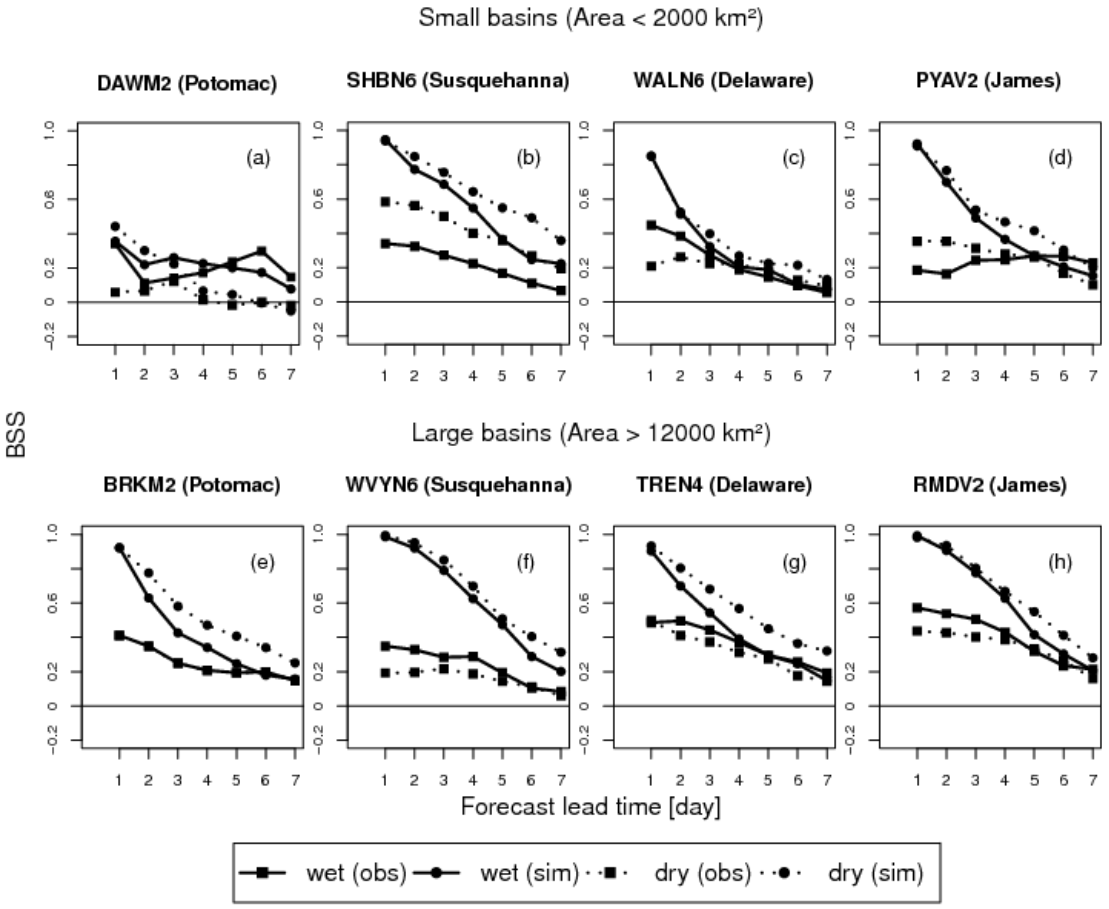


Figure 7. BSS of the raw, ensemble streamflow forecasts versus the forecast lead time, with reference to both the observed (lines with squared symbols) and simulated (lines with dotted symbols) climatological flows. Results are shown for the dry (dotted lines) and wet (solid lines) seasons, for small (a-d) and large (e-h) basins, under high flow conditions (flows with nonexceedance probability of 0.9). The dry season includes the months of June-November and the wet season the months of December-May.

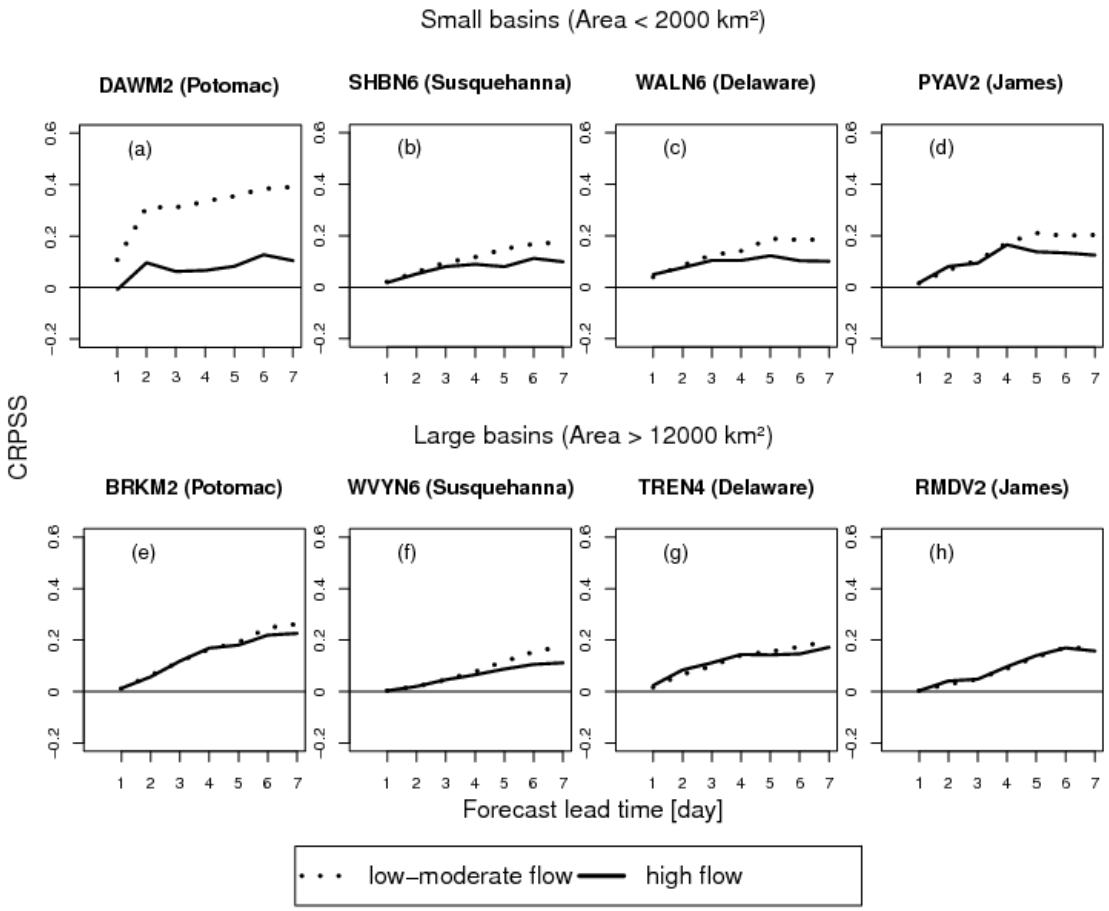


Figure 8. CRPSS of the raw, ensemble streamflow forecasts versus the forecast lead time, with reference to the deterministic forecasts. Results are shown for small (a-d) and large (e-h) basins, under low-moderate (dotted lines) and high (solid lines) flow conditions. The low-moderate and high flows are flows with nonexceedance probabilities of 0.5 and 0.9, respectively.

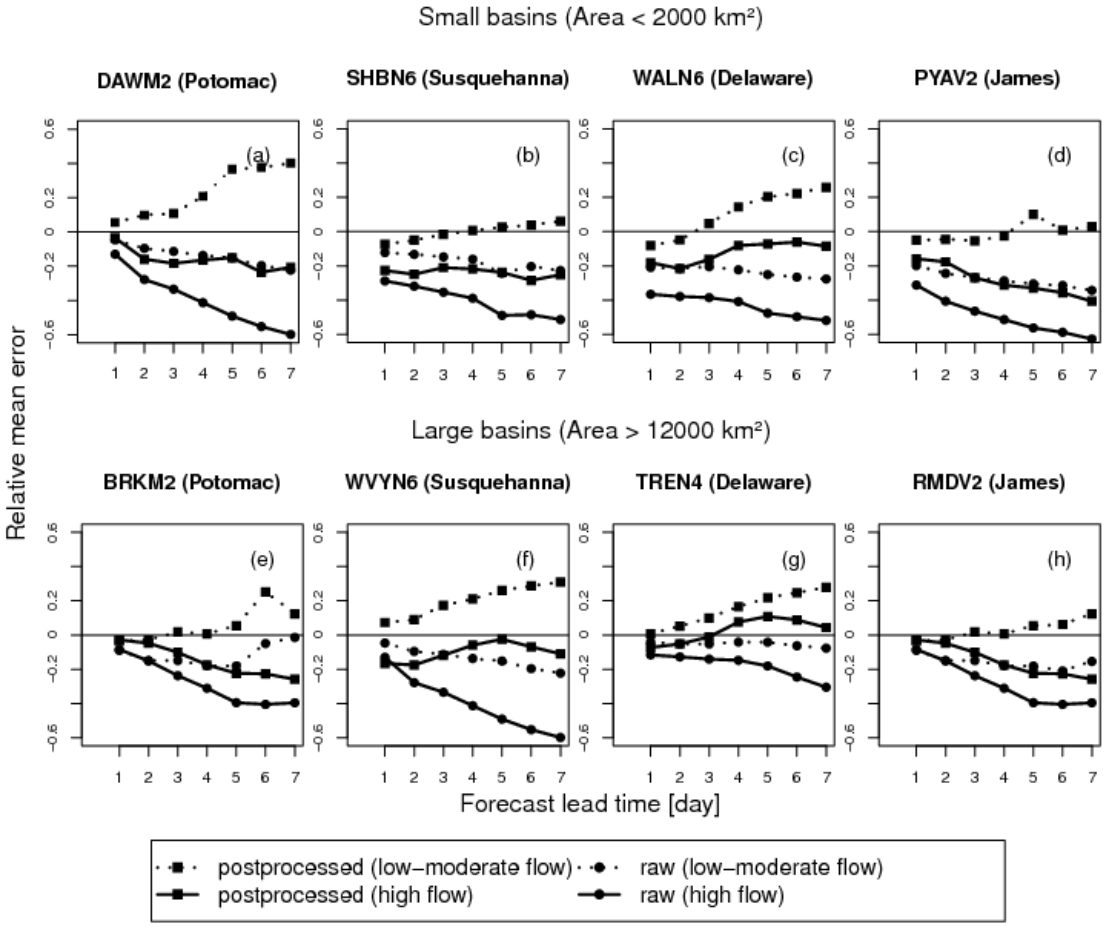


Figure 9. RME of the mean postprocessed (lines with squared symbols) and raw (lines with dotted symbols), ensemble streamflow forecasts versus the forecast lead time. Results are shown for small (a-d) and large (e-h) basins, under low-moderate (dotted lines) and high (solid lines) flow conditions. The low-moderate and high flows are flows with nonexceedance probabilities of 0.5 and 0.9, respectively.

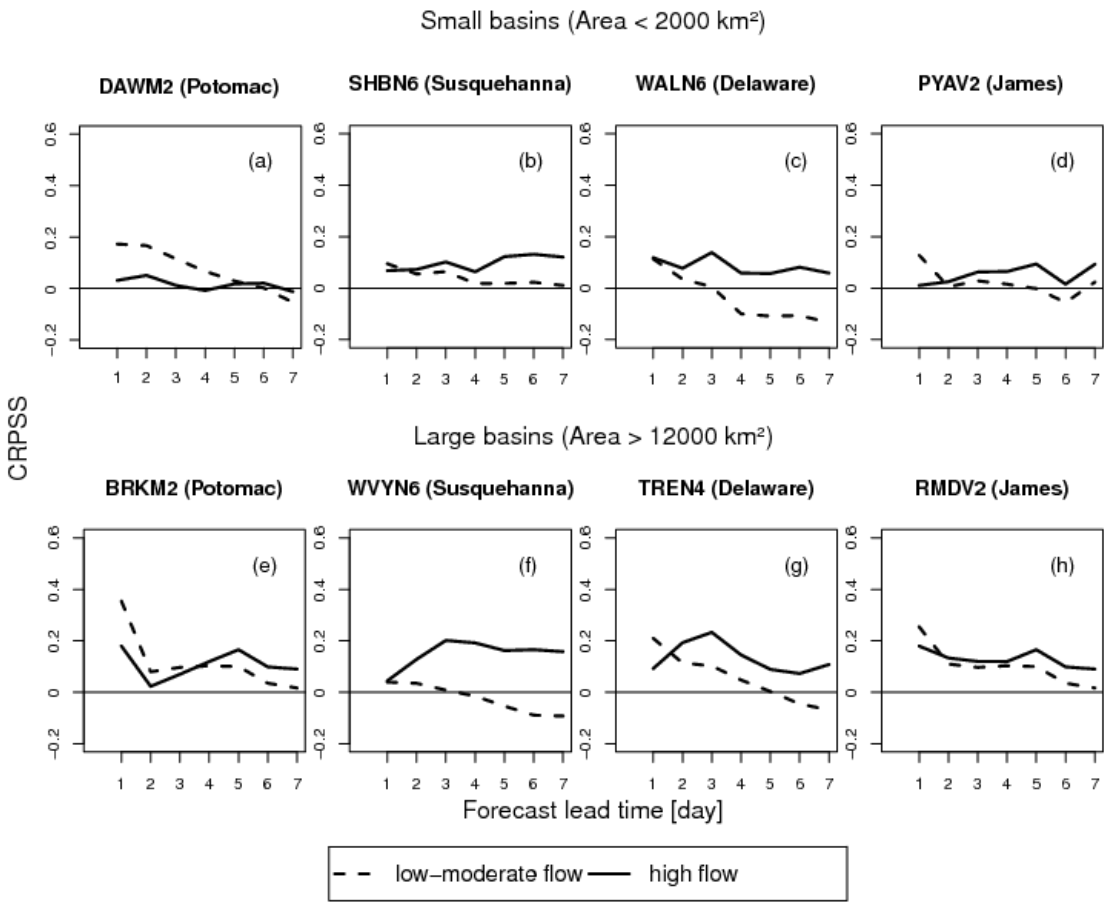


Figure 10. CRPSS of the postprocessed, ensemble streamflow forecasts versus the forecast lead time, with reference to the raw forecasts. Results are shown for small (a-d) and large (e-h) basins, under low-moderate (dashed lines) and high (solid lines) flow conditions. The low-moderate and high flows are flows with nonexceedance probabilities of 0.5 and 0.9, respectively.

Chapter 6: Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system

ABSTRACT

The relative roles of statistical weather preprocessing and streamflow postprocessing in hydrological ensemble forecasting at short- to medium-range forecast lead times (day 1-7) are investigated. For this purpose, a regional hydrologic ensemble prediction system (RHEPS) is developed and implemented. The RHEPS is comprised by the following components: i) hydrometeorological observations (multisensor precipitation estimates, gridded surface temperature, and gauged streamflow); ii) weather ensemble forecasts (precipitation and near-surface temperature) from the National Centers for Environmental Prediction 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2); iii) NOAA's Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM); iv) heteroscedastic censored logistic regression (HCLR) as the statistical preprocessor; v) two statistical postprocessors, an autoregressive model with a single exogenous variable (ARX(1,1)) and quantile regression (QR); and vi) a comprehensive verification strategy. To implement the RHEPS, 1 to 7 days weather forecasts from the GEFSRv2 are used to force HL-RDHM and generate raw ensemble streamflow forecasts. Forecasting experiments are conducted in four nested basins in the U.S. middle Atlantic region, ranging in size from 381 to 12,362 km^2.

Results show that the HCLR preprocessed ensemble precipitation forecasts have greater skill than the raw forecasts. These improvements are more noticeable in the warm season at the longer lead times (>3 days). Both postprocessors, ARX(1,1) and QR, show gains in skill relative to the raw ensemble flood forecasts but QR outperforms ARX(1,1). Preprocessing alone has little effect on improving the skill of the ensemble flood forecasts. Indeed, postprocessing alone performs similar, in terms of the relative mean error, skill, and reliability, to the more involved scenario that includes both preprocessing and postprocessing. We conclude that statistical preprocessing may not always be a necessary component of the ensemble flood forecasting chain.

1. Introduction

The intersection of climate variability and change, increased exposure from expanding urbanization, and sea level rise are increasing the frequency of damaging flood events and making their prediction more challenging across the globe (Dankers et al., 2014; Wheater and Gober, 2015; Ward et al., 2015). Accordingly, current research and operational efforts in hydrological forecasting are seeking to develop and implement enhanced forecasting systems, with the goals of improving the skill and reliability of short- to medium-range flood forecasts (0-14 days), and providing more effective early warning services (Pagano et al., 2014; Thiemig et al., 2015; Emerton et al., 2016; Siddique and Mejia, 2017). Ensemble-based forecasting systems have become the preferred paradigm, showing substantial improvements over single-valued deterministic ones (Schaake et al., 2007; Cloke and Pappenberger, 2009; Demirel et al., 2013; Fan et al., 2014; Demargne et al., 2014; Schwanenberg et al., 2015; Siddique and Mejia, 2017). Ensemble flood forecasts can be generated in a number of ways, being the most common approach the use of meteorological forecast ensembles to force a hydrological model (Cloke and Pappenberger, 2009; Thiemig et al., 2015). Such meteorological forecasts can be generated by multiple alterations of a numerical weather prediction model, including perturbed initial conditions and/or multiple model physics and parameterizations.

A number of ensemble prediction systems (EPSs) are being used to generate flood forecasts. In the United States (U.S.), the NOAA's National Weather Service River Forecast Centers are implementing and using the Hydrological Ensemble Forecast Service to incorporate meteorological ensembles into their flood forecasting operations (Demargne et al., 2014; Brown et al., 2014). Likewise, the European Flood Awareness System from the European Commission (Alfieri et al., 2014) and the Flood Forecasting and Warming Service from the Australia Bureau of Meteorology (Pagano et al., 2016) have adopted the ensemble paradigm. Furthermore, different regional EPSs have been designed and implemented for research purposes, meet specific regional needs, and/or real-time forecasting applications. Two examples, among several others (Zappa et al., 2008; Zappa et al., 2011; Hopson and Webster, 2010; Demuth and Rademacher, 2016; Addor et al., 2011; Golding et al., 2016; Bennett et al., 2014; Schellekens et al., 2011), are the Stevens Institute of Technology's Stevens Flood Advisory System for short-range flood forecasting (Saleh et al., 2016), and the National Center for Atmospheric Research (NCAR)'s System for Hydromet Analysis, Research, and Prediction for medium-range streamflow forecasting (NCAR, 2017). Further efforts are underway to operationalize global ensemble flood forecasting and early warning systems, e.g., through the Global Flood Awareness System (Alfieri et al., 2013; Emerton et al., 2016).

EPSs are comprised by several system components. In this study, the Regional Hydrological Ensemble Prediction System (RHEPS) is used (Siddique and Mejia, 2017). The RHEPS is an ensemble-based research forecasting system, aimed primarily at bridging the gap between hydrological forecasting research and operations by creating an adaptable and modular forecast emulator. The goal with the RHEPS is to facilitate the integration and rigorous verification of new system components, enhanced physical parameterizations, and novel assimilation strategies. For this study, the RHEPS is comprised by the following system components: i) precipitation and near surface temperature ensemble forecasts from the National Centers for Environmental Prediction 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2), ii) NOAA's Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM) (Reed et al., 2004; Smith et al., 2012a; Smith et al., 2012b), iii) statistical weather preprocessor (hereafter referred to as preprocessing), iv) statistical streamflow postprocessor (hereafter referred to as postprocessing), v) hydrometeorological observations, and vi) verification strategy. Recently, Siddique and Mejia (2017) employed the RHEPS to produce and verify ensemble streamflow forecasts over some of the major river basins in the U.S. middle Atlantic region. Here, the RHEPS is specifically implemented to investigate the relative roles played by preprocessing and postprocessing in enhancing the quality of ensemble flood forecasts.

The goal with statistical processing is to use statistical tools to quantify the uncertainty of and remove systematic biases in the weather and streamflow forecasts in order to improve the skill and reliability of forecasts. In weather and hydrological forecasting, a number of studies have demonstrated the benefits of separately implementing preprocessing (Sloughter et al., 2007; Verkade et al., 2013; Messner et al., 2014a; Yang et al., 2017) and postprocessing (Brown and Seo, 2010; Madadgar et al., 2014; Wang et al., 2016; Siddique and Mejia, 2017). However, only a very limited number of studies have investigated the combined ability of preprocessing and postprocessing to improve the overall quality of ensemble streamflow forecasts (Kang et al., 2010; Zalachori et al., 2012; Roulin and Vannitsem, 2015). At first glance, in the context of medium-range streamflow forecasting, preprocessing seems necessary and beneficial since meteorological forcing are often biased and their uncertainty more dominant than the hydrological one (Cloke and Pappenberger, 2009; Bennett et al., 2014; Siddique and Mejia,

2017). In addition, some streamflow postprocessors assume unbiased forcing (Zhao et al., 2011) and hydrological models can be sensitive to forcing biases.

The few studies that have analyzed the joint effects of preprocessing and postprocessing on short- to medium-range streamflow forecasts have mostly relied on weather ensembles from the European Centre for Medium-range Weather Forecasts (ECMWF) (Roulin and Vannitsem (2015), Zalachori et al. (2012) Benninga et al., 2016). Kang et al. (2010) used different forcing but focused on monthly, as opposed to daily, streamflow. The conclusions from these studies have been mixed (Benninga et al., 2016). Some have found statistical processing to be useful, particularly postprocessing, while others have found that it contributes little to forecast quality. Overall, studies indicate that the relative effects of preprocessing and postprocessing depend strongly on the forecasting system (e.g., forcing, hydrological model, statistical processing technique, etc.) and conditions (e.g., lead time, study area, season, etc.), underscoring the research need to rigorously verify and benchmark new forecasting systems that incorporate statistical processing.

The main objective of this study is to verify and assess the ability of preprocessing and postprocessing to improve ensemble flood forecasts from the RHEPS. This study differs from previous ones in several important respects. The assessment of statistical processing is done using a spatially distributed hydrological model whereas previous studies have tended to emphasize spatially lumped models. Much of the previous studies have used ECMWF forecasts, here we rely on GEFSRv2 precipitation and temperature outputs. Also, we test and implement a preprocessor, namely heteroscedastic censored logistic regression (HCLR), which has not been used before in streamflow forecasting. We also consider a relatively wider range of nested, basin sizes and longer study period than in previous studies. In particular, this paper addresses the following questions:

- What are the separate and joint contributions of preprocessing and postprocessing over the raw RHEPS outputs?
- What forecast conditions (e.g., lead time, season, flood threshold, and basin size) benefit potential increases in skill?
- How much skill improvement can be expected from statistical processing under different uncertainty scenarios (i.e., when skill is measured relative to observed or simulated flow conditions)?

The remainder of the paper is organized as follows. In section 2, the study area and datasets employed are described. Section 3 describes the methodology, including the preprocessor, postprocessor, hydrological model, and forecast verification strategy. The main results and their implications are examined in section 4. Lastly, section 5 summarizes key findings.

2. Study area and datasets

2.1 Study area

The North Branch Susquehanna River basin in the U.S. middle Atlantic region (MAR) is selected as the study area (Fig. 1), with an overall drainage area of 12,362 km^2. The MAR is selected as flooding is an important regional concern. The MAR has a high level of urbanization and high frequency of extreme weather events, making it particularly vulnerable to damaging flood events (Gitro et al., 2014; MARFC, 2017). In the North Branch Susquehanna River basin, four different U.S. Geological Survey (USGS) daily gauge stations, representing a system of nested subbasins, are selected as the forecast locations (Fig. 1). The selected locations are the Ostellic River at Cincinnatus (USGS gauge 01510000), Chenango River at Chenango Forks

(USGS gauge 01512500), Susquehanna River at Conklin (USGS gauge 01503000), and Susquehanna River at Waverly (USGS gauge 01515000) (Fig. 1). The drainage area of the selected basins ranges from 381 to 12,362 km^2. Table 1 outlines some key characteristics of the study basins.

2.2 Datasets

2.2.1 Hydrometeorological observations

Three main observation datasets are used: multisensor precipitation estimates (MPEs), gridded near-surface air temperature, and daily streamflow. MPEs and gridded near-surface air temperature are used to run the hydrological model in simulation mode for parameter calibration purposes and to initialize the RHEPS. Both the MPEs and gridded near-surface air temperature data at 4 x 4 km^2 resolution were provided by the NOAA's Middle Atlantic River Forecast Center (MARFC). Similar to the NCEP stage-IV dataset (Moore et al., 2015; Prat and Nelson, 2015), the MARFC's MPEs represent a continuous time series of hourly, gridded precipitation observations at 4 x 4 km^2 cells, which are produced by combining multiple radar estimates and rain gauge measurements. The gridded near-surface air temperature data at 4 x 4 km^2 resolution were developed by combining multiple temperature observation networks as described by Siddique and Mejia (2017). Daily streamflow observations for the selected basins were obtained from the USGS. The streamflow observations are used to verify the simulated flows, and the raw and postprocessed ensemble streamflow forecasts.

2.2.2 Meteorological forecasts

GEFSRv2 data are used for the ensemble precipitation and near-surface air temperature forecasts. The GEFSRv2 uses the same atmospheric model and initial conditions as the version 9.0.1 of the Global Ensemble Forecast System and runs at T254L42 (~0.50^o Gaussian grid spacing or ~55 km) and T190L42 (~0.67^o Gaussian grid spacing or ~73 km) resolutions for the first and second 8 days, respectively (Hamill et al., 2013). The reforecasts are initiated once daily at 00 Coordinated Universal Time. Each forecast cycle consists of 3 hourly accumulations for day 1 to day 3 and 6 hourly accumulations for day 4 to day 16. In this study, we use 9 years of GEFSRv2 data, from 2004 to 2012, and forecast lead times from 1 to 7 days. The period 2004 to 2012 is selected to take advantage of data that were previously available to us (i.e., GEFSRv2 and MPEs for the MAR) from a recent verification study (Siddique et al., 2015). Forecast lead times of up to 7 days are chosen since we previously found that the GEFSRv2 skill is low after 7 days (Siddique et al., 2015; Sharma et al., 2017).

3. Methodology

3.1 Distributed hydrological model

NOAA's HL-RDHM is used as the spatially distributed hydrological model (Koren et al., 2004). Within HL-RDHM, the Sacramento Soil Moisture Accounting model with Heat Transfer (SAC-HT) is used to represent hillslope runoff generation, and the SNOW-17 module is used to represent snow accumulation and melting.

HL-RDHM is a spatially distributed 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 comprised of storage components that store and transmit water; i.e., each cell acts as a hillslope capable of generating surface, interflow, and groundwater runoff that discharges directly into the streams. The cells are connected to each other through the stream

network system. Through the SNOW-17 module, each cell can also accumulate snow and generate hillslope snow melt based on the near-surface air 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., 2012a). Likewise, flows in the stream network are routed downstream using a nonlinear kinematic wave algorithm that accounts for parameterized stream cross-section shapes (Smith et al., 2012a; Koren et al., 2004). Here we run HL-RDHM in a fully distributed manner at a spatial resolution of 2 x 2 km^2. Note that HL-RDHM requires the forcing to be input at the 4 x 4 km^2 resolution but the model itself can actually be ran at different resolutions. The 2 x 2 km^2 resolution mainly allows for a more realistic representation of the stream network. Further information about the HL-RDHM can be found elsewhere (Koren et al., 2004; Reed et al., 2007; Smith et al., 2012a; Fares et al., 2014; Rafieeinasab et al., 2015; Thorstensen et al., 2016; Siddique and Mejia 2017).

To calibrate HL-RHDM, we first run the model using a-priori parameter estimates previously derived from available datasets (Koren et al., 2000; Reed et al., 2004; Anderson et al., 2006). We then select 10 out of the 17 SAC-HT parameters for calibration based upon prior experience and preliminary sensitivity tests. During the calibration process, each a-priori parameter field is multiplied by a factor. Therefore, we calibrate these factors instead of the parameter values at all grid cells, assuming that the a-priori parameter distribution is true (e.g., Mendoza et al., 2012).The multiplying factors are adjusted manually first; once the manual changes do not yield noticeable improvements in model performance, the factors are tuned-up using stepwise line search (SLS; Kuzmin et al., 2008; Kuzmin, 2009). 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 is optimized:

$$OF = \sqrt{\sum_{i=1}^{m} \left[q_i - s_i(\Omega) \right]^2}, \quad (1)$$

where q_i and s_i denote the daily observed and simulated flows at time i, respectively; Ω is the parameter vector being estimated; and m is the total number of days used for calibration. Three years (2003-2005) of streamflow data are used to calibrate the HL-RDHM for the selected basins. The first year (year 2003) is used to warm-up HL-RDHM. To assess the model performance during calibration, we use the percent bias (PB), modified correlation coefficient (R_m), and Nash-Sutcliffe efficiency (NSE) (see appendix for details). Note that these metrics are used during the manual phase of the calibration process, and to assess the final results from the implementation of the SLS. However, the actual implementation of the SLS is based on the objective function in Eq. (1).

3.2 Statistical weather preprocessor

Heteroscedastic censored logistic regression (HCLR) (Messner et al., 2014a; Yang et al., 2017) is implemented to preprocess the ensemble precipitation forecasts from the GEFSRv2. HCLR is selected since it offers the advantage, over other regression-based preprocessors (Wilks, 2009), of obtaining the full, continuous predictive probability density function (pdf) of precipitation forecasts (Messner et al., 2014b). Also, HCLR has been shown to outperform other widely used preprocessors (Yang et al., 2017). In principle, HCLR fits the conditional logistic probability distribution function to the transformed (here the square root) ensemble mean and bias corrected precipitation ensembles. Note that we tried different transformations (square root, cube root, and fourth root), and found a similar performance between the square and cube root, both outperforming the fourth root. In addition, HCLR uses the ensemble spread as a predictor, which allows the use of uncertainty information contained in the ensembles.

The development of the HCLR follows the logistic regression model initially proposed by Hamill et al. (2004) as well as the extended version of that model proposed by Wilks (2009). The extended logistic regression of Wilks (2009) is used to model the probability of binary responses such that

$$P(y \le z \mid x) = \Lambda[\omega(z) - \delta(x)], \tag{2}$$

where $\Lambda(.)$ denotes the cumulative distribution function of the standard logistic distribution, y is the transformed precipitation, z is a specified threshold, x is a predictor variable that depends on the forecast members, $\delta(x)$ is a linear function of the predictor variable x, and the transformation $\omega(.)$ is a monotone nondecreasing function. Messner et al. (2014a) proposed the heteroscedastic extended logistic regression (HELR) preprocessor with an additional predictor variable φ to control the dispersion of the logistic predictive distribution,

$$P(y \le z \mid x) = \Lambda\left\{\frac{\omega(z) - \delta(x)}{\exp[\eta(\varphi)]}\right\}, \tag{3}$$

where $\eta(.)$ is a linear function of φ. The functions $\delta(.)$ and $\eta(.)$ are defined as:

$$\delta(x) = a_0 + a_1 x, \text{ and} \tag{4}$$

$$\eta(\varphi) = b_0 + b_1 \varphi, \tag{5}$$

where a_0, a_1, b_0, and b_1 are parameters that need to be estimated; $x = 1/K \sum_{k=1}^{K} f_k^{1/2}$, i.e., the predictor variable x is the mean of the transformed, via the square root, ensemble forecasts f; K is the total number of ensemble members; and φ is the standard deviation of the square root transformed, precipitation ensemble forecasts.

To estimate the parameters associated with Eq. (3), maximum likelihood estimation with the log-likelihood function is used (Messner et al., 2014a; Messner et al., 2014b). For this, one needs to determine the predicted probability π_i of the ith observed outcome. One variation of the HELR postprocessor that can straightforwardly accommodate nonnegative variables that are continuous for positive values and have a natural threshold at zero, such as precipitation amounts, is censored regression or, as termed by Messner et al. (2014a), HCLR. For HCLR, π_i can be expressed as (Messner et al., 2014a)

$$\pi_i = \begin{cases} \Lambda\left[\dfrac{\omega(0) - \delta(x)}{\exp[\eta(\varphi)]}\right] & y_i = 0 \\ \lambda\left[\dfrac{\omega(y_i) - \delta(x)}{\exp[\eta(\varphi)]}\right] & y_i > 0, \end{cases} \tag{6}$$

where $\lambda[.]$ denotes the likelihood function of the standard logistic function. As indicated by equation (6), HCLR fits a logistic error distribution with point mass at zero to the transformed predictand.

HCLR is applied here to each GEFSRv2 grid cell within the selected basins. At each cell, HCLR is implemented for the period 2004-2012 using a leave-one-out approach. For this, we select 7 years for training and the two remaining years for verification purposes. This is repeated until all the 9 years have been preprocessed and verified independently of the training period. This is done so that no training data is discarded and the entire 9-year period of analysis can be used to generate the precipitation forecasts. HCLR is employed for 6-hourly precipitation accumulations for lead times from 6 to 168 hours. To train the preprocessor, we use a stationary training period, as opposed to a moving window, for each season and year to be forecasted, comprised by the seasonal data from all the 7 training years. Thus, to forecast a given season and

specific lead time, we use ~6930 forecasts (i.e., 11 members x 90 days per season x 7 years). We previously tested using a moving window training approach and found that the results were similar to the stationary window one (Yang et al., 2017). To make the implementation of HCLR as straightforward as possible, the stationary window is used here. Finally, the Schaake Shuffle method as applied by Clark et al. (2004) is implemented to maintain the observed space-time variability in the preprocessed GEFSRv2 precipitation forecasts. At each individual forecast time, the Schaake Shuffle is applied to produce a spatial and temporal rank structure for the ensemble precipitation values that is consistent with the ranks of the observations.

3.3 Statistical streamflow postprocessors

To statistically postprocess the flow forecasts generated by the RHEPS, two different approaches are tested, namely a first-order autoregressive model with a single exogenous variable, ARX(1,1), and quantile regression (QR). We select the ARX(1,1) postprocessor since it has been suggested and implemented for operational applications in the U.S. (Regonda et al., 2013). QR is chosen because it is of similar complexity as the ARX(1,1) postprocessor but for some forecasting conditions it has been shown to outperform it (Mendoza et al., 2016). Furthermore, the ARX (1,1) and QR postprocessors have not been compared against each other for the forecasting conditions specified by the RHEPS. The postprocessors are implemented for the years 2004-2012, using the same leave-one-out approach used for the preprocessor. The postprocessors are applied at each individual lead time from day 1 to 7. For this, the 6-houlry streamflow forecasts from HL-RDHM are averaged over 24 hours to get the streamflow forecast for a particular day.

3.3.1 First-order autoregressive model with a single exogenous variable

To implement the ARX(1,1) postprocessor, the observation and forecast data are first transformed into standard normal deviates using the normal quantile transformation (NQT) (Krzysztofowicz, 1997; Bogner et al., 2012). The transformed observations and forecasts are then used as predictors in the ARX(1,1) model (Siddique and Mejia, 2017). Specifically, for each forecast lead time, the ARX (1,1) postprocessor is formulated as follows:

$$q_{i+1}^T = (1-c_{i+1})q_i^T + c_{i+1} f_{i+1}^T + \xi_{i+1}, \tag{7}$$

where q_i^T and q_{i+1}^T are the NQT transformed observed flows at time steps i and i+1, respectively; c is the regression coefficient; f_{i+1}^T is the NQT transformed forecast flow at time step i+1; and ξ is the residual error term. In Eq. (7), assuming that there is significant correlation between ξ_{i+1} and q_i^T, ξ_{i+1} can be calculated as:

$$\xi_{i+1} = \frac{\sigma_{\xi_{i+1}}}{\sigma_{\xi_i}} \rho(\xi_{i+1}, \xi_i)\xi_i + \vartheta_{i+1}, \tag{8}$$

where σ_{ξ_i} and $\sigma_{\xi_{i+1}}$ are the standard deviation of ξ_i and ξ_{i+1}, respectively; $\rho(\xi_{i+1}, \xi_i)$ is the serial correlation between ξ_{i+1} and ξ_i; and ϑ_{i+1} is a random Gaussian error generated from $¥(0, \sigma_{\vartheta_{i+1}}^2)$. To estimate $\sigma_{\eta_{i+1}}^2$, the following equation is used:

$$\sigma_{\vartheta_{i+1}}^2 = [1-\rho^2(\xi_{i+1}, \xi_i)]\sigma_{\xi_{i+1}}^2. \tag{9}$$

To implement Eq. (7), ten equally spaced values of c_{i+1} are selected from 0.1 to 0.9. For each value of c_{i+1}, $\sigma_{\vartheta_{i+1}}^2$ is determined from Eq. (9) using the training data to determine the other

variables in Eq. (9). Then, ϑ_{i+1} is generated from $¥(0,\sigma^2_{\vartheta_{i+1}})$ and ξ_{i+1} is calculated from Eq. (8). The result from Eq. (8) is used with Eq. (7) to generate a trace of q^T_{i+1} which is transformed back to real space using the inverse NQT. These steps are repeated to generate multiple traces for each value of c_{i+1}. Lastly, the value of c_{i+1} that produces the ensemble forecast with the smallest mean continuous ranked probability skill (CRPS) is selected. The ARX (1,1) postprocessor is applied at each individual lead time. For lead times beyond the initial one (day 1), one day-ahead predictions are used as the observed streamflow. For the cases where q^T_{i+1} falls beyond the historical maxima, extrapolation is used by modeling the upper tail of the forecast distribution as hyperbolic (Journel and Huijbregts, 1978).

3.3.2 Quantile regression

Quantile regression (QR; Koenker and Bassett Jr, 1978; Koenker, 2005) is employed to determine the error distribution, conditional on the ensemble mean, resulting from the difference between observations and forecasts (Dogulu et al., 2015; López et al., 2014; Weerts et al., 2011; Mendoza et al., 2016). QR is applied here in streamflow space, since it has been shown that, in hydrological forecasting applications, QR has similar skill performance in streamflow and normal space (López et al., 2014). Another advantage of QR is that it does not make any prior assumptions regarding the shape of the distribution. Further, since QR results in conditional quantiles rather than conditional means, QR is less sensitive to the tail behavior of the streamflow dataset, and consequently, less sensitive to outliers. Note that although QR is here implemented separately for each lead time, the mathematical notation does not reflect this for simplicity.

The QR model is given by

$$\varepsilon'_\tau = d_\tau + e_\tau \overline{f}, \tag{10}$$

where ε'_τ is the error estimate at quantile interval τ; $\overline{f}$ is the ensemble mean; and d_τ and e_τ are the linear regression coefficients at τ. The coefficients are determined by minimizing the sum of the residuals based on the training data as follows:

$$\min \sum_{i=1}^{N} w_\tau [\varepsilon_{\tau,i} - \varepsilon'_\tau(i, \overline{f}_i)], \tag{11}$$

$\varepsilon_{\tau,i}$ and $\overline{f}_i$ are the i^{th} paired samples from a total of N samples; $\varepsilon_{\tau,i}$ is computed as the observed flow minus the forecasted one, $q_\tau - f_\tau$; and w_τ is the weighting function for the τ^{th} quantile defined as:

$$w_\tau(\varsigma_i) = \begin{cases} (\tau-1)\varsigma_i & \text{if } \varsigma_i \le 0 \\ \tau\varsigma_i & \text{if } \varsigma_i > 0 \end{cases}. \tag{12}$$

ς_i is the residual term defined as the difference between $\varepsilon_{\tau,i}$ and $\varepsilon'_\tau(i, \overline{f}_i)$ for the quantile τ. The minimization in Eq. (11) is solved using linear programming (Koenker, 2005).

Lastly, to obtain the calibrated forecast f_τ, the following equation is used:

$$f_\tau = \overline{f} + \varepsilon'_\tau. \tag{13}$$

In Eq. (13), the estimated error quantiles and the ensemble mean are added to form a calibrated discrete quantile relationship for a particular forecast lead time and thus generate an ensemble streamflow forecast.

3.4 Forecast experiments and verification

The verification analysis is carried out using the Ensemble Verification System (Brown et al., 2010). For the verification, the following metrics are considered: relative mean error (RME), Brier skill score (BSS), mean continuous ranked probability skill score (CRPSS), and the decomposed components of the CRPS (Hersbach, 2000), i.e., the CRPS reliability ($CRPS_{rel}$) and CRPS potential ($CRPS_{pot}$). The definition of each of these metrics is provided in the appendix. Additional details about the verification metrics can be found elsewhere (Wilks, 2011;Jolliffe and Stephenson, 2012). Confidence intervals for the verification metrics are determined using the stationary block bootstrap technique (Politis and Romano, 1994), as done by Siddique et al. (2015). The verification is focused on flood events by choosing flow amounts greater than that implied by a non-exceedance probability, in the sampled climatological probability distribution, of ~0.95. Thus, hereafter the term floods is used instead of streamflow to denote the forecasts generated by HL-RDHM. All the forecast verifications are done for lead times from 1 to 7 days.

To verify the forecasts for the period 2004-2012, six different forecasting scenarios are considered (Table 2). The first (S1) and second (S2) scenarios verify the raw and preprocessed ensemble precipitation forecasts, respectively. Scenarios 3 (S3), 4 (S4) and 5 (S5) verify the raw, preprocessed, and postprocessed ensemble flood forecasts, respectively. The last scenario, S6, verifies the combined preprocessed and postprocessed ensemble flood forecasts. In S1 and S2, the raw and preprocessed ensemble precipitation forecasts are verified against the MPEs. For the verification of S1 and S2, each grid cell is treated as a separate verification unit. Thus, for a particular basin, the average performance is obtained by averaging the verification results from different verification units. The flood forecast scenarios, S3-S6, are verified against daily streamflow observations from the USGS. The quality of the flood forecasts is evaluated conditionally upon forecast lead time, season (cool and warm), and flow threshold.

4. Results and discussion

This section is divided into four subsections. The first subsection demonstrates the performance of the spatially distributed model, HL-RDHM. The second subsection describes the performance of the raw and preprocessed GEFSRv2 ensemble precipitation forecasts (forecasting scenarios S1 and S2). In the third subsection, the two statistical postprocessing techniques are compared. Lastly, the verification of different ensemble flood forecasting scenarios is shown in the fourth subsection (forecasting scenarios S3-S6).

4.1 Performance of the distributed hydrological model

To assess the performance of HL-RDHM, the model is used to generate streamflow simulations which are verified against daily observed flows, covering the entire period of analysis (years 2004-2012). Note that the simulated flows are obtained by forcing HL-RDHM with gridded precipitation and near surface temperature observations. The verification is done for the four basin outlets shown in Fig. 1. To perform the verification and assess the quality of the streamflow simulations, the following statistical measures of performance are employed: modified correlation coefficient, R_m; Nash-Sutcliffe efficiency, NSE; and percent bias, PB. The mathematical definition of these metrics is provided in the appendix. The verification is done for both uncalibrated and calibrated simulation runs for the entire period of analysis. The main results from the verification of the streamflow simulations are summarized in Fig. 2.

The performance of the calibrated simulation runs is satisfactory, with R_m values ranging from ~0.75 to 0.85 (Fig. 2a). Likewise, the NSE, which is sensitive to both the correlation and bias, ranges from ~0.69 to 0.82 for the calibrated runs (Fig. 2b), while the PB ranges from ~5 to -

11% (Fig. 2c). Relative to the uncalibrated runs, the R_m, NSE, and PB values improve by ~18, 29, and 47%, respectively. Further, the performance of the calibrated simulation runs is similar across the four selected basins, although the largest size basin, WVYN6 (Fig. 2), seems to slightly outperform the other basins with R_m, NSE, and PB values of 0.85, 0.82, and -3% (Fig. 2), respectively. The lowest performance is seen in CNON6 with R_m, NSE, and PB values of 0.75, 0.7, and -11% (Fig. 2), respectively. Nonetheless, the performance metrics for both the uncalibrated and calibrated simulation runs do not deviate widely from each other in the selected basins, with perhaps the only exception being PB (Fig. 2c).

As part of the calibration process, we adjusted 10 out the 17 SAC-HT parameters associated with each model grid cell. Note that we adjusted the parameter fields rather than the actual parameter values. The adjusted parameters were associated with baseflow, percolation, evaporation, storm runoff, and the channel routing process. The most sensitive parameters were found to be the lower zone supplemental withdrawal rate (LZSK), upper zone free water maximum storage (UZFWM), and the channel routing parameters. In addition, the performance of the model was lower during the cool season. By comparing the performance of individual hydrographs during the cool season, it was observed that high flow events are consistently underestimated with the a-priori parameter set. To improve this, the SNOW-17 parameters were also adjusted. The adjusted SNOW-17 parameters included the maximum negative melt factor (NMF), temperature threshold that separates rain from snow (PXTMP), and the snow fall correction factor (SCF). Adjusting these parameters improve the performance some but, in the future, snow data when available could be used to directly assess the modeling of the snow dynamics. Nonetheless, the performance of the HL-RDHM at the selected outlet locations is reasonably good.

4.2 Verification of the raw and preprocessed ensemble precipitation forecasts

To examine the skill of both the raw and preprocessed GEFSRv2 ensemble precipitation forecasts, we plot in Fig. 3 the CRPSS (relative to sampled climatology) as a function of the forecast lead time (day 1 to 7) and season for the selected basins. Two seasons are considered: cool (October-March) and warm (April-September). Note that a CRPSS value of zero means no skill (i.e., same skill as the reference system) and a value of one indicates maximum skill. The CRPSS is computed using 6 hourly precipitation accumulations and high precipitation events. High precipitation events are here defined by an amount greater than that implied by a non-exceedance probability, in the sampled climatological probability distribution, of ~0.95.

The skill of both the raw and preprocessed ensemble precipitation forecasts tends to decline with increasing forecast lead time (Fig. 3). In the warm season (Figs. 3a-d), the CRPSS values vary overall, across all the basins, in the range from ~0 to 0.4 and from ~-0.2 to 0.3 for the preprocessed and raw forecasts, respectively; while in the cool season (Figs. 3e-h) the CRPSS values vary overall in the range from ~0.1 to 0.6 and from 0 to 0.5 for the preprocessed and raw forecasts, respectively. The skill of the preprocessed ensemble precipitation forecasts tends to be greater than the raw ones across basins, seasons, and forecast lead times. Comparing the raw and preprocessed forecasts against each other, the relative skill gains from preprocessing are somewhat more apparent in the medium-range lead times (>3 days) and warm season. That is, the differences in skill seem not as significant in the short-range lead times (≤3 days). This seems particularly the case in the cool season where the confidence intervals for the raw and preprocessed forecasts tend to overlap.

Indeed, seasonal skill variations are noticeable in all the basins. Even though the relative gain in skill from preprocessing is slightly greater in the warm season, the overall skill of both the raw and preprocessed forecasts is better in the cool season than the warm one. This may be due, among other potential factors, to the greater uncertainty associated with modeling convective precipitation, which is more prevalent in the warm season, by the NWP model used to generate the GEFSRv2 outputs (Hamill et al., 2013; Baxter et al., 2014). Nonetheless, the warm season preprocessed forecasts show gains in skill across all the lead times and basins. For a particular season, the forecast ensembles across the different basins tend to display similar performance; i.e. the analysis does not reflect skill sensitivity to the basin size as in other studies (Siddique et al., 2015; Sharma et al., 2017). This is expected here since the verification is performed for each GEFSRv2 grid cell, rather than verifying the average for the entire basin. That is, the results in Fig. 3 are for the average skill performance obtained from verifying each individual grid cell within the selected basins.

Based on the results presented in Fig. 3, we may expect some skill contribution to the flood ensembles from forcing the HL-RDHM with the preprocessed precipitation, as opposed to using the raw forecast forcing. Although the contribution may not be as large, since the differences between the preprocessed and raw precipitation forecasts are only mild. It may also be expected that the contributions are greater for the medium-range lead times and warm season. This will be examined in subsection 4.4, prior to that we compare next the two postprocessors, namely ARX(1,1) and QR.

4.3 Selection of the flood postprocessor

The ability of the ARX(1,1) and QR postprocessors to improve ensemble flood forecasts is investigated here. The postprocessors are applied to the raw flood ensembles at each forecast lead time from day 1 to 7. To examine the skill of the postprocessed flood forecasts, Fig. 4 displays the CRPSS (relative to the raw ensemble flood forecasts) versus the forecast lead time for all the selected basins, for both cool (Fig. 4a-d) and warm (Fig. 4e-h) seasons. The overall tendency is for both postprocessing techniques to demonstrate improved forecast skill across all the basins, seasons, and most of the lead times. The skill can improve as much as 40% at the later lead times (Fig. 4b). The general trend in Fig. 4 is for the skill of the postprocessors to increase with increasing lead time. Note that this is the case since the skill is here measured relative to the raw flood forecasts which is done to better isolate the effect of the postprocessors on the flood forecasts. This means that the postprocessors are more able to improve the medium-range (>3 days) forecasts than the short-range (≤3 days) ones.

The gains in skill from QR vary from ~5% (Fig. 4a at the day 1 lead time) to 40% (Fig. 4b at the day 5 lead time) depending upon the season and lead time. While the gains from ARX(1,1) vary from ~4% (Fig. 4e at the day 1 lead time) to a much lower level of ~22% (Fig. 4c at the day 2 lead time). In most cases, both postprocessors exhibit somewhat similar performance at the initial lead times (days 1-2), with skills varying from nearly 0.1 (e.g., Figs. 4a and 4e) to 0.4 (Fig. 4f at the day 2 lead time). At the later lead times (4-7 days), QR tends to outperform ARX(1,1), with the difference in performance being as high as 30% (Fig. 4d at the day 7 lead time). This is noticeable across all the basins and for both seasons. The skill improvement of QR over ARX(1,1) is significant at later lead times (> day 3), as indicated by the fact that the confidence intervals for the curves representing the postprocessors in Fig. 4 often do not overlap. There are also seasonal differences in the performance of the postprocessors. In particular, the gains in skill from ARX(1,1) in the warm season can be quite low (Figs. 4a and 4c).

As discussed and demonstrated in Fig. 4, QR performs better than ARX(1,1). Indeed, we also found (plots not shown) that QR displays better reliability than ARX(1,1) across lead times, basins, and seasons. Therefore, we select QR as the statistical flood postprocessor to examine the interplay between preprocessing and postprocessing in the RHEPS.

4.4 Verification of the ensemble flood forecasts for different statistical processing scenarios

In this subsection, we examine the effects of different statistical processing scenarios on the ensemble flood forecasts from the RHEPS. Recall that, to consider flood events, the verification is done for flow events with an amount greater than that implied by a non-exceedance probability, in the sampled climatological probability distribution, of ~0.95. The forecasting scenarios considered here are S3-S6 (Table 1 defines the scenarios). To facilitate presenting the verification results, this subsection is divided into the following four parts: relative mean error, CRPSS, CRPS decomposition, and BSS.

4.4.1 Relative mean error

To examine the bias associated with the mean ensemble flood forecasts under scenarios S3-S6, we plot the RME versus the forecast lead time for all the basins (Fig. 5), and the warm (Fig. 5a-d) and cool seasons (Fig. 5e-h). Results in Fig. 5 show that, under all the considered scenarios, the mean ensemble flood forecasts exhibit underforecasting bias across basins, lead times, and seasons. The underforecasting bias increases with the lead time, and decreases somewhat with the increase in basin size. For example, the bias for the largest basin, WVYN6, is -0.1 at the day 1 lead time and scenario S3 (Fig. 5d), while for the same lead time and scenario the bias is -0.35 for the smallest basin (Fig. 5a). In essence, the GEFSRv2-based flood ensembles exhibit a conditional bias that is consistent with the conditional bias (i.e., to significantly underforecast large events) for the GEFSRv2 precipitation ensembles (Siddique et al., 2015; Sharma et al., 2017).

The two most striking features of Fig. 5 are: i) the significant difference in performance between the pair S3-S4 and S5-S6 and, in contrast, ii) the similarity in performance between S5 and S6. The former confirms that statistical processing, in particular postprocessing, has a significant effect on the flood ensembles. Recall that to generate the ensemble flood forecasts S5 only employs postprocessing, while S6 considers both preprocessing and postprocessing (Table 1). Yet, the RME across basins, lead times, and seasons for both S5 and S6 are quite similar, with differences tending to be not as significant. The similarity between S5 and S6 indicates that in this case preprocessing has a mild effect on the flood forecasts.

As a corollary to the latter comment, it can be argued that by only postprocessing the raw flood ensembles most of the benefits from statistical processing can be realized. This seems also supported by the results for S3 and S4 (Fig. 5). The differences between the RME of the flood forecasts generated by forcing the HL-RDHM with raw, S3, versus preprocessed precipitation ensembles, S4, are only significant at lead times greater than 4 days. In addition, the differences are not as large, with the largest one being ~-0.18 at the day 5 lead time in CNON6 (Fig. 5b). This is not entirely surprising as we previously saw (Fig. 3) that differences between the raw and preprocessed precipitation ensembles are only significant at the later lead times where the skill of the forecast is, in any case, already somewhat low. In terms of the seasonal analysis, both S5 and S6 tend to be less biased in the cool season than in the warm one, particularly at the short-range lead times (<3 days). This can be seen by comparing Fig. 4c against Fig. 4g at the day 1 lead

time. The role played by preprocessing and postprocessing in ensemble flood forecasting is further evaluated next in terms of the forecast skill.

4.4.2 CRPSS

The skill of the ensemble flood forecasts for S3-S6 is assessed using the CRPSS relative to the sampled climatology (Fig. 6). Fig. 6 shows that, across lead times, basin sizes, and seasons, the results for the CRPPS are qualitatively similar to those for the RME (Fig. 5). That is, the most salient feature of Fig. 6 is that the performance of the flood forecasts tends to progressively improve from S3 to S6. This means that the forecast skill tends to improve across lead times, basin sizes, and seasons as additional statistical processing steps are included in the RHEPS' forecasting chain. The skill first increases from the raw scenario (i.e., S3 where no statistical processing is done) to the scenario where only preprocessing is performed, S4. However, the gain in skill between S3 and S4 is generally small, particularly at the short lead times, reinforcing the fact that preprocessing may have little effect on the flood forecasts. The skill then shows a more significant improvement for both S5 and S6, relative to S4. As was the case with the RME, the differences in skill between S5 and S6 are not as significant, suggesting that postprocessing alone (i.e., without preprocessing) may be sufficient to remove systematic biases in the flood forecasts.

In terms of the warm and cool seasons, at the initial forecast lead times ($\leq$ 2 days), the skill of the flood forecasts tends to be slightly greater in the cool season (Figs. 6e-h) than in the warm one (Figs. 6a-d), with the exception of CNON6. As was the case in the calibration results (e.g., in Fig. 2c), during the cool season CNON6 has a lower performance prior to postprocessing (S3 or S4 in Fig. 6f) than the other basins. Interestingly, after postprocessing (S5 in Fig. 6f), the skill of CNON6 is as good as that of CINN6, even though at the day 1 lead time the skill for S3 is ~0.3 for CNON6 (Fig. 6f) and ~0.5 for CINN6 (Fig. 6e). Hence, the postprocessor seems capable to compensate some for the lesser performance of CNON6 during calibration.

4.4.3 CRPS decomposition

Fig. 7 displays different components of the mean CRPS against lead times of 1, 3, and 7 days for all the basins according to both the warm (Figs. 7a-d) and cool (Figs. 7e-h) seasons. The components presented here are reliability ($CRPS_{rel}$) and potential CRPS ($CRPS_{pot}$) (Hersbach, 2000). $CRPS_{rel}$ measures the average reliability of the ensemble forecasts across all the possible events, i.e., it examines whether the fraction of observations that fall below the *j*-th of *n* ranked ensemble members is equal to *j*/*n* on average. $CRPS_{pot}$ represents the lowest possible CRPS that could be obtained if the forecasts were made perfectly reliable (i.e., $CRPS_{rel}$=0). Note that the CRPS, $CRPS_{rel}$, and $CRPS_{pot}$ are all negatively oriented, with perfect score of zero. Overall, as was the case with the RME (Fig. 5) and CRPSS (Fig. 6), the CRPS decomposition reveals that forecasts reliability increases from S3 to S6.

Interestingly, improvements in forecast quality for S5 and S6, relative to the raw flood forecasts of S3, are mainly due to reductions in $CRPS_{rel}$ (i.e., by making the forecasts more reliable), whereas for S4 better forecast quality is achieved by reductions in $CRPS_{pot}$. The latter is seen across all basins, lead times, and seasons. The explanation for this lies in the implementation of the HCLR preprocessor, which uses the ensemble spread as a predictor of the dispersion of the predictive pdf and the $CRPS_{pot}$ is sensitive to the spread (Messner et al., 2014a). Although the forecasts from S3 have lower $CRPS_{pot}$, the forecasts including postprocessing, S5

and S6, ultimately result in lower CRPS. This indicates that the forecasts for S5 and S6 are more reliable than for S3 and S4.

4.4.4 BSS

In our final verification comparison, the BSS of the ensemble flood forecasts for S5 (Figs. 8a-d) and S6 (Figs. 8e-h) are plotted against the non-exceedance probability associated with different flood thresholds ranging from 0.95 to 0.99. The BSS is computed for all the basins, warm season, and lead times of 1, 3 and 7 days. In addition, the BSS is computed relative to both observed (solid lines in Fig. 8) and simulated (dashed lines in Fig. 8) floods. When the BSS is computed relative to observed floods, it considers the effect on forecast skill of both meteorological and hydrological uncertainties. While the BSS relative to simulated floods is mainly affected by meteorological uncertainties. The difference between the two, i.e., the BSS relative to observed floods minus the BSS relative to simulated ones, provides an estimate of the effect of hydrological uncertainties on the skill of the flood forecasts. Similar to the CRPSS, the BSS value of zero means no skill (i.e., same skill as the reference system) and a value of one indicates perfect skill.

In general, the skill of flood forecasts tends to decrease with lead time across the flow thresholds and basins. As was the case with the CRPSS (Fig. 6), the BSS values appear similar for S5 (Figs. 8a-d) and S6 (Figs. 8e-h). The only exception is CKLN6 (Figs. 8c and 8g) where, at the higher flood thresholds, S6 has better skill than S5 at the day 1 and 3 lead times. With respect to the basin size, the skill tends to improve some from the small to the large basin. For instance, for non-exceedance probabilities of 0.95 and 0.99 at the day 1 lead time, the BSS values for the smallest basin (Fig. 8a), measured relative to the observed flows, are ~0.49 and 0.35, respectively. For the same conditions, both values increase to ~0.65 for the largest basin (Fig. 8d).

Indeed, the most notable feature in Fig. 8 is that the effect of hydrological uncertainties on forecast skill is evident at the day 1 lead time, while meteorological uncertainties clearly dominate at the day 7 lead time. With respect to the latter, notice that the solid and dashed green lines for the day 7 lead time tend to be very close to each other in Fig. 8, indicating that hydrological uncertainties are relatively small compared to meteorological ones. Hydrological uncertainties are largest at the day 1 lead time, particularly for the small basins (Figs. 8a-b and 8e-f). For example, for a non-exceedance probability of 0.95 and at a day 1 lead time (Fig. 8b), the BSS value relative to the simulated and observed floods are ~0.79 and 0.38, respectively, suggesting a reduction of ~50% skill due to hydrological uncertainties.

5. Summary and conclusion

In this study, we used the RHEPS to investigate the effect of statistical processing on short- to medium-range ensemble flood forecasts. First, we assessed the raw precipitation forecasts from the GEFSRv2 (S1), and compared them with the preprocessed precipitation ensembles (S2). Then, flood ensembles were generated with the RHEPS for four different forecasting scenarios involving no statistical processing (S3), preprocessing alone (S4), postprocessing alone (S5), and both preprocessing and postprocessing (S6). The verification of ensemble precipitation and flood forecasts was done for the years 2004-2012, using four basins in the U.S. MAR. We found that – for the models, datasets, and study domain used here - the skill gains from joint preprocessing and postprocessing are similar to those from postprocessing alone. Other specific findings are as follows:

- The HCLR preprocessed ensemble precipitation forecasts show improved skill relative to the raw forecasts. The improvements are more noticeable in the warm season at the longer lead times (>3 days).
- Both postprocessors, ARX(1,1) and QR, show gains in skill relative to the raw ensemble flood forecasts. For the medium-range lead times (>3 days), the gains with QR, however, tend to be greater than with ARX(1,1), particularly during the warm season.
- By comparing different statistical processing scenarios for the ensemble flood forecasts, it was found that the scenario with preprocessing alone has little effect on improving the skill of the flood forecasts in contrast with the postprocessing alone scenario.
- The scenario including only postprocessing performs similar, in terms of the relative mean error, CRPSS, and reliability, to the more complex scenario consisting of both preprocessing and postprocessing. It thus seems for our conditions, using GEFSRv2 forecasts, that preprocessing may be unnecessary.
- The skill of the postprocessing alone scenario and the scenario that combines preprocessing and postprocessing was further assessed using the Brier skill score for different flood thresholds. This assessment further confirmed that both scenarios have similar skill and performance behavior.

These conclusions are specific to the RHEPS forecasting system, which is mostly relevant to the U.S. research and operational communities as it relies on a weather and a hydrological model that are used in this domain. However, the use of a global weather forecasting system illustrates the potential of applying the statistical techniques tested here in other regions worldwide.

The emphasis of this study has been on benchmarking the contributions of statistical processing to the RHEPS. To accomplish this, our approach required that the quality of ensemble flood forecasts be verified over multiple years (i.e., across many flood cases) to obtain robust verification statistics. Future research, however, could be focused on studying how distinct hydrological processes contribute or constrain forecast quality. This effort could be centered around specific flood events rather than in the statistical, many-cases approach taken here. To further assess the relative importance of the various components of the RHEPS, additional tests involving the uncertainty to initial hydrologic conditions and hydrological parameters could be performed. For instance, the combined use of data assimilation and postprocessing has been shown to produce more reliable and sharper streamflow forecasts (Bourgin et al., 2014). The potential for the interaction of preprocessing and postprocessing with data assimilation to significantly enhance streamflow predictions, however, has not been investigated. This could be investigated in the future with the RHEPS, as the pairing of data assimilation with preprocessing and postprocessing could facilitate translating the improvements in the preprocessed meteorological forcing down the hydrological forecasting chain.

Data availability: Daily streamflow observation data for the selected forecast stations can be obtained from the USGS website (https://waterdata.usgs.gov/nwis/). Multisensor precipitation estimates are obtained from the NOAA's Middle Atlantic River Forecast Center. Precipitation and temperature forecast datasets can be obtained from the NOAA Earth System Research Laboratory website (https://www.esrl.noaa.gov/psd/forecasts/reforecast2/download.html).

Appendix A: Verification metrics

Modified correlation coefficient (R_m):

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 flow is introduced to refine the conventional correlation coefficient R. The modified correlation coefficient R_m is defined as:

$$R_m = R \frac{\min\{\sigma_s, \sigma_q\}}{\max\{\sigma_s, \sigma_q\}}, \quad \text{(A1)}$$

where σ_s and σ_q denote the standard deviation of the simulated and observed flows, respectively.

Percent bias (PB):

PB measures the average tendency of the simulated flows to be larger or smaller than their observed counterparts. Its optimal value is 0.0 where positive values indicate model overestimation bias, and negative values indicate model underestimation bias. The PB is estimated as follows:

$$PB = \frac{\sum_{i=1}^{N}(s_i - q_i)}{\sum_{i=1}^{N} q_i} \times 100, \quad \text{(A2)}$$

where s_i and q_i denote the simulated and observed flow, respectively, at time *i*.

Nash-Sutcliffe efficiency (NSE):

The NSE (Nash and Sutcliffe, 1970) is defined as the ratio of the residual variance to the initial variance. It is widely used to indicate how well the simulated flows fit the observations. The range of NSE can vary between negative infinity to 1.0, with 1.0 representing the optimal value and values should be larger than 0.0 to indicate minimally acceptable performance. The NSE is computed as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(s_i - q_i)^2}{\sum_{i=1}^{N}(q_i - \bar{q}_i)^2} \quad \text{(A3)}$$

where s_i, q_i, and $\bar{q}_i$ are the simulated, observed, and mean observed flow, respectively, at time *i*.

Relative mean error (RME):

RME quantifies the average error between the ensemble mean forecast and their corresponding observation 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:

$$RME = \frac{\sum_{i=1}^{n}(\bar{f}_i - q_i)}{\sum_{i=1}^{n} q_i}, \quad \text{(A4)}$$

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

Brier Skill Score (BSS):

The Brier score (BS; Brier, 1950) is analogous to the mean squared error, but where the forecast is a probability and the observation is either a 0.0 or 1.0 (Brown and Seo 2010). The BS is given by

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}[F_{f_i}(z) - F_{q_i}(z)]^2, \tag{A5}$$

where the probability of f_i to exceed a fixed threshold z is

$$F_{f_i}(z) = \mathrm{P_r}[f_i > z], \tag{A6}$$

n is again the total number of forecast-observation pairs, and

$$F_{q_i}(z) = \begin{cases} 1, & q_i > z; \\ 0, & \text{otherwise.} \end{cases} \tag{A7}$$

In order 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):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS_{main}}}{\mathrm{BS_{reference}}}, \tag{A8}$$

where $\mathrm{BS_{main}}$ and $\mathrm{BS_{reference}}$ 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.

Mean Continuous Ranked Probability Skill Score (CRPSS):

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, $F_f(z)$, and the corresponding cdf of the observation, $F_q(z)$. The CRPS is given by

$$\mathrm{CRPS} = \int_{-\infty}^{\infty}\left[F_f(z) - F_q(z)\right]^2 dz. \tag{A9}$$

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:

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS_{main}}}{\mathrm{CRPS_{reference}}}, \tag{A10}$$

where the CRPS is averaged across n pairs of forecasts and observations to calculate the mean CRPS of the main forecast system ($\mathrm{CRPS_{main}}$) and reference forecast system ($\mathrm{CRPS_{reference}}$). The CRPSS ranges from -∞ 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.

In addition, to further explore the effect of postprocessing on forecast skill, we separate the $\mathrm{CRPS_{main}}$ into different components according to the procedure developed by Hersbach (2000). Specifically, we consider the CRPS reliability ($\mathrm{CRPS_{rel}}$) and potential ($\mathrm{CRPS_{pot}}$) such that

$$\mathrm{CRPS_{main}} = \mathrm{CRPS_{rel}} + \mathrm{CRPS_{pot}}. \tag{A11}$$

The $\mathrm{CRPS_{rel}}$ measures the ability of the precipitation ensembles to generate cumulative distributions that have, on average, the correct or desired statistical properties. The reliability is

closely connected to the rank histogram, which shows whether the frequency that the verifying analysis was found in a given bin is equal for all bins (Hersbach 2000). The $CRPS_{pot}$ measures the CRPS that one would obtain for a perfect reliable system. It is sensitive to the average spread of the ensemble and outliers. For instance, the narrower the spread of the ensemble is, the smaller the $CRPS_{pot}$ becomes. As indicated by Hersbach (2000), provided a certain degree of unpredictability, a balance between the ensemble spread and the statistics of outliers will result in the optimal value of the $CRPS_{pot}$.

References

Addor, N., Jaun, S., Fundel, F., and Zappa, M.: An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios, Hydrology and Earth System Sciences, 15, 2327, 2011.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., and Pappenberger, F.: GloFAS-global ensemble streamflow forecasting and flood early warning, Hydrology and Earth System Sciences, 17, 1161, 2013.

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., and Salamon, P.: Evaluation of ensemble streamflow predictions in Europe, Journal of Hydrology, 517, 913-922, 2014.

Anderson, R. M., Koren, V. I., and Reed, S. M.: Using SSURGO data to improve Sacramento Model a priori parameter estimates, Journal of Hydrology, 320, 103-116, 2006.

Baxter, M. A., Lackmann, G. M., Mahoney, K. M., Workoff, T. E., & Hamill, T. M.: Verification of quantitative precipitation reforecasts over the southeastern United States,Weather and Forecasting, 29(5), 1199-1207,2014.

Bennett, J. C., Robertson, D. E., Shrestha, D. L., Wang, Q., Enever, D., Hapuarachchi, P., and Tuteja, N. K.: A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9days, Journal of Hydrology, 519, 2832-2846, 2014.

Benninga, H.-J. F., Booij, M. J., Romanowicz, R. J., and Rientjes, T. H. M.: Performance of ensemble streamflow forecasts under varied hydrometeorological conditions, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2016-584, in review, 2016.

Bogner, K., Pappenberger, F., and Cloke, H.: Technical Note: The normal quantile transformation and its application in a flood forecasting system, Hydrology and Earth System Sciences, 16, 1085-1094, 2012.

Bourgin, F., Ramos, M.-H., Thirel, G., and Andreassian, V.: Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting, Journal of Hydrology, 519, 2775-2784, 2014.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Monthly weather review, 78, 1-3,1950.

Brown, J. D., and Seo, D.-J.: A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts, Journal of Hydrometeorology, 11, 642-665, 2010.

Brown, J. D., He, M., Regonda, S., Wu, L., Lee, H., and Seo, D.-J.: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification, Journal of Hydrology, 519, 2847-2868, 2014.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R.: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields, Journal of Hydrometeorology, 5, 243-262, 2004.

Cloke, H., and Pappenberger, F.: Ensemble flood forecasting: a review, Journal of Hydrology, 375, 613-626, 2009.

Dankers, R., Arnell, N. W., Clark, D. B., Falloon, P. D., Fekete, B. M., Gosling, S. N., Heinke, J., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., and Wisser, D.: First look at changes in flood hazard in the Inter-Sectoral Impact Model Intercomparison Project ensemble, Proceedings of the National Academy of Sciences, 111, 3257-3261, 10.1073/pnas.1302078110, 2014.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., and Fresch, M.: The science of NOAA's operational hydrologic ensemble forecast service, Bulletin of the American Meteorological Society, 95, 79-98, 2014.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models, Water resources research, 49, 4035-4053, 2013.

Demuth, N., and Rademacher, S.: Flood Forecasting in Germany—Challenges of a Federal Structure and Transboundary Cooperation, Flood Forecasting: A Global Perspective, 125, 2016.

Dogulu, N., López López, P., Solomatine, D., Weerts, A., and Shrestha, D.: Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments, Hydrology and Earth System Sciences, 19, 3181-3201, 2015.

Emerton, R. E., Stephens, E. M., Pappenberger, F., Pagano, T. C., Weerts, A. H., Wood, A. W., Salamon, P., Brown, J. D., Hjerdt, N., and Donnelly, C.: Continental and global scale flood forecasting systems, Wiley Interdisciplinary Reviews: Water, 2016.

Fan, F. M., Collischonn, W., Meller, A., and Botelho, L. C. M.: Ensemble streamflow forecasting experiments in a tropical basin: The São Francisco river case study, Journal of Hydrology, 519, 2906-2919, 2014.

Fares, A., Awal, R., Michaud, J., Chu, P.-S., Fares, S., Kodama, K., and Rosener, M.: Rainfall-runoff modeling in a flashy tropical watershed using the distributed HL-RDHM model, Journal of Hydrology, 519, 3436-3447, 2014.

Gitro, C. M., Evans, M. S., & Grumm, R. H.: Two Major Heavy Rain/Flood Events in the Mid-Atlantic: June 2006 and September 2011, Journal of Operational Meteorology, 2(13), 2014.

Golding, B., Roberts, N., Leoncini, G., Mylne, K., and Swinbank, R.: MOGREPS-UK convection-permitting ensemble products for surface water flood forecasting: Rationale and first results, Journal of Hydrometeorology, 17, 1383-1406, 2016.

Hamill, T. M., Whitaker, J. S., and Wei, X.: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts, Monthly Weather Review, 132, 1434-1447, 2004.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau Jr, T. J., Zhu, Y., and Lapenta, W.: NOAA's second-generation global medium-range ensemble reforecast dataset, Bulletin of the American Meteorological Society, 94, 1553-1565, 2013.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather and Forecasting, 15, 559-570, 2000.

Hopson, T. M., and Webster, P. J.: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07, Journal of Hydrometeorology, 11, 618-641, 2010.

Jolliffe, I. T., and Stephenson, D. B.: Forecast verification: a practitioner's guide in atmospheric science, John Wiley & Sons, 2012.

Journel, A. G., and Huijbregts, C. J.: Mining geostatistics, Academic press, 1978.
Kang, T. H., Kim, Y. O., and Hong, I. P.: Comparison of pre- and post- processors for ensemble streamflow prediction, Atmospheric Science Letters, 11, 153-159, 2010.
Koenker, R., and Bassett Jr, G.: Regression quantiles, Econometrica: journal of the Econometric Society, 33-50, 1978.
Koenker, R.: Quantile regression, 38, Cambridge university press, 2005.
Koren, V., Smith, M., Wang, D., and Zhang, Z.: 2.16 USE OF SOIL PROPERTY DATA IN THE DERIVATION OF CONCEPTUAL RAINFALL-RUNOFF MODEL PARAMETERS, 2000.
Koren, V., Reed, S., Smith, M., Zhang, Z., and Seo, D.-J.: Hydrology laboratory research modeling system (HL-RMS) of the US national weather service, Journal of Hydrology, 291, 297-318, 2004.
Krzysztofowicz, R.: Transformation and normalization of variates with specified distributions, Journal of Hydrology, 197, 286-292, 1997.
Kuzmin, V., Seo, D.-J., and Koren, V.: Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search, Journal of Hydrology, 353, 109-128, 2008.
Kuzmin, V.: Algorithms of automatic calibration of multi-parameter models used in operational systems of flash flood forecasting, Russian Meteorology and Hydrology, 34, 473-481, 2009.
López, P. L., Verkade, J., Weerts, A., and Solomatine, D.: Alternative configurations of quantile regression for estimating predictive uncertainty in water forecasts for the upper Severn River: a comparison, Hydrology and Earth System Sciences, 18, 3411-3428, 2014.
Madadgar, S., Moradkhani, H., and Garen, D.: Towards improved post- processing of hydrologic forecast ensembles, Hydrological Processes, 28, 104-122, 2014.
MARFC: http://www.weather.gov/marfc/Top20, accesed on April 1, 2017.
McCuen, R. H., and Snyder, W. M.: A proposed index for comparing hydrographs, Water Resources Research, 11, 1021-1024,1975.
Mendoza, P. A., McPhee, J., and Vargas, X.: Uncertainty in flood forecasting: A distributed modeling approach in a sparse data catchment, Water Resources Research, 48(9),2012.
Mendoza, P.A.,Wood, A., Clark, E., Nijssen, B., Clark,M., Ramos,MH.,and Voisin N.: Improving medium-range ensemble streamflow forecasts through statistical postprocessing. Presented at 2016 Fall Meeting, AGU,San Francisco, Calif., 11-15 Dec, 2016.
Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A.: Extending extended logistic regression: Extended versus separate versus ordered versus censored, Monthly Weather Review, 142, 3003-3014, 2014a.
Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S.: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance, Monthly Weather Review, 142, 448-456, 2014b.
Moore, B. J., Mahoney, K. M., Sukovich, E. M., Cifelli, R. , and Hamill T. M.: Climatology and environmental characteristics of extreme precipitation events in the southeastern United States, Monthly Weather Review, 143, 718–741,2015.
Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282-290,1970
NCAR:https://ral.ucar.edu/projects/system-for-hydromet-analysis-research-and-prediction-sharp, accessed on April 1, 2017.

Pagano, T., Elliott, J., Anderson, B., and Perkins, J.: Australian Bureau of Meteorology Flood Forecasting and Warning, Flood Forecasting: A Global Perspective, 1, 2016.

Pagano, T. C., Wood, A. W., Ramos, M.-H., Cloke, H. L., Pappenberger, F., Clark, M. P., Cranston, M., Kavetski, D., Mathevet, T., and Sorooshian, S.: Challenges of operational river forecasting, Journal of Hydrometeorology, 15, 1692-1707, 2014.

Politis, D. N., and Romano, J. P.: The stationary bootstrap, Journal of the American Statistical association, 89, 1303-1313, 1994.

Prat, O. P., and Nelson, B. R.: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012), Hydrology and Earth System Sciences, 19, 2037–2056,2015.

Rafieeinasab, A., Norouzi, A., Kim, S., Habibi, H., Nazari, B., Seo, D.-J., Lee, H., Cosgrove, B., and Cui, Z.: Toward high-resolution flash flood prediction in large urban areas–Analysis of sensitivity to spatiotemporal resolution of rainfall input and hydrologic modeling, Journal of Hydrology, 531, 370-388, 2015.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., and Participants, D.: Overall distributed model intercomparison project results, Journal of Hydrology, 298, 27-60, 2004.

Reed, S., Schaake, J., and Zhang, Z.: A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations, Journal of Hydrology, 337, 402-420, 2007.

Regonda, S. K., Seo, D. J., Lawrence, B., Brown, J. D., and Demargne, J.: Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts–A Hydrologic Model Output Statistics (HMOS) approach, Journal of hydrology, 497, 80-96, 2013.

Roulin, E., and Vannitsem, S.: Post- processing of medium- range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors, Hydrological Processes, 29, 1434-1449, 2015.

Saleh, F., Ramaswamy, V., Georgas, N., Blumberg, A. F., and Pullen, J.: A retrospective streamflow ensemble forecast for an extreme hydrologic event: a case study of Hurricane Irene and on the Hudson River basin, Hydrol. Earth Syst. Sci., 20, 2649-2667, doi:10.5194/hess-20-2649-2016, 2016.

Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M.: HEPEX: the hydrological ensemble prediction experiment, Bulletin of the American Meteorological Society, 88, 1541-1547, 2007.

Schellekens, J., Weerts, A., Moore, R., Pierce, C., and Hildon, S.: The use of MOGREPS ensemble rainfall forecasts in operational flood forecasting systems across England and Wales, Advances in Geosciences, 29, 77-84, 2011.

Schwanenberg, D., Fan, F. M., Naumann, S., Kuwajima, J. I., Montero, R. A., and Dos Reis, A. A.: Short-term reservoir optimization for flood mitigation under meteorological and hydrological forecast uncertainty, Water Resources Management, 29, 1635-1651, 2015.

Sharma, S., Siddique,R., Balderas,N., Fuentes, J.D., Reed, S., Ahnert, P., Shedd, R., Astifan, B., Cabrera, R., Laing, A., Klein, M., and Mejia, A.: Eastern U.S. Verification of Ensemble Precipitation Forecasts. Wea. Forecasting, 32, 117–139, 2017.

Siddique, R., Mejia, A., Brown, J., Reed, S., and Ahnert, P.: Verification of precipitation forecasts from two numerical weather prediction models in the Middle Atlantic Region of the USA: A precursory analysis to hydrologic forecasting, Journal of Hydrology, 529, 1390-1406, 2015.

Siddique, R., and Mejia, A.: Ensemble streamflow forecasting across the US middle Atlantic region with a distributed hydrological model forced by GEFS reforecasts, Journal of Hydrometeorology, 2017.
Sloughter, J. M. L., Raftery, A. E., Gneiting, T., and Fraley, C.: Probabilistic quantitative precipitation forecasting using Bayesian model averaging, Monthly Weather Review, 135, 3209-3220, 2007.
Smith, M. B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., Cui, Z., Mizukami, N., Anderson, E. A., and Cosgrove, B. A.: The distributed model intercomparison project–Phase 2: Motivation and design of the Oklahoma experiments, Journal of Hydrology, 418, 3-16, 2012a.
Smith, M. B., Koren, V., Zhang, Z., Zhang, Y., Reed, S. M., Cui, Z., Moreda, F., Cosgrove, B. A., Mizukami, N., and Anderson, E. A.: Results of the DMIP 2 Oklahoma experiments, Journal of Hydrology, 418, 17-48, 2012b.
Thiemig, V., Bisselink, B., Pappenberger, F., and Thielen, J.: A pan-African medium-range ensemble flood forecast system, Hydrology and Earth System Sciences, 19, 3365, 2015.
Thorstensen, A., Nguyen, P., Hsu, K., and Sorooshian, S.: Using Densely Distributed Soil Moisture Observations for Calibration of a Hydrologic Model, Journal of Hydrometeorology, 17, 571-590, 2016.
Verkade, J., Brown, J., Reggiani, P., and Weerts, A.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, Journal of Hydrology, 501, 73-91, 2013.
Wang, Q., Bennett, J. C., and Robertson, D. E.: Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting, Hydrology and Earth System Sciences, 20, 3561, 2016.
Ward, P. J., Jongman, B., Salamon, P., Simpson, A., Bates, P., De Groeve, T., Muis, S., De Perez, E. C., Rudari, R., and Trigg, M. A.: Usefulness and limitations of global flood risk models, Nature Climate Change, 5, 712-715, 2015.
Weerts, A., Winsemius, H., and Verkade, J.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrology and Earth System Sciences, 15, 255, 2011.
Wheater, H. S., and Gober, P.: Water security and the science agenda, Water Resources Research, 51, 5406-5424, 2015.
Wilks, D. S.: Extending logistic regression to provide full- probability- distribution MOS forecasts, Meteorological Applications, 16, 361-368, 2009.
Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic press, 2011.
Yang, X., Sharma, S., Siddique, R., Greybush, S. J., and Mejia, A.: Postprocessing of GEFS Precipitation Ensemble Reforecasts over the US Mid-Atlantic Region, Monthly Weather Review, 145, 1641-1658, 2017.
Zalachori, I., Ramos, M., Garçon, R., Mathevet, T., and Gailhard, J.: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies, Advances in Science & Research, 8, p. 135-p. 141, 2012.
Zappa, M., Rotach, M. W., Arpagaus, M., Dorninger, M., Hegg, C., Montani, A., Ranzi, R., Ament, F., Germann, U., and Grossi, G.: MAP D- PHASE: real- time demonstration of hydrological ensemble prediction systems, Atmospheric Science Letters, 9, 80-87, 2008.

Zappa, M., Jaun, S., Germann, U., Walser, A., and Fundel, F.: Superposition of three sources of uncertainties in operational flood forecasting chains, Atmospheric Research, 100, 246-262, 2011.

Zhao, L., Duan, Q., Schaake, J., Ye, A., and Xia, J.: A hydrologic post-processor for ensemble streamflow predictions, Advances in Geosciences, 29, 51-59, 2011.

Table 1. Main characteristics of the four study basins.

Location of outlet	Cincinnatus, New York	Chenango Forks, New York	Conklin, New York	Waverly, New York
NWS id	CINN6	CNON6	CKLN6	WVYN6
USGS id	01510000	01512500	01503000	01515000
Area [km^2]	381	3841	5781	12362
Latitude	$42^0$32’28”	$42^0$13’05”	$42^0$02’07”	$41^0$59’05”
Longitude	$75^0$53’59”	$75^0$50’54”	$75^0$48’11”	$76^0$30’04”
Minimum daily flow* [m^3/s]	0.31 (0.11)	4.05 (2.49)	6.80 (5.32)	13.08 (6.71)
Maximum daily flow* [m^3/s]	172.73 (273.54)	1248.77 (1401.68)	2041.64 (2174.734)	4417.42 (4417.42)
Mean daily flow* [m^3/s]	8.89 (9.17)	82.36 (81.66)	122.93 (121.99)	277.35 (215.01)
Climatological flow (Pr=0.95)** [m^3/s]	29.45	266.18	382.28	843.84

*The number in parenthesis is the historical (based on entire available record, as opposed to the period 2004-2012 used in this study) daily minimum, maximum, or mean recorded flow.
**Pr=0.95 indicates flows with exceedance probability of 0.05.

Table 2. Summary and description of the verification scenarios.

Scenario	Description
S1	Verification of the raw ensemble precipitation forecasts from the GEFSRv2
S2	Verification of the preprocessed ensemble precipitation forecasts from the GEFSRv2: GEFSRv2+HCLR
S3	Verification of the raw ensemble flood forecasts: GEFSRv2+HL-RDHM
S4	Verification of the preprocessed ensemble flood forecasts: GEFSRv2+HCLR+HL-RDHM
S5	Verification of the postprocessed ensemble flood forecasts: GEFSRv2+HL-RDHM+QR
S6	Verification of the preprocessed and postprocessed ensemble flood forecasts: GEFSRv2+HCLR+HL-RDHM+QR

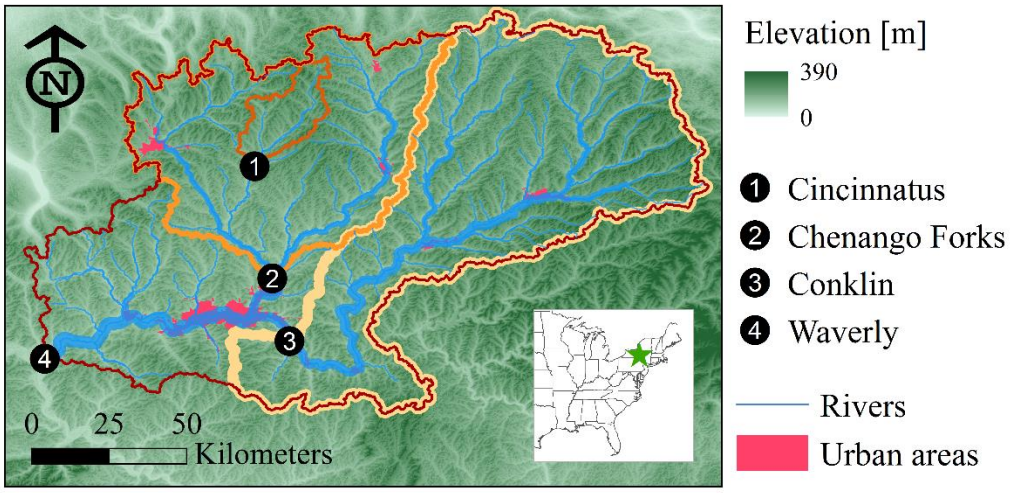


Figure 1. Map illustrating the location of the four selected river basins in the U.S. middle Atlantic region.

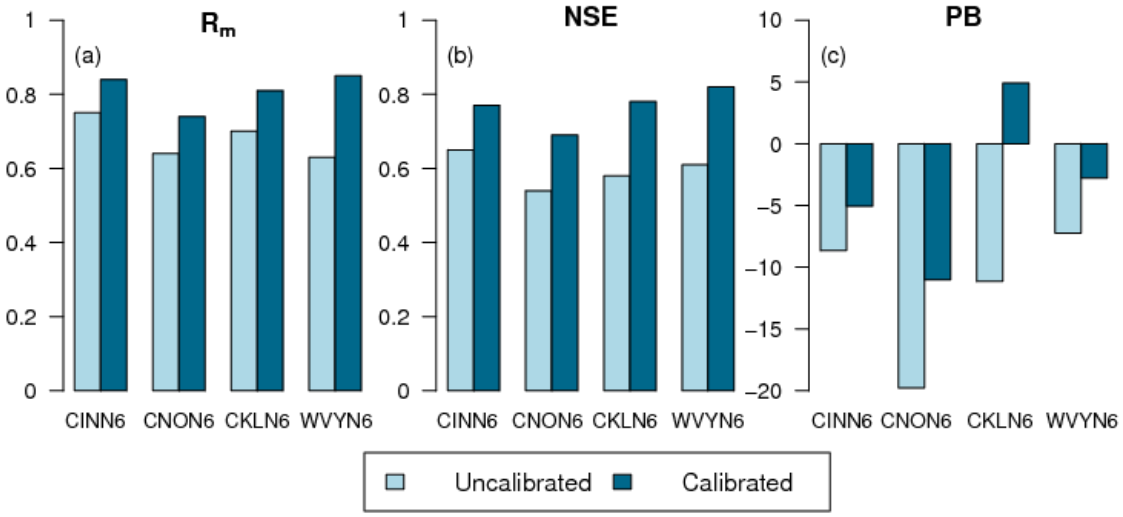


Figure 2. Performance statistics for the uncalibrated and calibrated simulation runs for the entire period of analysis (years 2004-2012): (a) R_m, (b) NSE, and (c) PB.

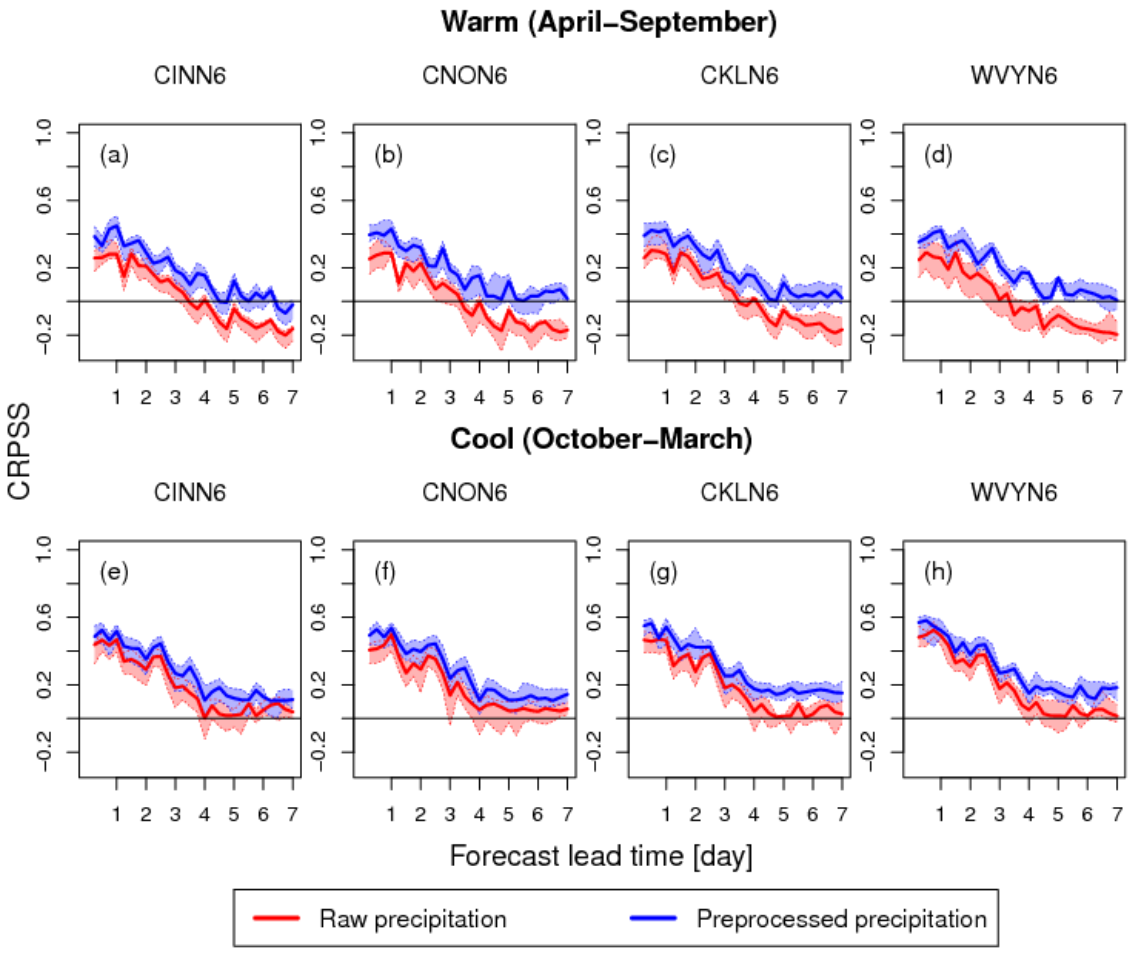


Figure 3. CRPSS (relative to sampled climatology) of the raw (red curves) and preprocessed (blue curves) ensemble precipitation forecasts from the GEFSRv2 vs the forecast lead time during the (a)-(d) warm (April-September) and (e)-(h) cool season (October-March) for the selected basins.

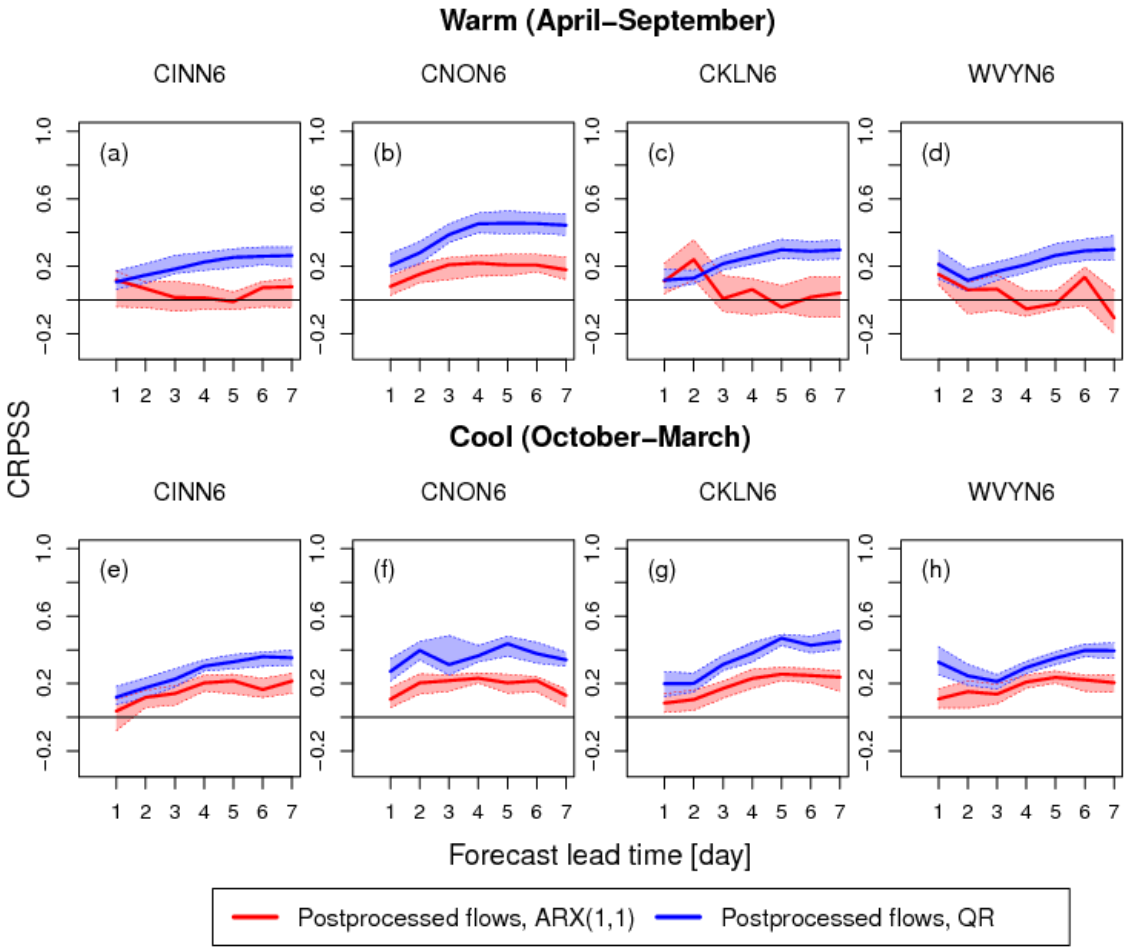


Figure 4. CRPSS (relative to the raw forecasts) of the ARX(1,1) (red curves) and QR (blue curves) postprocessed ensemble flood forecasts vs the forecast lead time during the (a)-(d) warm (April-September) and (e)-(h) cool season (October-March) for the selected basins.

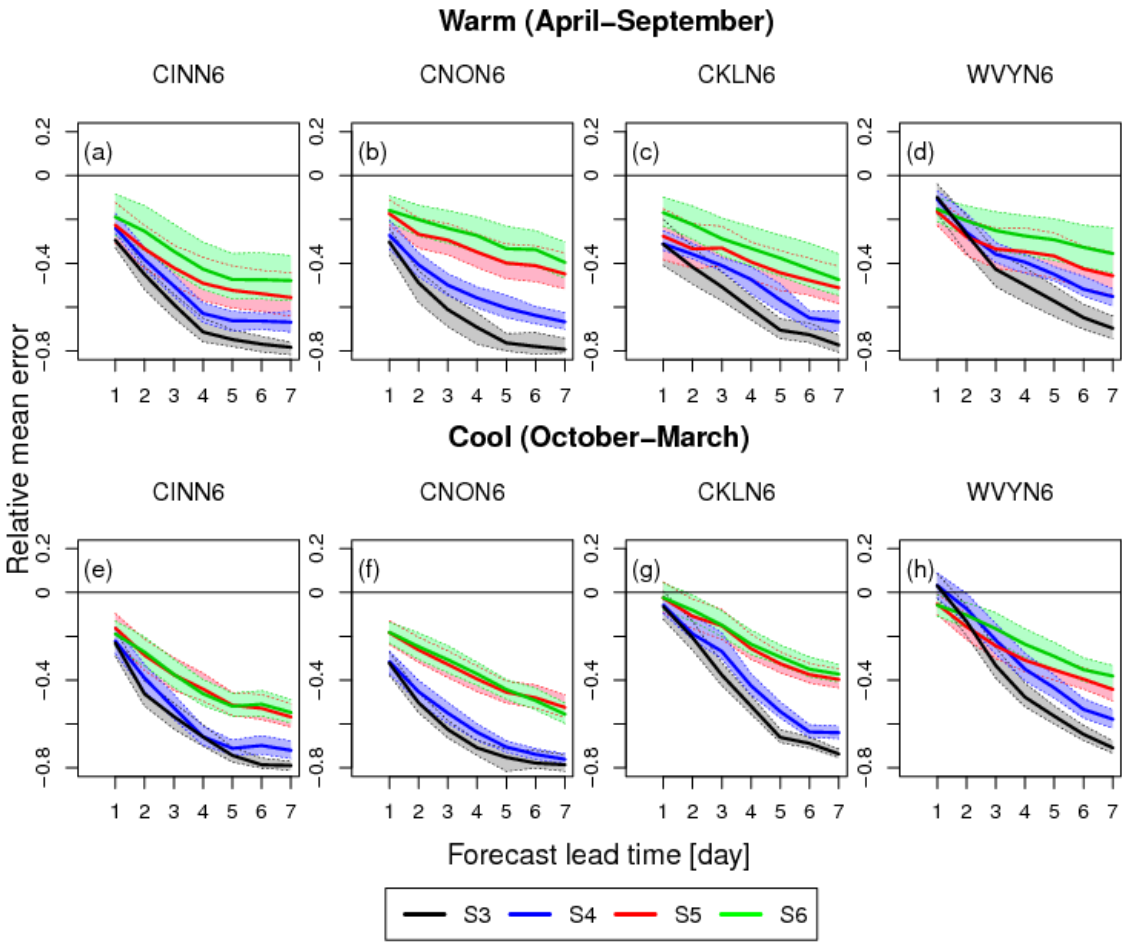


Figure 5. Relative mean error (RME) of the mean ensemble flood forecasts vs the forecast lead time during the (a)-(d) warm (April-September) and (e)-(h) cool season (October-March) for the selected basins. The curves represent the different forecasting scenarios S3-S6. Note that S3 consists of GEFSRv2+HL-RDHM, S4 of GEFSRv2+HCLR+HL-RDHM, S5 of GEFSRv2+HL-RDHM+QR, and S6 of GEFSRv2+HCLR+HL-RDHM+QR.

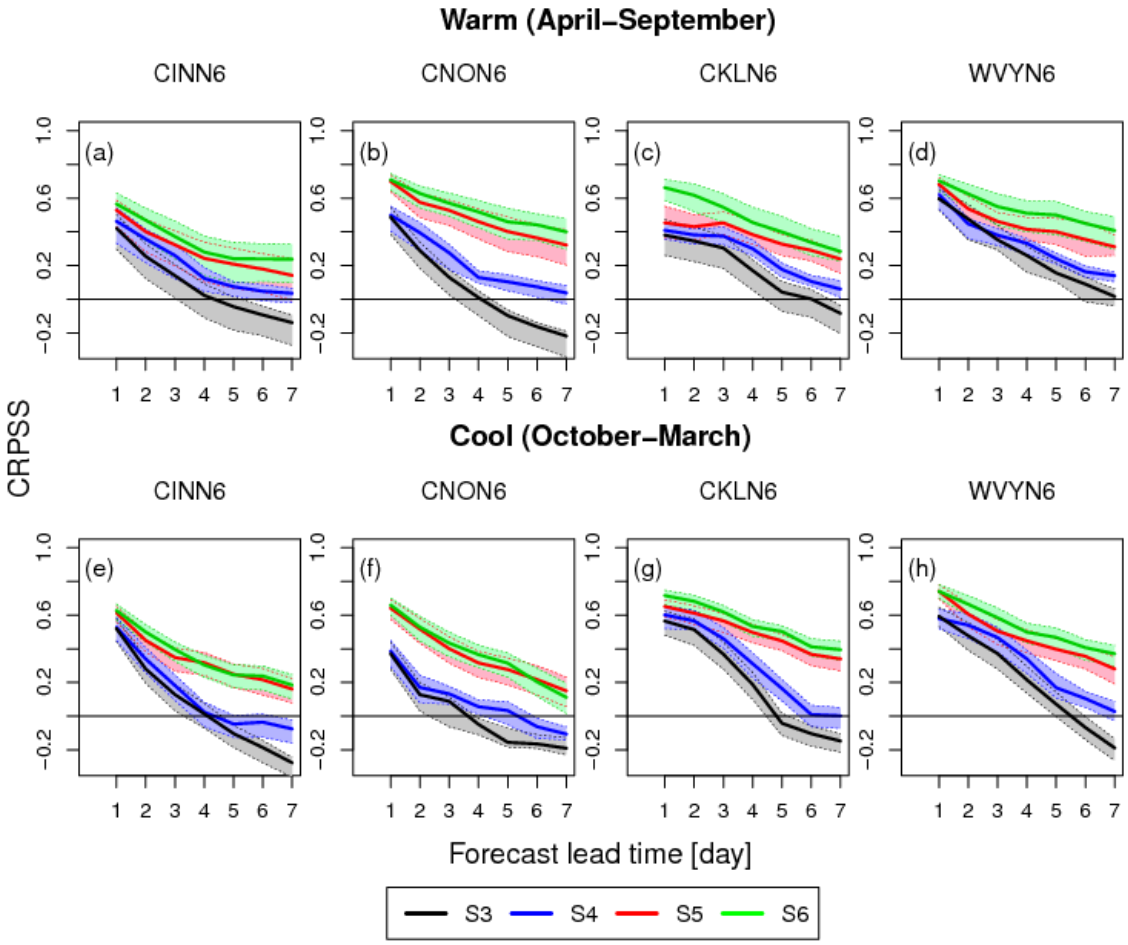


Figure 6. As in Fig. 5, but for the CRPSS (relative to sampled climatology) of the ensemble flood forecasts vs the forecast lead time.

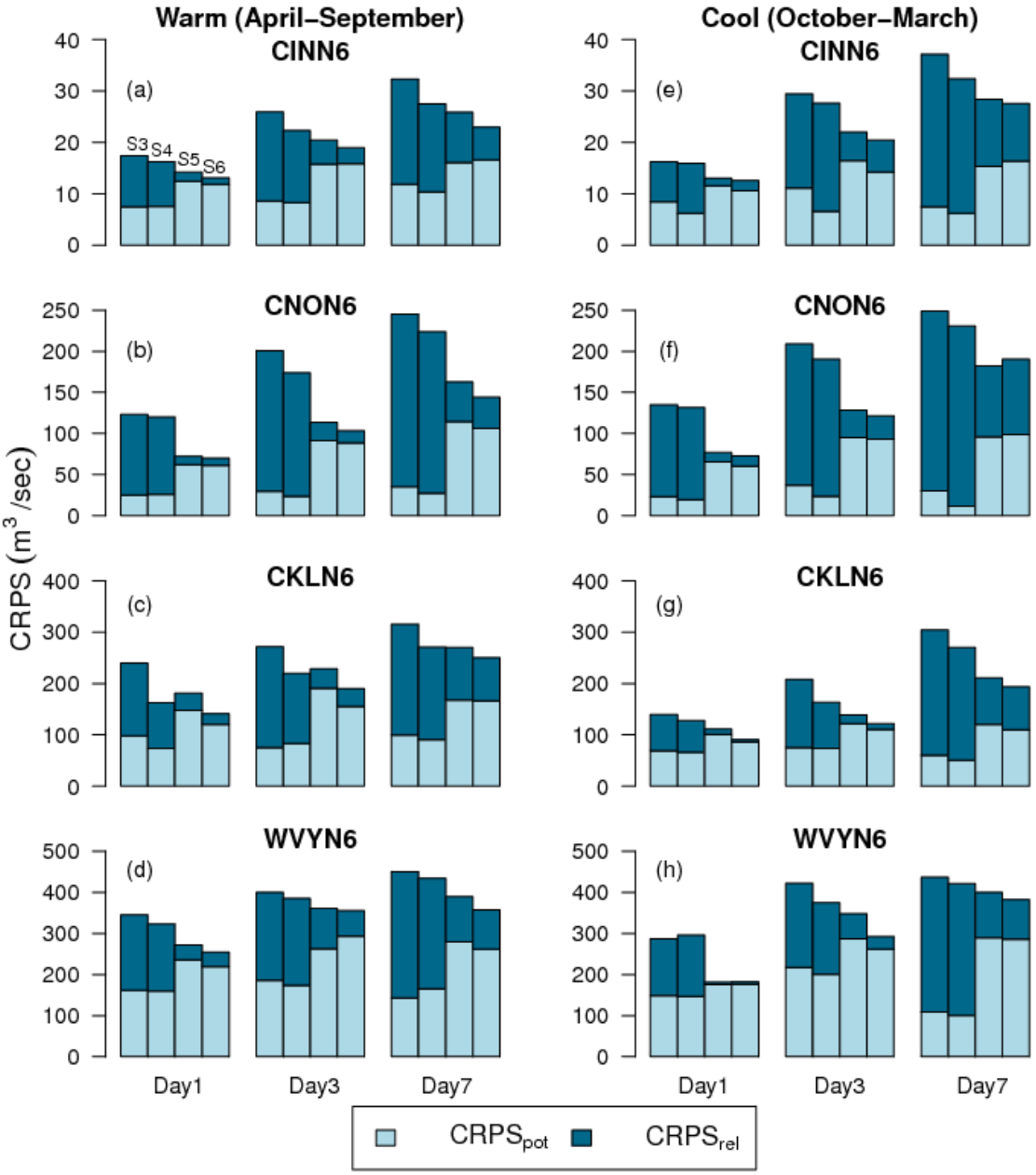


Figure 7. Decomposition of the CRPS into CRPS potential ($CRPS_{pot}$) and CRPS reliability ($CRPS_{rel}$) for forecasts lead times of 1, 3, and 7 days during the warm (a)-(d) (April-September) and cool season (e)-(h) (October-March) for the selected basins. The four columns associated with each forecast lead time represent the forecasting scenarios S3-S6 (from left to right). Note that S3 consists of GEFSRv2+HL-RDHM, S4 of GEFSRv2+HCLR+HL-RDHM, S5 of GEFSRv2+HL-RDHM+QR, and S6 of GEFSRv2+HCLR+HL-RDHM+QR.

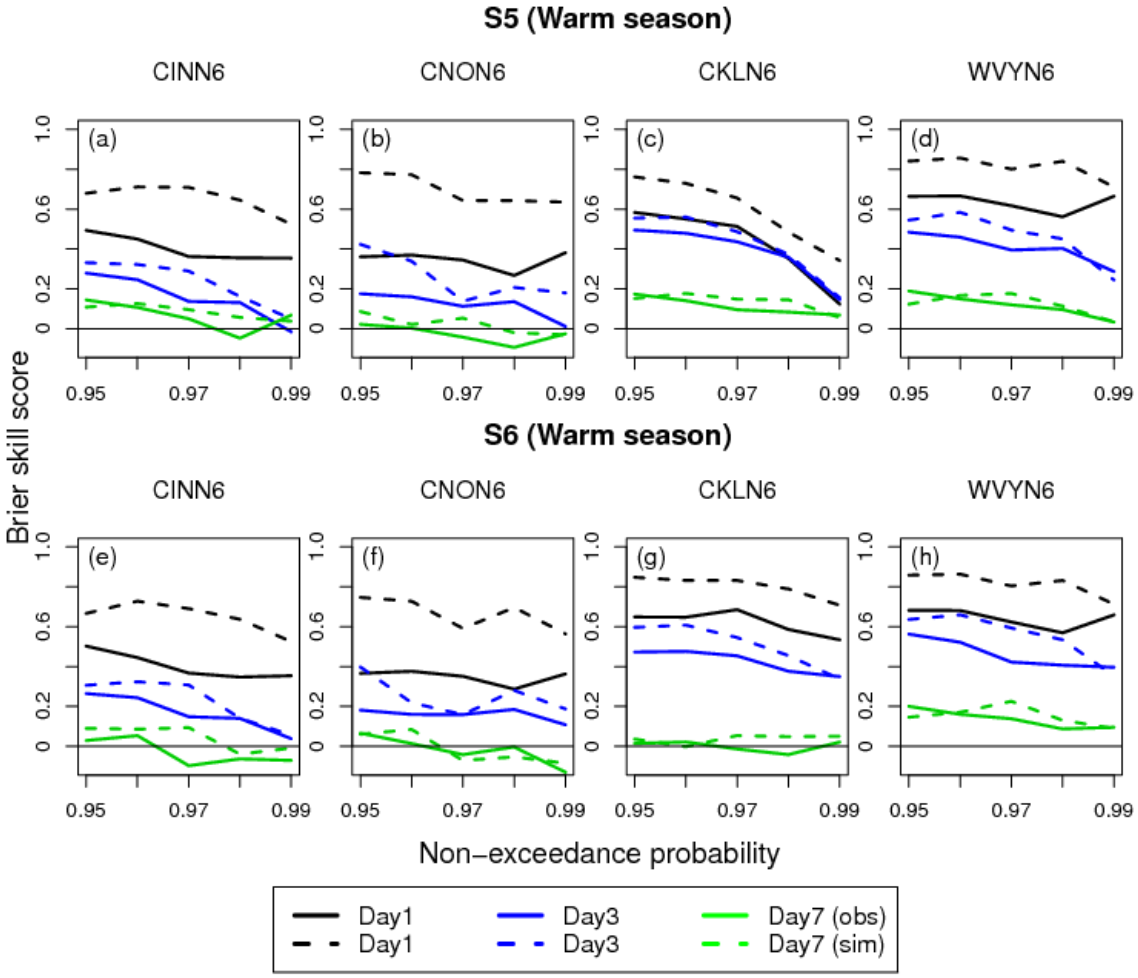


Figure 8. Brier skill score (BSS) of the mean ensemble flood forecasts for S5 (a-d) and S6 (e-h) vs the flood threshold for forecast lead times of 1, 3, and 7 days during the warm (April-September) season for the selected basins. The BSS is shown relative to both observed (solid lines) and simulated floods (dashed lines).

Chapter 7: Hydrological Model Diversity Enhances Streamflow Forecast Skill More than the Ensemble Size at Short- to Medium-Range Timescales

ABSTRACT

We investigate the ability of hydrological multimodel ensemble predictions to enhance the skill of streamflow forecasts at short- to medium-range timescales. To generate the multimodel ensembles, we implement a new statistical postprocessor, namely quantile regression-Bayesian model averaging (QR-BMA). QR-BMA uses QR to bias correct the ensemble streamflow forecasts from the individual models and BMA to optimally combine their probability density functions. Additionally, we use an information-theoretic measure, namely conditional mutual information, to quantify the skill enhancements from the multimodel forecasts. We generate ensemble streamflow forecasts at lead times from 1 to 7 days using three hydrological models: i) Antecedent Precipitation Index (API)-Continuous, ii) Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM), and iii) Weather Research and Forecasting Hydrological (WRF-Hydro) modeling system. As forcing to the hydrological models, we use weather ensemble forecasts from the National Centers for Environmental Prediction 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2). The forecasting experiments are performed for four nested basins of the North Branch Susquehanna River, USA. We find that after bias-correcting the streamflow forecasts from each model their skill performance becomes comparable. We find that the multimodel ensemble forecasts have higher skill than the best single-model forecasts. Furthermore, the skill enhancements obtained by the multimodel ensemble forecasts are found to be dominated by model diversity, rather than by increased ensemble size alone. This result, obtained using conditional mutual information, indicates that each hydrological model contributes additional information to enhance forecast skill. Overall, our results highlight benefits of hydrological multimodel forecasting for improving streamflow predictions.

1. Introduction

Multimodel forecasting is a well-established technique in atmospheric science (Bosart, 1975; Gyakum, 1986; Krishnamurti, 2003; Sanders, 1973; Weisheimer et al., 2009), which consists of using the outputs from several models to make and improve predictions about future events (Fritsch et al., 2000). The motivation for multimodel forecasting is that for a complex system, such as the atmosphere or a river basin, comprised by multiple processes interacting nonlinearly and with limited observability, predictions solely based on the outputs from a single model will be prone to errors and biases (Fritsch et al., 2000). Indeed, early experiments comparing blended forecasts from different weather models against single-model predictions demonstrated the ability of multimodel predictions to improve the skill and reduce the errors of weather forecasts (Bosart, 1975; Gyakum, 1986; Sanders, 1973; Thompson, 1977; Winkler et al., 1977). This was found to be the case for both forecasts issued by humans (Sanders, 1963, 1973) and from numerical models (Bosart, 1975; Fraedrich & Leslie, 1987; Fraedrich & Smith, 1989; Fritsch et al., 2000; Gyakum, 1986; Krishnamurti et al., 1999, 2000; Sanders, 1973).

Initial meteorological multimodel experiments accounted for model-related uncertainties but not for uncertainties in the initial states. To account for the latter, multimodel ensembles were introduced, where multiple ensemble members from individual models are generated for the

same lead time and geographic area by perturbing the models' initial states (Hamill & Colucci, 1997; Stensrud et al., 1999; Toth & Kalnay, 1993). An illustrative example of a recent, successful multimodel framework is the North American Multimodel Ensemble experiment for subseasonal to seasonal timescales (Bastola et al., 2013; Becker et al., 2014; Kirtman et al., 2013). Indeed, most of the established operational systems across the globe for short- to medium-range weather forecasting are multimodel, multiphysics ensemble systems (Buizza et al., 2005; Du et al., 2003; Hamill et al., 2013; Palmer et al., 2004). In contrast, hydrological multimodel ensemble prediction systems (HMEPS) have not been widely implemented and remain an underexplored area of research. To our knowledge, there is currently no operational HMEPS in the world, despite their success in weather (Hagedorn et al., 2012; Hamill et al., 2013) and climate forecasting (Bastola et al., 2013; Becker et al., 2014; Kirtman et al., 2013).

HMEPS can be classified into the following three general categories, depending on whether multiple weather and/or hydrological models are used: i) a single hydrological model forced by outputs from multiple numerical weather prediction (NWP) models (Thirel et al., 2008, 2010), ii) multiple hydrological models forced by outputs from a single NWP model (Randrianasolo et al., 2010), and iii) multiple hydrological models forced by outputs from multiple NWP models (Velázquez et al., 2011). As is the case in meteorology, hydrological multimodel outputs can be deterministic or probabilistic, depending on how many and the manner in which ensembles are generated from each model (Davolio et al., 2008). It is important to note that, although hydrological multimodel approaches have been investigated before (Ajami et al., 2007; Duan et al., 2007; Vrugt & Robinson, 2007), the vast majority of those studies have been performed in simulation mode (i.e., by forcing the hydrological models with observed weather variables), as opposed to forecasting mode. Simulation studies may provide useful information for near-real time hydrological forecasting conditions. However, at medium-range timescales ($\geq$ 3 days), where weather uncertainties tend to be as important or more dominant than hydrological uncertainties, hydrological simulations provide considerably less information about forecast behavior (Sharma et al., 2018; Siddique & Mejia, 2017).

One of the earliest attempt at hydrological multimodel prediction is that of Shamseldin and O'Connor (1999). They combined streamflow simulations from different rainfall-runoff models by assigning different weights to the models based on their performance during historical runs. Since then, several simulation studies have been performed to address the potential of hydrological multimodel approaches to improve understanding and prediction of hydrological variables (Ajami et al., 2007; Bohn et al., 2010; Duan et al., 2007; Georgakakos et al., 2004; Regonda et al., 2006; Vrugt & Robinson, 2007). In hydrological forecasting, recent implementations of the multimodel approach have been focused on seasonal or longer timescales (Nohara et al., 2006; Yuan & Wood, 2013), while very few studies are available at short- to medium-range timescales (Hopson & Webster, 2010; Velázquez et al., 2011). Furthermore, a shortcoming of the latter studies has been the use of similar hydrological models to generate the multimodel forecasts. For example, Hopson and Webster (2010) as well as Velázquez et al. (2011) used similar spatially lumped or semi-distributed hydrological models for their respective multimodel experiments.

To maximize the benefits from a multimodel approach, it is critical to use dissimilar models (Thompson, 1977), a property that is referred to as model diversity (DelSole et al., 2014). In hydrological science, different model types are available that could be used to fulfill model diversity, e.g., spatially lumped, spatially distributed, process-based, or land-surface models (Reed et al., 2004; Smith et al., 2012). These different types of models tend to differ markedly in

their spatial discretization, physical parameterizations, and numerical schemes (Kollet et al., 2017), potentially making them good candidates for multimodel forecasting. Another important concern with the multimodel approach is that of distinguishing whether any gains in skill from the multimodel are due to model diversity itself or are related to increases in the ensemble size. Recently, an information-theoretic measure, namely conditional mutual information (*CMI*), was proposed to address this issue in climate forecasts (DelSole et al., 2014). *CMI* is implemented here with hydrological multimodel forecasts for the first time.

Any multimodel forecast requires some type of statistical technique (with simple averaging being the simplest approach (DelSole, 2007; DelSole et al., 2013)) or postprocessor (Duan et al., 2007; Fraley et al., 2010; Raftery et al., 1997) to optimally combine the ensemble forecasts from the individual models. Multimodel postprocessing is typically employed to accomplish several objectives: i) reduce systematic biases in the outputs from each model, ii) assign each model a weight that measures its contribution to the final multimodel forecast, and iii) quantify the overall forecast uncertainty. Although a number of multimodel postprocessors have been developed and implemented for dealing with hydrological simulations (Duan et al., 2007; Hsu et al., 2009; Madadgar & Moradkhani, 2014; Najafi et al., 2011; Shamseldin et al., 1997; Steinschneider et al., 2015; Vrugt & Robinson, 2007; Xiong et al., 2001), few have been applied in a forecasting context (Hopson & Webster, 2010). In this study, we implement a new quantile regression-Bayesian model averaging (QR-BMA) postprocessor. The postprocessor uses QR to bias correct the streamflow forecasts from the individual models (Sharma et al., 2018) and BMA to optimally combine their probability density functions (pdfs) (Duan et al., 2007; Vrugt & Robinson, 2007). QR-BMA takes advantage of the proven effectiveness and simplicity of QR to remove systematic biases (Sharma et al., 2018) and of BMA to produce optimal weights (Duan et al., 2007; Liang et al., 2013).

Our primary goal with this study is to understand the ability of hydrological multimodel ensemble predictions to improve the skill of streamflow forecasts at short- to medium-range timescales. With this goal, we seek to answer the following two main questions: Are multimodel ensemble streamflow forecasts more skillful than single-model forecasts? Are any skill improvements from the multimodel ensemble streamflow forecasts dominated by model diversity or increasing ensemble size? Answering the latter is relevant to operational forecasting because generating many ensembles in real time is often not feasible or realistic, and may not be as effective if skill enhancements are dominated by model diversity. The paper is structured as follows. Section 2 describes our methodology. Section 3 describes the experimental setup. The main results and their implications are presented in Section 4. Lastly, Section 5 summarizes our conclusions.

2. Methodology

2.1. Statistical Multimodel Postprocessor

The proposed postprocessor uses QR to bias correct the ensemble forecasts from individual models and BMA to combine the bias-corrected forecasts. We begin by briefly revisiting the BMA technique. BMA generates an overall forecast pdf by taking a weighted average of the conditional pdfs associated with the individual model forecasts. Letting Δ be the forecasted variable, D the training data, and $M = [M_1, M_2, ..., M_k]$ the independent predictions from a total of *K* hydrological models, the pdf of the BMA probabilistic prediction of Δ can be expressed by the law of total probability as

$$P\left(\Delta \mid (M_1, M_2, ..., M_K)\right) = \sum_{k=1}^{K} P\left(\Delta \mid M_k\right) P\left(M_k \mid D\right), \quad (1)$$

where $P\left(\Delta \mid M_k\right)$ is the posterior distribution of Δ given the model prediction M_k, and $P\left(M_k \mid D\right)$ is the posterior probability of model M_k being the best one given the training data *D*. $P\left(M_k \mid D\right)$ reflects the performance of model M_k in predicting the forecast variable during the training period.

The posterior model probabilities are nonnegative and add up to one (Raftery et al., 2005), such that

$$\sum_{k=1}^{K} P\left(M_k \mid D\right) = 1. \quad (2)$$

Thus, $P\left(M_k \mid D\right)$ can be viewed as the model weight, w_k, reflecting an individual model's relative contribution to predictive skill over the training period. The BMA pdf is therefore a weighted average of the conditional pdfs associated with each of the individual model forecasts, weighted by their posterior model probabilities. Since model predictions are time variant, letting *t* be the forecast lead time, equation (1) can be written as

$$P\left(\Delta^t \mid (M_1^t, M_2^t, ..., M_K^t)\right) = \sum_{k=1}^{K} w_k^t P\left(\Delta^t \mid M_k^t\right). \quad (3)$$

The efficient application of BMA requires bias-correcting the ensemble forecasts from the individual models and optimizing their weights w_k^t (Raftery et al., 2005). We used QR to bias-correct the forecasts. QR has several advantages as compared to the linear regression bias correction used in the original BMA approach (Raftery et al., 2005). It does not make any prior assumptions regarding the shape of the distribution and, since QR results in conditional quantiles rather than conditional means, QR is less sensitive to the tail behavior of the streamflow data and, consequently, more robust to outliers.

To implement QR, the bias-corrected ensemble forecasts from each model *k* and forecast lead time *t*, $f_{k,\tau}^t$, are determined using

$$f_{k,\tau}^t = \overline{f}_k^t + \hat{\xi}_{k,\tau}, \quad (4)$$

where $\overline{f}_k^t$ is the ensemble mean forecast of model k at time t, and $\hat{\xi}_{k,\tau}$ is the error estimate at the quantile interval τ defined as

$$\hat{\xi}_{k,\tau} = a_{k,\tau}^t + b_{k,\tau}^t \overline{f}_k^t. \quad (5)$$

In equation (5), $a_{k,\tau}^t$ and $b_{k,\tau}^t$ are the regression parameters for model *k* and quantile interval τ at time *t*. The parameters associated with each model are determined separately by minimizing the sum of the residuals from a training dataset as follows

$$\arg\min_{f \in \mathrm{i}} \sum_{j=1}^{J} \Gamma_\tau^t [\xi_{\tau,j}^t - \hat{\xi}_{\tau,j}(j, \overline{f}_j)]. \quad (6)$$

$\xi_{\tau,j}^t$ and $\overline{f}_j$ are the j^{th} paired samples from a total of J samples; $\xi_{\tau,j}^t$ is computed as the observed flow minus the forecasted one at time t; Γ_τ^t is the QR function for the τ^{th} quantile at time t defined as

$$\Gamma_{\tau}^{t}(\Psi_{j}^{t}) = \begin{cases} (\tau-1)\Psi_{j}^{t} & \text{if } \Psi_{j}^{t} \le 0 \\ \tau\Psi_{j}^{t} & \text{if } \Psi_{j}^{t} > 0 \end{cases}, \quad (7)$$

and Ψ_{j}^{t} is the residual term computed as the difference between $\xi_{\tau,j}^{t}$ and $\hat{\xi}_{\tau,j}(j, \overline{f_j})$ for any quantile $\tau \in [0,1]$. The resulting minimization problem in equation (6) is solved using linear programming via the interior point method (Koenker, 2005). By varying the values of τ, QR allows describing the entire conditional distribution of the error estimate in equation (5).

After bias-correcting the single-model forecasts using equations (4)-(7), the posterior distribution of each model is assumed Gaussian. Thus, before implementing equation (3), both the observations and bias-corrected forecasts are transformed into standard normal deviates using the normal quantile transformation (NQT) (Krzysztofowicz, 1997). The NQT matches the empirical cumulative distribution function (cdf) of the marginal distribution to the standard normal distribution such that

$$f_{k,NQT}^{t} = G^{-1}\left(cdf(f_{k}^{t})\right), \quad (8)$$

where $cdf(.)$ is the cdf of the bias-corrected forecasts from model k at time t, f_{k}^{t}; G is the standard normal distribution and G^{-1} its inverse; and $f_{k,NQT}^{t}$ are the transformed, bias-corrected forecasts from model k at time t. When applying the NQT, extrapolation is used to model the tails of the forecast distribution for those cases where a sampled data point in normal space falls outside the range of the training data maxima or minima. For the upper tail, a hyperbolic distribution (Journel & Huijbregts, 1978) is used while linear extrapolation is used for the lower tail.

Lastly, to determine the BMA probabilistic prediction in equation (3), the weight w_{k}^{t} and variance $\sigma_{k}^{2,t}$ of model k at the forecast lead time t are estimated using the log likelihood function. Note that $\sigma_{k}^{2,t}$ is the variance associated with the Gaussian posterior distribution of model k. Setting the parameter vector $\theta = \left\{w_{k}^{t},\ \sigma_{k}^{2,t},\ k = 1,\ 2, ...,\ K\right\}$, the log likelihood function of θ at the forecast lead time t is approximated as

$$l(\theta) = \log\left(\sum_{k=1}^{K} w_{k}^{t}\, g(\Delta_{NQT}^{t} \mid f_{k,NQT}^{t})\right), \quad (9)$$

where $g(.)$ denotes a Gaussian pdf, and Δ_{NQT}^{t} is the forecasted variable in Gaussian space. Because of the high dimensionality of this problem, the log likelihood function typically cannot be maximized analytically. Thus, the maximum likelihood estimates of θ are determined using the expectation maximization (EM) optimization algorithm (Bilmes, 1998). The steps required to implement the EM algorithm are provided in Appendix A.

Our proposed QR-BMA approach consists of implementing equations (3)-(9). To apply QR-BMA, we used a leave-one-out approach where part of the forecast dataset was used to train QR-BMA and the rest to verify the multimodel ensemble forecasts. We applied QR-BMA at each forecast lead time t of interest for selected forecast locations. As part of our forecast experiments, we generated both single-model and multimodel ensemble forecasts. The single-model forecasts were postprocessed using QR following the same leave-one-out approach used with QR-BMA.

2.2. Measures of Forecast Skill

2.2.1. Conditional Mutual Information

CMI is used as a measure of skill improvement following the approach by DelSole et al. (2014). The approach allows to distinguish whether multimodel skill improvements are dominated by model diversity (i.e., additional information provided by the different models) or increased ensemble size. To present the *CMI* measure, we first introduce three related information-theoretic measures: entropy, conditional entropy, and mutual information.

In the case of a continuous random variable (e.g., the streamflow forecasts F with pdf $P(f)$, where uppercase is used to denote the random variable and lowercase its realizations), the amount of average information required to describe *F* is given by the entropy $\mathrm{H}(F)$ defined as

$$\mathrm{H}(F) = -\int P(f) \ln P(f)\, df. \tag{10}$$

Entropy measures the uncertainty of *F* (Cover & Thomas, 1991). The entropy of a random variable conditional upon the knowledge of another can be defined by the conditional entropy. The conditional entropy between the streamflow observations O and forecasts F can be calculated using the chain rule

$$\mathrm{H}(O \mid F) = \mathrm{H}(O, F) - \mathrm{H}(F). \tag{11}$$

With equations (10)-(11), the mutual information (*MI*) between the streamflow observations and the forecasts, $MI(O;F)$, is given by (Cover & Thomas, 1991)

$$\begin{aligned} MI(O;F) &= \mathrm{H}(O) + \mathrm{H}(F) - \mathrm{H}(O, F) \\ &= \iint P(o, f) \log\left[\frac{P(o, f)}{P(o)P(f)}\right] do\, df, \end{aligned} \tag{12}$$

where $P(o, f)$ is the joint pdf of *O* and *F*, with marginal pdfs $P(o)$ and $P(f)$, respectively. *MI* is an elegant and powerful measure to quantify the amount of information that one random variable contains about another random variable. It is nonnegative and equal to zero if and only if *O* and *F* are independent from each other. *MI* has several important benefits. It is a domain independent measure such that the information provided is relatively insensitive to the size of datasets and outliers, unaffected by systematic errors, and invariant to any nonlinear transformations of the variables (Cover & Thomas, 1991; Kinney & Atwal, 2014).

In the case of multimodel combinations, where F_1 represents the single-model ensemble mean and F_2 represents the multimodel mean of the remaining models, the *CMI* between *O* and F_2, conditioning out F_1, is given by

$$CMI(O; F_2 \mid F_1) = MI(O; (F_1, F_2)) - MI(O; F_1), \tag{13}$$

where the mutual information $MI(O;(F_1, F_2))$ measures the degree of dependence between the observation and the joint variability of the forecasts F_1 and F_2. According to equation (13), *CMI* quantifies the additional decrease in uncertainty due to adding a single model forecast to the multimodel forecast mean of the other models. When the distributions are Gaussian, the *CMI* reduces to a simple function of partial correlation as follows (Sedghi & Jonckheere, 2014)

$$CMI(O; F_2 \mid F_1) = -\frac{1}{2}\log\left(1 - \rho^2_{O2|1}\right), \tag{14}$$

where $\rho_{O2|1}$ denotes the partial correlation between O and F_2 conditioned on F_1. The partial correlation is related to the pairwise correlations by (Abdi, 2007)

$$\rho_{O2|1} = \frac{\rho_{O2} - \rho_{O1}\rho_{12}}{\sqrt{\left(1-\rho_{O1}^2\right)\left(1-\rho_{12}^2\right)}}, \tag{15}$$

where ρ_{O1} and ρ_{O2} are the correlation skills of F_1 and F_2, respectively; and ρ_{12} is the correlation between F_1 and F_2. Hereafter the subscript 1 denotes single-model forecasts and the subscript 2 denotes either single-model forecasts or multimodel forecasts, depending on whether one is assessing the skill of single-model or multimodel forecasts.

To further understand any skill enhancements provided by a multimodel forecast, the streamflow forecasts and observations can be partitioned into a conditional mean, called the signal variable α, and a deviation about the conditional mean, called the noise variable β. As shown by DelSole et al. (2014), in the case that all the ensemble members are drawn from the same model and the forecasts are computed with means of ensemble size E_1 and E_2, the partial correlation in equation (15) becomes

$$\rho_{O2|1}^{noise} = \frac{\rho_{\alpha O}}{E_1} \frac{\sqrt{SNR}}{\sqrt{SNR\left(\frac{1}{E_1}+\frac{1}{E_2}\right)+\frac{1}{E_1E_2}}\ \sqrt{SNR\left(1-\rho_{\alpha O}^2\right)+\frac{1}{E_1}}}, \tag{16}$$

where the signal-to-noise ratio *SNR* is defined as the ratio of signal variance to noise variance, and $\rho_{\alpha O}$ is the correlation between the signal variable and streamflow observation. The partial correlation in equation (16) is nonzero when a predictable signal exists (i.e., $SNR \neq 0$), forecast skill exists ($\rho_{\alpha O} \neq 0$), and the ensemble sizes are finite. To the extent that forecast skill exceeds predictability skill,

$$|\rho_{\alpha O}| \leq \sqrt{\frac{SNR}{SNR+1}}. \tag{17}$$

Equation (17) implies that an upper bound on $\rho_{\alpha O}$ results in an upper bound on the partial correlation in equation (16). Thus, an upper bound on the skill improvement due to adding new ensemble members from the same model can be estimated by combining equations (16)-(17) and taking the limit $SNR \to \infty$,

$$\rho_{O2|1}^{noise} \leq \sqrt{\frac{E_2}{\left(E_1+E_2\right)\left(E_1+1\right)}}. \tag{18}$$

Thus, any skill enhancement measured by equation (15) that exceeds the upper bound of equation (18) is dominated by the addition of new predictable signals (DelSole et al., 2014).

We computed *CMI* using equations (14)-(15), together with the streamflow ensemble forecasts and observations. We used equation (18) to obtain an upper bound for the skill improvement due to increased ensemble size. Any improvements beyond this upper bound, we attributed to the addition of new signals or model diversity. When using equations (14)-(15) and (18), the subscript 1 refers to the single model forecasts F_1 that one is trying to improve and the subscript 2 the multimodel forecasts F_2 or, in the case of a single-model experiment, the addition of new members from the same model. *CMI* was computed for each individual model and multimodel combination at every lead time of interest for selected forecast locations. Before

computing *CMI*, both the streamflow observations and forecasts were transformed into Gaussian space using NQT.

Additionally, we estimated *CMI* in streamflow space using the approach discussed by Meyer (2008). The approach relies on the Miller-Madow asymptotic bias-corrected empirical estimator for entropy estimation (Meyer, 2008; Miller, 1955) and an equal frequency binning algorithm for data discretization (Meyer, 2008). This approach does not require transforming streamflow into Gaussian space but has the drawback that an exact upper bound, akin to equation (18), is not available. The *CMI* in streamflow space was computed using the same experimental conditions described before for *CMI* in Gaussian space.

2.2.2. Continuous Ranked Probability Skill Score

Besides using *CMI* to measure skill improvements, we used the mean Continuous Ranked Probability Skill Score (*CRPSS*) (Hersbach, 2000) since this is a commonly used verification metric to assess the quality of ensemble forecasts (Brown et al., 2014). The *CRPSS* is derived from the Continuous Ranked Probability Skill Score (*CRPS*). The *CRPS* evaluates the overall accuracy of a probabilistic forecast by estimating the quadratic distance between the forecasts' cdf and the corresponding observations. The *CRPS* is defined as

$$CRPS = \int_{-\infty}^{\infty} \left[cdf(f) - \Pi(f - o) \right]^2 df, \tag{19}$$

where

$$\Pi(f) = \begin{cases} 0 & \text{for } f < 0 \\ 1 & \text{otherwise} \end{cases}. \tag{20}$$

$\Pi(.)$ is the Heaviside step function.

To evaluate the skill of the forecasting system relative to a reference system, the associated skill score or *CRPSS* is computed as

$$CRPSS = 1 - \frac{\overline{CRPS}_m}{\overline{CRPS}_r}, \tag{21}$$

where the *CRPS* is averaged across *n* pairs of forecasts and observations to calculate the mean *CRPS* of the main forecast system, $\overline{CRPS}_m$, and reference forecast system, $\overline{CRPS}_r$. The *CRPSS* ranges from $[-\infty, 1]$. Positive *CRPSS* values indicate the main forecasting system has higher skill than the reference forecasting system, with 1 indicating perfect skill. In this study, we used sampled climatology as the reference forecasting system. Similar to our implementation of *CMI*, the *CRPSS* was computed for both single-model and multimodel ensemble streamflow forecasts at each lead time of interest for selected forecast locations. Our proposed multimodel forecasting approach is summarized in Figure 1.

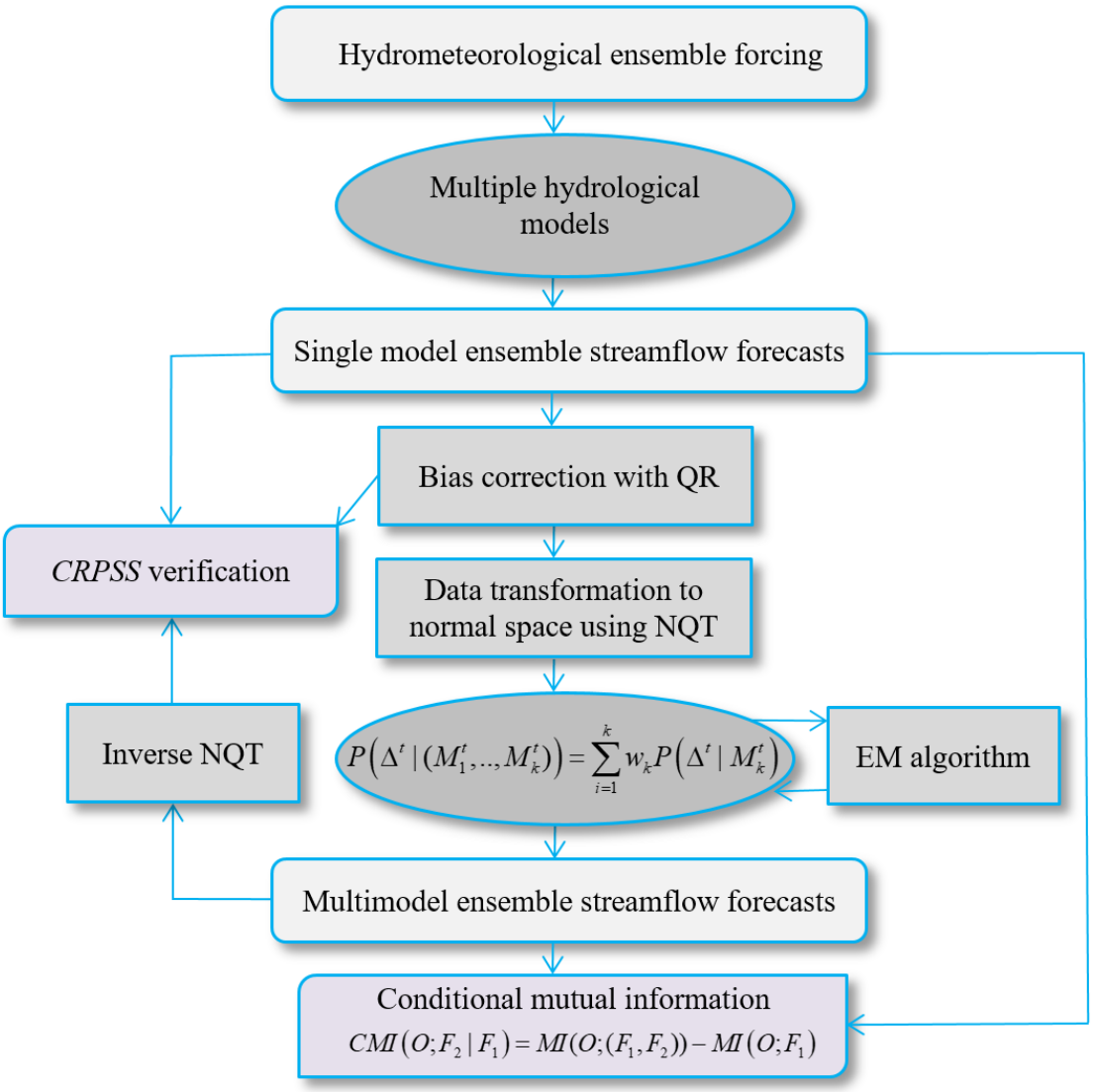


Figure 1. Diagrammatic representation of the proposed multimodel forecasting approach. The approach starts with the hydrometeorological ensemble forcing. The forcing is used to drive different hydrological models to generate single model ensemble streamflow forecasts. The single model forecasts are subsequently bias-corrected, transformed to Gaussian space, and combined using BMA to generate multimodel ensemble streamflow forecasts. Lastly, both the single model and multimodel forecasts are verified using the *CRPSS* and *CMI*.

3. Experimental Setup

3.1. Study Area

The North Branch Susquehanna River (NBSR) basin in the United States (US) Middle Atlantic Region (MAR) was selected as the study area (Figure 2) (Nelson, 1966). Severe weather and flooding hazards are an important concern in the NBSR, e.g., the City of Binghamton, New York, has been affected by multiple damaging flood events over recent years (Gitro et al., 2014;

Location of outlet	Cincinnatus, New York	Chenango Forks, New York	Conklin, New York	Waverly, New York
NWS id	CINN6	CNON6	CKLN6	WVYN6
USGS id	01510000	01512500	01503000	01515000
Area [km^2]	381	3841	5781	12362
Outlet latitude [North]	$42^0$32’28”	$42^0$13’05”	$42^0$02’07”	$41^0$59’05”
Outlet longitude [West]	$75^0$53’59”	$75^0$50’54”	$75^0$48’11”	$76^0$30’04”
Minimum daily flow[a]	0.31	4.05	6.80	13.08
[$m^3 s^{-1}$]	(0.11)	(2.49)	(5.32)	(6.71)
Maximum daily flow[a]	172.73	1248.77	2041.64	4417.42
[$m^3 s^{-1}$]	(273.54)	(1401.68)	(2174.734)	(4417.42)
Mean daily flow[a]	8.89	82.36	122.93	277.35
[$m^3 s^{-1}$]	(9.17)	(81.66)	(121.99)	(215.01)

[a]The number in parenthesis is the historical (based on the entire available record, as opposed to the period 2004-2009 used in this study) daily minimum, maximum, or mean recorded flow.

Jessup & DeGaetano, 2008). In the NBSR, four different US Geological Survey (USGS) daily gauge stations were selected as the forecast locations (Figure 2). The selected locations are the Ostelic

River at Cincinnatus (USGS gauge 01510000), Chenango River at Chenango Forks (USGS gauge 01512500), Susquehanna River at Conklin (USGS gauge 01503000), and Susquehanna River at Waverly (USGS gauge 01515000). These forecast locations represent a system of nested subbasins with drainage areas ranging from ~381 to 12,362 km^2. A summary of the main characteristics of the selected gauge locations is provided in Table 1.

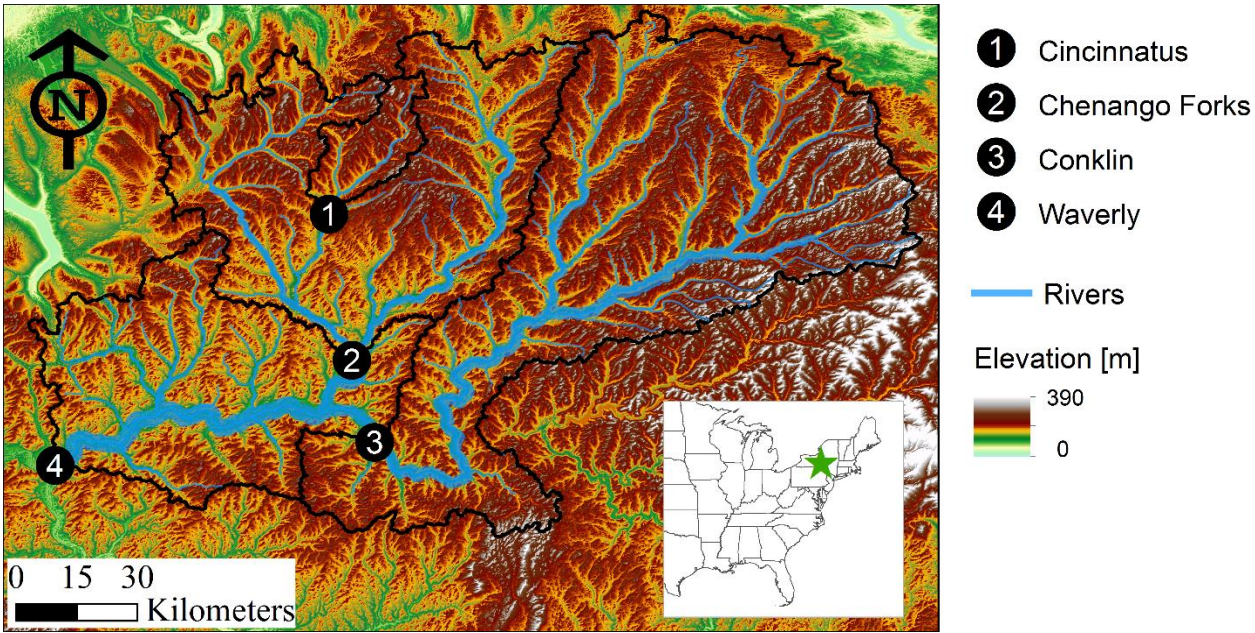
Figure 2. Map of the study area showing the terrain elevations, stream network, and the location of the selected gauged stations. The inset map shows the approximate location of the stduy area in the US.

Table 1. *Characteristics of the Selected Gauged Locations.*

3.2. Datasets

3.2.1 Meteorological Forecasts

NOAA’s latest global, medium-range ensemble reforecast dataset, the Global Ensemble



Forecast System Reforecast version 2 (GEFSRv2; https://www.esrl.noaa.gov/psd/forecasts/reforecast2/), was used as the forecast forcing. The following GEFSRv2 variables were used: precipitation, specific humidity, surface pressure, downward short and long wave radiation, u-v components of wind speed, and near-surface air temperature. The GEFSRv2 is an 11-member ensemble forecast generated by stochastically perturbing the initial numerical weather prediction model conditions using the ensemble transform technique with rescaling (Wei et al., 2008). The GEFSRv2 data are based on the same atmospheric model and initial conditions as the version 9.0.1 of the NOAA's Global Ensemble Forecast System, and runs at T254L42 (0.50^0 Gaussian grid spacing or ~ 55 km) resolution up to day 8. The 11-member reforecasts are generated every day at 00 Coordinated Universal Time. The GEFSRv2 forecast cycle consists of 3-hourly accumulations for the first 3 days and 6-hourly accumulations after that. To generate the ensemble streamflow forecasts, we used the first 7 days of GEFSRv2 data for the period 2004-2009. Table 2 summarizes key information about the GEFSRv2 dataset. Additional details about the GEFSRv2 can be found elsewhere (Hamill et al., 2013).

Table 2. *Summary and Main Characteristics of the Datasets Used in the Study*.

Dataset	Source	Horizontal Resolution [km^2]	Temporal Resolution [hour]	Variables
		Meteorological forecasts		
GEFSRv2	NCEP	~55 x 55 (0.5^0 x 0.5^0)	3 (days 1-3) and 6 (days 4-7)hourly accumulations	Precipitation, near-surface temperature, specific humidity, surface pressure, downward short and long wave radiation, and u-v components of wind speed
		Hydrometeorological observations		
NLDAS-2	NASA	~13 x 13 (0.125^0 x 0.125^0)	Hourly	Near-surface temperature, specific humidity, surface pressure, downward long and short wave radiation, and u-v components of wind speed
MPEs	MARFC	~4 x 4	Hourly	Gridded precipitation
Temperature	MARFC	~4 x 4	Hourly	Gridded temperature
Gauge discharge	USGS	-	Hourly	Streamflow

3.2.2. Hydrometeorological Observations

Four main observational datasets were used: multi-sensor precipitation estimates (MPEs), gridded near-surface air temperature, Phase 2 of the North American Land Data Assimilation System (NLDAS-2; https://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php), and daily streamflow. These observational datasets were used to calibrate and verify the hydrological models, perform the hydrological model simulations, and obtain initial conditions for the forecasting runs for the period 2004-2009. Both the MPEs and gridded near-surface air temperature data at 4 x 4 km^2

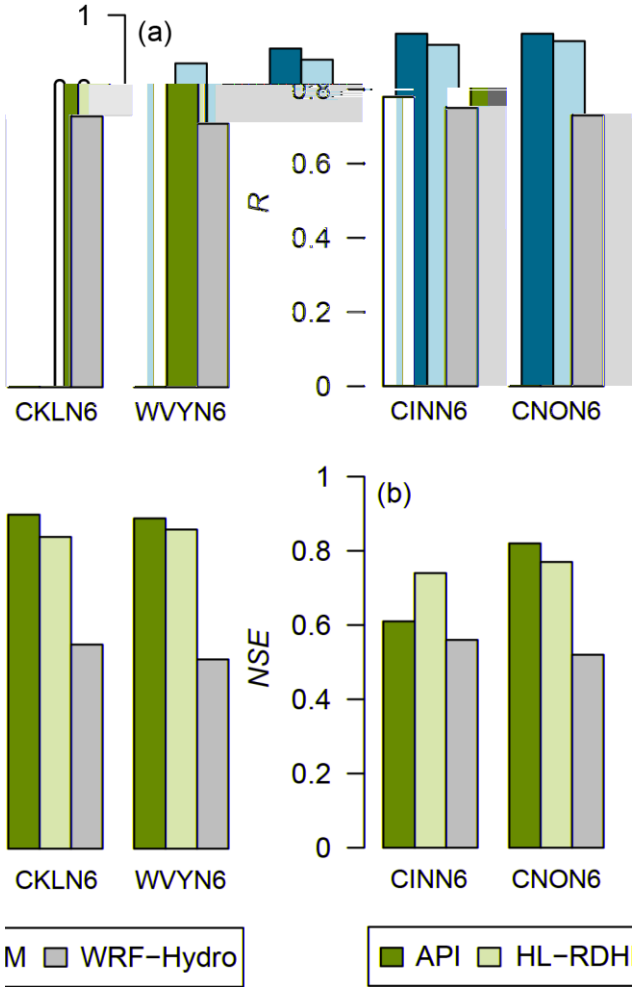
were obtained from the MARFC. Similar to the NCEP stage IV MPEs (Moore et al., 2014; Prat & Nelson, 2015), the MARFC MPE product combines radar estimated precipitation with in-situ gauge measurements to create a continuous time series of hourly, gridded precipitation observations. The gridded near-surface air temperature data were produced by the MARFC using multiple observation networks, including the meteorological terminal aviation routine weather report (METAR), USGS stations, and National Weather Service Cooperative Observer Program (Siddique & Mejia, 2017). Additionally, we used NLDAS-2 data for near-surface air temperature, specific humidity, surface pressure, downward long and short wave radiation, and u-v components of wind speed. The spatial resolution of the NLDAS-2 data is $1/8^{th}$-degree grid spacing while the temporal resolution is hourly. Further details about the NLDAS-2 data can be found elsewhere (Mitchell et al., 2004). To calibrate the hydrological models and verify the streamflow simulations and forecasts, daily streamflow observations for the selected gauged locations were obtained from the USGS. In total, 6 years (2004-2009) of hydrometeorological observations were used. Table 2 summarizes the observational datasets.

3.3. Hydrological Models

To generate the multimodel forecasts, we used the following three hydrological models: Antecedent Precipitation Index (API)-Continuous (Moreda et al., 2006), NOAA's Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM) (Koren et al., 2004), and the Weather Research and Forecasting Hydrological (WRF-Hydro) modeling system (Gochis et al., 2015). We selected these three hydrological models because they are relevant to operational forecasting in the US and represent varying levels of model structural complexity as well as different spatial resolutions and parameterizations. The selected models collectively represent a sufficiently diverse set of models favorable for multimodel forecasting. The description of each model and the details about the configuration, calibration, and performance of the models in simulation mode are provided in the supplemental information.

The models were used to simulate and forecast flows over the entire period of analysis (years 2004-2009, warm season only, May-October) at the selected gauge locations (Figure 1). The simulated flows were obtained by forcing the hydrological models with meteorological observations. The streamflow simulations were verified against daily observed flows for the entire period of analysis. Figure 3 summarizes the models' performance in simulation mode using the Pearson's correlation coefficient, *R*, and Nash-Sutcliffe efficiency, *NSE*, between the simulated and observed streamflows at daily resolution for the entire analysis period. The overall performance of the models was satisfactory (Figures 3a-b). API and HL-RDHM exhibited comparable performance while WRF-Hydro tended to underperform relative to API and HL-RDHM. The performance of the models is discussed further in Section 4.

Figure 3. Performance of the hydrological models in simulation mode over the entire period of analysis (2004-2009, May-October): (a) Pearson's correlation coefficient, *R*, and (b) Nash-Sutcliffe efficiency, *NSE*, between the daily simulated and observed flows.



3.4. Ensemble Streamflow Forecasts

To perform our forecast experiments, we generated and verified the following three different datasets of ensemble streamflow forecasts: i) raw single-model, ii) postprocessed single-model, and iii) multimodel. The raw single-model dataset consists of ensemble streamflow forecasts from each hydrological model without postprocessing. The postprocessed single-model dataset was generated by using QR to postprocess the raw ensemble streamflow forecasts from each

hydrological model. Lastly, the multimodel dataset was generated by optimally combining the ensemble forecasts from the different hydrological models using QR-BMA. As part of the multimodel dataset, we also generated an equal weight multimodel forecast by using the same weight, $1/K$, to combine the models rather than the optimal weights from QR-BMA. Additionally, for both the single-model and multimodel forecast datasets, we varied the number of ensemble members used (9 to 33 members) to perform different experiments.

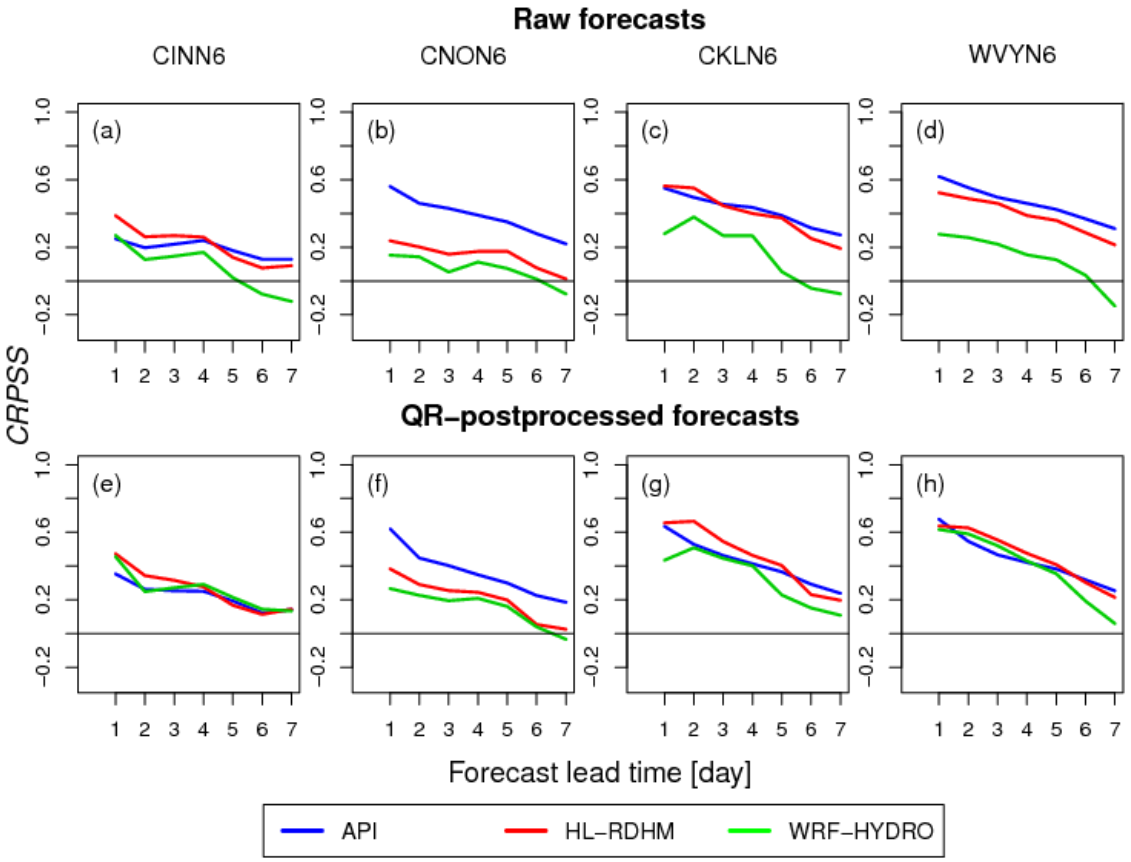
All the forecast datasets were verified across lead times of 1 to 7 days using 6 years of data (2004-2009) for the warm season only (May-October). To postprocess and verify both the single model and multimodel ensemble streamflow forecasts, a leave-one-out approach was implemented by using 4 years of forecast data to train the postprocessor and the remaining 2 years to verify the forecasts. The subdaily streamflow forecasts generated by the hydrological models were averaged over 24 hours to get the mean daily flow. Six-hourly streamflow forecasts were generated from API and HL-RDHM, and 3-hourly forecasts from WRF-Hydro. The mean daily ensemble streamflow forecasts were verified against mean daily streamflow observations for the selected gauged locations.

4. Results and Discussion

4.1. CRPSS Verification of the Single-Model Forecasts

4.1.1. Raw Ensemble Streamflow Forecasts

In terms of the *CRPSS* (relative to sampled climatology), the raw single-model ensemble streamflow forecasts remain skillful across lead times (1-7 days) and basins (Figures 4a-d), with the exception of WRF-Hydro that has slightly negative *CRPSS* values at the longer lead times (6-7 days). In Figures 4a-d, the *CRPSS* values tend overall to decline with increasing lead time, as might be expected since the weather uncertainties tend to grow and become more dominant of forecast skill as the lead time progresses (Siddique & Mejia, 2017). There is also a slight tendency for the *CRPSS* values to exhibit spatial scale-dependency. The *CRPSS* values for each



model tend to increase from the smallest (Figure 4a) to the largest (Figure 4d) basin across lead times. This tendency is, however, rather weak throughout all of our forecasts and it is somewhat more apparent for the API and HL-RHDM forecasts than for the WRF-Hydro (Figures 4a-d).

Figure 4. *CRPSS* (relative to sampled climatology) of the (a)-(d) raw and (e)-(h) QR-postprocessed single model ensemble streamflow forecasts versus the forecast lead. The *CRPSS* are shown for the four selected basins.

Across all lead times and basins (Figures 4a-d), the *CRPSS* values vary approximately from -0.15 (WRF-Hydro at the day 7 lead time; Figure 4d) to 0.6 (API at the day 1 lead time; Figure 4d). Contrasting the hydrological models, the performance of API and HL-RDHM is comparable, with the exception of CNON6 (Figure 4b) where API outperforms HL-RDHM. This is due to HL-RDHM having an unusually high percent simulation bias of -14.3 for CNON6 relative to API whose simulation bias is -5.8. The performance of the models in forecasting mode tends to mimic their performance in simulation mode (Figure 3). That is, API tends to

perform better than HL-RDHM and, in turn, both of these models tend to outperform WRF-Hydro. Deviations from this tendency, however, do emerge. For example, WRF-Hydro has similar forecasting skill as HL-RDHM at the day 1 lead time in CINN6 (Figure 4a), even though in this basin HL-RDHM performs better than WRF-Hydro in simulation mode. Similarly, API performs slightly better than HL-RDHM in forecasting mode at the later lead times (>4 days) in CINN6 (Figure 4a) but HL-RDHM shows better performance in simulation mode. Thus, the results obtained here in simulation mode do not always translate to similar performance in forecasting mode. This is not surprising given the nonlinear relationship between hydrological processes and weather forcings. It reinforces the need to verify hydrological models in both simulation and forecasting mode to gain a more complete understanding of model behavior.

The underperformance of WRF-Hydro, in both simulation and forecasting mode, in comparison to API and HL-RDHM may be due to several factors. One factor is likely to be the additional model complexity of WRF-Hydro. That is, WRF-Hydro requires more forcing inputs and parameters to be specified than the other two models. For example, in terms of forcings, HL-RDHM requires only precipitation and near-surface air temperature to be specified, whereas WRF-Hydro requires 7 different forcings. It is possible that any biases in the NLDAS-2 or GEFSRv2 forcings used here to configure the WRF-Hydro simulations and forecasts, respectively, could be affecting its performance. However, we evaluated (results not shown) for the WRF-Hydro streamflow forecasts the effect of each individual forcing on the *CRPSS* values and found that precipitation was the most dominant forcing. At least in forecasting mode, the additional forcings used by WRF-Hydro do not seem to have a strong influence on its forecast skill.

The determination of model parameter values for the WRF-Hydro is another factor that is likely affecting its performance. Although we calibrated selected WRF-Hydro parameter values, both manually and numerically (see supplemental information), there is generally less community knowledge about and experience with WRF-Hydro than API and HL-RDHM. The latter two have been around for much longer (e.g., Moreda et al., 2006; Koren et al., 2004; Anderson et al., 2006; Reed et al., 2004) than WRF-Hydro. In the future, a more in-depth sensitivity analysis of the WRF-Hydro model parameters could be beneficial. Nonetheless, the performance of WRF-Hydro in this study is comparable to those previously reported in the literature (Givati et al., 2016; Kerandi et al., 2017; Naabil et al., 2017; Salas et al., 2018; Silver et al., 2017; Yucel et al., 2015).

4.1.2. Postprocessed (Single-Model) Ensemble Streamflow Forecasts

We used QR to postprocess the raw single-model ensemble streamflow forecasts. Using the *CRPSS* (relative to sampled climatology) to assess the forecast skill (Figures 4e-h), we found that the postprocessed single-model ensemble streamflow forecasts show, overall, skill improvements relative to the raw forecasts. The relative improvements are more noticeable for the WRF-Hydro. For example, at WVYN6 (Figure 4d), the raw WRF-Hydro forecasts have a *CRPSS* value of ~0.27 at the day 1 lead time, and that value increases to ~0.6 after postprocessing (Figure 4h).

Interestingly, the *CRPSS* values for the postprocessed single-model forecasts reveal that, after postprocessing, the models have comparable skill across lead times and basins (Figures 4e-h), perhaps with the exception of CNON6 (Figure 4f) where API tends to outperform the other models. This indicates that the streamflow forecasts are influenced by systematic biases and, in this case, those biases are stronger in WRF-Hydro than in the other models. Such streamflow forecast biases result from the combined effect of biases in the weather forcings and hydrological

models. In regards to the former, precipitation forecasts from the GEFSRv2 are characterized by an underforecasting bias in our study region (Sharma et al., 2017; Siddique et al., 2015), particularly at the longer lead times. This underforecasting bias affects all of our hydrological model forecasts so it is unlikely to be the cause of the strong biases seen in the WRF-Hydro forecasts.

Hydrological model biases appear to have a strong effect on the performance of WRF-Hydro, given the relatively mild skill gains from postprocessing for the API and HL-RHDM models and the larger gains for WRF-Hydro (Figures 4e-h). Nonetheless, the QR postprocessor is able in this case to handle those biases. This suggests that models with simple structure (e.g., API which is spatially lumped and has fewer parameters) may benefit less from postprocessing while models with complex structure (e.g., WRF-Hydro which is spatially distributed and has more parameters) may be good candidates for postprocessing. It is also possible that systematic biases in the WRF-Hydro could be reduced through improved parameter sensitivity analysis and calibration, as opposed to statistical postprocessing.

Another interesting outcome from the postprocessed single-model results is that the ranking of the models, in terms of the *CRPSS*, varies depending on the lead time and basin. For example, both HL-RDHM and WRF-Hydro tend to slightly outperform API at the day 1 lead time in Figure 4e, but API outperforms both models at the later lead times (>6 days) in Figures 4f-h. This is important because it indicates that there is no single model that consistently outperforms the other models. In other words, it is not possible, at least in terms of the *CRPSS*, to choose one model as the best in all cases. This suggests that it may be possible to maximize forecast skill across lead times and basins by optimally combining the outputs from the different models, as opposed to relying on a single model. It shows that multimodel forecasting may be a viable option to enhance streamflow predictions.

4.2. CRPSS Verification of the Multimodel Forecasts

We now examine with the *CRPSS* the ability of multimodel forecasts to improve streamflow predictions. For this, the *CRPSS* is again plotted against the forecast lead time for the selected basins (Figure 5). In Figure 5, the following three different multimodel forecasting experiments are shown: i) equal weight, ii) 9-m, and iii) 33-m. For the equal weight experiment, the same weight, $1/K$, was used to combine the predictive distribution of the streamflow forecasts from each hydrological model. That is, instead of using the optimal weights from QR-BMA, the same weight was used to form a 9-member multimodel forecast. For the 9-m and 33-m experiments, we used 3 and 11 members per model, respectively, to obtain a multimodel forecast with QR-BMA. In the 9-m and 33-m experiments, the weights were optimized with QR-BMA employing different number of ensemble members, 3 members per model in the case of 9-m and 11 members per model in 33-m. Additionally, the reference system used to compute the *CRPSS* values in Figure 5 consists of the postprocessed ensemble streamflow forecasts from API, as opposed to sampled climatology. We selected API as the reference system since this is currently the regional operational model being used to generate streamflow forecasts in our study area.

We found that the 33-m multimodel forecasts result in higher *CRPSS* values than API across lead times and basins (Figure 5). The 9-m multimodel forecasts perform similarly to the 33-m forecasts, but in a few cases (e.g., Figure 5c at the day 5 lead time) the 9-m forecasts result in lower (negative) *CRPSS* values than API. The equal weight experiment is only able to improve the *CRPSS* values at the initial lead times (<3 or 4 days; Figure 5), while at the later lead times its *CRPSS* values are lower than API. CNON6 offers an interesting case to further compare the

single-model and multimodel forecasts. In the single-model forecasts for CNON6 (Figure 4f), API tends to clearly outperform the other models. Despite the better performance of API, the multimodel forecasts are still able to improve the skill for CNON6 relative to the performance of API, with the largest improvement being ~0.16 at the day 7 lead time for the 33-m experiment.

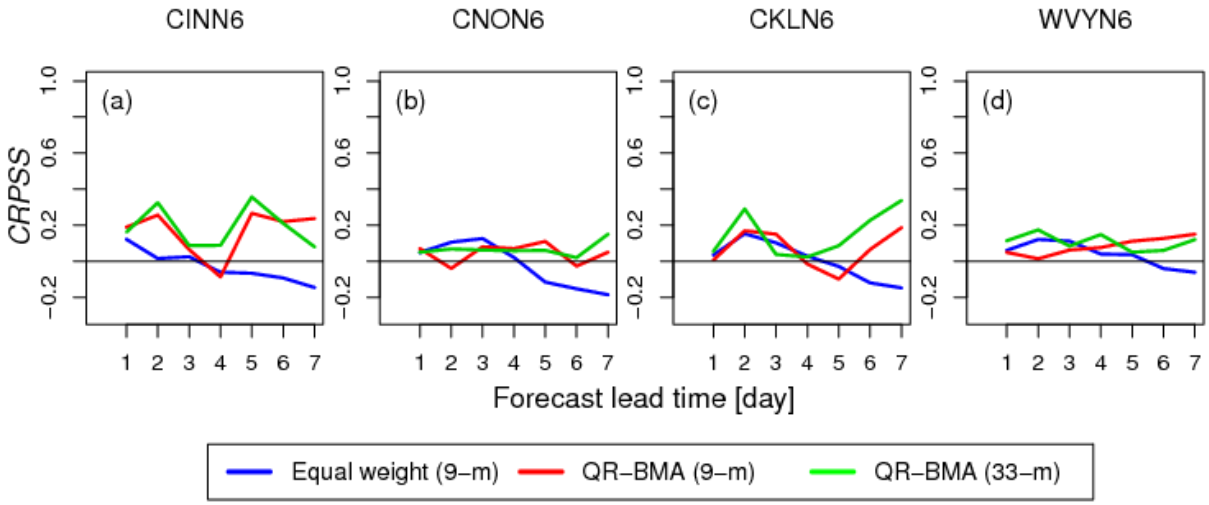


Figure 5. *CRPSS* of the multimodel ensemble streamflow forecasts versus the forecast lead time for a) CINN6, b) CNON6, c) CKLN6, and d) WVYN6. The *CRPSS* is plotted with reference to the QR-postprocessed API forecasts. Three different experiments are shown: equal weight (9-m), QR-BMA (9-m), and QR-BMA (33-m). The equal weight experiment uses the same weight to combine the predictive distribution of the streamflow forecasts from each hydrological model. The 9-m and 33-m experiments use 3 and 11 members per model, respectively, to obtain a multimodel forecast with optimal weights using QR-BMA.

In sum, the multimodel forecasts reveal skill improvements relative to API, which may be considered here the best performing model in terms of the overall simulation and raw forecasts results; the optimal weights from QR-BMA result in more skillful multimodel forecasts than using equal weights, particularly at the later lead times (>3 days); and increasing the ensemble size of the multimodel forecasts results in relatively mild skill gains. Several studies have investigated the source of improvements (skill gains) from multimodel forecasts (Hagedorn et al., 2012; Weigel et al., 2008, 2009). Those studies have found that multimodel forecasts can improve predictions by error cancellation and correcting deficiencies (underdispersion) in the ensemble spread of the single models. These sources of skill gain appear to be mainly statistical. This way of understanding the benefits of multimodel forecasts does not consider whether a particular model contributes additional information to the forecasts. Considering the latter is important to be able to justify adding any new models to an existing forecasting system. Another way to assess the source of improvements from multimodel forecasts that accounts for the contribution of model information, signal as opposed to noise, is through *CMI*, which we do next.

4.3. Skill Assessment Using Conditional Mutual Information

We used *CMI* to determine whether the skill improvements from the multimodel forecasts are dominated by model diversity or increased ensemble size. To this end, *CMI* was computed

using equations (14) and (15), together with the ensemble mean forecast, at lead times of 1-7 days for the selected basins (Figure 6). In Figure 6, the following three different experiments are shown: i) 9-m single model, ii) 9-m multimodel, and iii) 33-m multimodel.

The 9-m single-model experiment consists of a 3-member single-model forecast combined with a 6-member ensemble from the same model. The experiment was repeated for each of the models used. The results from this experiment are shown in Figures 6a-c. In the 9-m multimodel experiment, a 3-member single model ensemble from one of the models was combined with a 6-member ensemble from the remaining other two models (3 members per model; Figures 6d-f). The last experiment, 33-m multimodel, was the same as the 9-m multimodel experiment but using instead 33 members (Figures 6g-i). That is, an 11-member single-model ensemble from one of the models was combined with a 22-member ensemble from the remaining two other models (11 members per model). Thus, the ensemble size for the first and second experiments was 9, and for the third experiment was 33. Note that in all three experiments the ensemble members were randomly selected to form any of the model combinations considered. The single-model ensembles were selected from the raw forecasts, whereas the multimodel forecasts were selected from the postprocessed QR-BMA output. Raw single-model and postprocessed multimodel forecasts were combined to emulate basic operational conditions. Note that we also tried combining only postprocessed single-model and multimodel forecasts and the results (not shown) were similar.

For the first experiment, we used equations (14) and (18) to obtain a theoretical upper bound for *CMI*. This theoretical bound represents the potential skill gain from the ensemble size alone. We found that the theoretical bound is in this case equal to 0.090. Figure 6a-c shows that indeed the empirical *CMI* values for the 9-m single-model forecasts tend to be less than or around 0.090 for all three models across lead times and basins. The 9-m single-model *CMI* values tend to be greater for API than HL-RDHM and WRF-Hydro. This indicates that the less complex model, API, is able to maximize the skill gains from the ensemble size alone. For example, in terms of the *CRPSS*, the raw single-model forecasts from API and HL-RDHM have comparable skill in the case of CKLN6 (Figure 4c) and WVYN6 (Figure 4d). In contrast, the 9-m single model *CMI* values tend to be greater for API than HL-RDHM in both cases, CKLN6 and WVYN6 (Figures 6a and 6b), particularly at the longer lead times. This ability of API to maximize the benefits from ensemble size alone may be due to API having a lesser impact on the weather ensembles, i.e., contributing less uncertainty to the streamflow forecasts. Also, in Figures 6a-c, the tendency

is for the *CMI* values to increase some with the lead time for all the basins. This is more apparent for API and HL-RDHM than WRF-Hydro.

Figure 6. *CMI* of the ensemble streamflow forecasts versus both the basin and forecast lead time for three different experiments: a-c) 9-m single model, d-f) 9-m multimodel, and g-i) 33-m multimodel forecasts. The 9-m single model experiment consists of a 3-member single model forecast from one of the hydrological models combined with a 6-member ensemble from the same model. In the 9-m multimodel experiment, a 3-member single model ensemble forecast from one of the models is combined with a 6-member ensemble from the remaining other two models (3 members from each model). The last experiment, 33-m multimodel, is the same as the 9-m multimodel experiment but using instead 33 members (11 members from each model).

Contrasting the *CMI* values between the 9-m single-model (Figures. 6a-c) and 9-m multimodel (Figures. 6d-f) experiment, it is apparent that the multimodel forecasts have

substantially greater *CMI* values than the single-model forecasts across lead times and basins. This indicates that any of the single-model forecasts (API, HL-RDHM or WRF-Hydro) can be improved by combining them with forecasts from the other models. Indeed, this improvement is dominated by model diversity rather than increased ensemble size alone. Although the multimodel forecasts show skill gains at all the lead times, the tendency is for the *CMI* values to increase with the lead time, suggesting that the multimodel forecasts may be particularly useful for improving medium-range streamflow forecasts.

To further examine the hypothesis that improvements in *CMI* are dominated by model diversity rather than the ensemble size, the *CMI* values from the 9-m multimodel experiment (Figures 6d-f) can be compared against the values from the 33-m multimodel experiment (Figures 6g-i). From this comparison, it is seen that the *CMI* values for these two experiments are, overall, very similar across lead times and basins. This further supports that incorporating additional information by adding new models plays a bigger role than the ensemble size in enhancing the skill of the multimodel forecasts. The results in Figure 6 indicate that hydrological multimodel forecasting can be a viable approach to improve streamflow forecasts at short- and medium-range timescales. They suggest that model diversity may be a more important consideration than the ensemble size when trying to enhance the skill of streamflow forecasts.

We also tested the effect on the *CMI* values of using postprocessed single-model forecasts, as opposed to raw forecasts. Thus, we calculated *CMI* (results not shown) for each basin and lead time using the QR postprocessed single-model forecasts, i.e., the experiments in Figure 6 were repeated using the postprocessed single-model forecasts. We found that, as was the case with the raw forecasts, the *CMI* values for the multimodel combinations exceeded the theoretical upper bound of 0.090, and the *CMI* values remained very similar after increasing the ensemble size, i.e., between the 9-m and 33-m multimodel experiments. Thus, the ability of model diversity to enhance the skill of the streamflow forecasts is independent of whether raw or postprocessed single-model forecasts are used.

Additionally, the *CMI* values for all the different experiments in Figure 6 were recomputed (results not shown) in streamflow space using the approach by Meyer (2008). Although a theoretical upper bound is not available for this approach, the *CMI* values in streamflow space for the multimodel forecasts tended to be noticeably greater than the values for the single-model forecasts for most lead times. Moreover, differences in the *CMI* values between the 9-m and 33-m multimodel forecasts were only marginal. Thus, the results for the experiments in Figure 6 using *CMI* values computed in both real (streamflow) and Gaussian space, overall, exhibited similar trends. This is again indicative of the ability of model diversity to enhance forecast skill beyond the improvements achievable by ensemble size alone.

5. Summary and Conclusions

In this study, we generated single-model ensemble streamflow forecasts at short- to medium-range lead times (1-7 days) from three different hydrological models: API, HL-RDHM, and WRF-Hydro. These models were selected because they represent different types of hydrological models with varying structures and parameterizations. API is a spatially lumped model; HL-RDHM is a conceptual, spatially distributed hydrological model; and WRF-Hydro is a land surface model. By forcing each hydrological model with GEFSRv2 data, single-model ensemble streamflow forecasts were generated for four nested basins of the US NBSR basin over the period 2004-2009, and the warm season (May-October). The single-model forecasts were used to generate multimodel forecasts using a new statistical postprocessor, namely QR-BMA. QR-

BMA uses first QR to correct systematic biases in the single-model forecasts and, in a subsequent step, BMA to optimally combine the predictive distribution from each model. To further understand the performance and behavior of the multimodel forecasts, we performed different ensemble streamflow forecast experiments by varying the number of ensemble members, models, and weights used to create the multimodel forecasts.

From the forecast experiments performed, we found that the raw single-model ensemble streamflow forecasts from both API and HL-RHDM tended to outperform, in terms of the CRPSS, the forecasts from WRF-Hydro across lead times and basins. However, after postprocessing the raw single-model forecasts using QR, we found that the CRPSS performance of the individual models was mostly comparable across lead times and basins. In terms of the multimodel ensemble streamflow forecasts, we found that the implementation of QR-BMA tended to improve the skill of the forecasts relative to the performance of API, which can be considered here the best performing model in terms of the raw single-model forecasts. Additionally, we compared the forecasts from QR-BMA against an equal-weight experiment, where each model was assigned the same weight. We found from this experiment that the optimal-weight forecasts from QR-BMA outperform the equal-weight forecasts. The latter was particularly evident at the later lead times (> 3 days).

Lastly, we used *CMI* to distinguish the source of the improvements for the multimodel forecasts. We found that skill enhancements across lead times and basins are largely dominated by model diversity and that increasing the ensemble size has only a small influence on the *CMI* values. This is important because it indicates that in an operational setting the combination of different hydrological models, as opposed to increasing the ensemble size of a single model, may be a more effective approach to improve forecast skill. It also highlights that there is no single model that can be considered best in all forecasting cases, instead the benefits or strengths of different models can be combined to produce the best forecast. Importantly, the benefits from using different models are, in this case, not due to the noise reduction associated with the ensemble size but with the ability of each model to contribute additional information to the forecasts.

Appendix A: Implementation of the Expectation Maximization Algorithm

We describe here the steps followed to implement the EM algorithm. The description uses the variables and notation previously defined in Subsection 2.1. To implement the EM algorithm, the latent variable $z_k^{t,i}$ is introduced, which has a value of 1 if the k^{th} model ensemble is the best prediction at time step i and a value of 0 otherwise. The EM algorithm starts with an initial weight and variance for each model set to

$$w_{k,Iter-1}^{t} = \frac{1}{K}, \text{ and} \tag{A1}$$

$$\sigma_{k,Iter-1}^{2,t} = \frac{1}{K}\sum_{i=1}^{T}\frac{\sum_{k=1}^{K}(\Delta_{NQT}^{t,i} - f_{k,NQT}^{t,i})^2}{T}, \tag{A2}$$

allowing the calculation of an initial log-likelihood

$$l(\theta_{Iter-1}) = \sum_{i=1}^{T}\log(\sum_{k=1}^{K} w_{k,Iter-1}^{t} g(\Delta_{NQT}^{t,i} \mid f_{k,NQT}^{t,i}, \sigma_{k,Iter-1}^{2,t})), \tag{A3}$$

where T is the length of the training period extending over the time steps $i \in [1,T]$. After initializing the weight and variance for each model, the EM algorithm alternates iteratively between an expectation and maximization step until a convergence criteria is satisfied. In the expectation step, the $\hat{z}_k^{t,i}$ for each time step is estimated given the initial values of the weight and variance as

$$\hat{z}_{k,Iter}^{t,i} = \frac{w_{k,Iter-1}^{t} g\left(\Delta_{NQT}^{t,i} \mid f_{k,NQT}^{t,i}, \sigma_{k,Iter-1}^{2,t}\right)}{\sum_{k=1}^{K} w_{k,Iter-1}^{t} g\left(\Delta_{NQT}^{t,i} \mid f_{k,NQT}^{t,i}, \sigma_{k,Iter-1}^{2,t}\right)}. \quad \text{(A4)}$$

In the subsequent maximization step, the values of the weight and variance are updated using the current estimate of $z_{k,Iter}^{t,i}$ as follows

$$w_{k,Iter}^{t} = \frac{1}{T}\sum_{i=1}^{T} \hat{z}_{k,Iter}^{t,i}, \text{ and}$$

$$\sigma_{k,Iter}^{2,t} = \frac{\sum_{i=1}^{T} \hat{z}_{k,Iter}^{t,i}\left(\Delta_{NQT}^{t,i} - f_{k,NQT}^{t,i}\right)^2}{\sum_{i=1}^{T} \hat{z}_{k,Iter}^{t,i}}. \quad \text{(A5)}$$

The log-likelihood function in equation (A3) is then recomputed using the updated weight and variance as

$$l(\theta_{Iter}) = \sum_{i=1}^{T} \log\left(\sum_{k=1}^{K} w_{k,Iter}^{t} g(\Delta_{NQT}^{t,i} \mid f_{k,NQT}^{t,i}, \sigma_{k,Iter}^{2,t})\right). \quad \text{(A6)}$$

The expectation and maximization steps are iterated until the improvement in the log-likelihood is no less than some pre-defined tolerance, i.e. $\left(\left|\left(l\left(\theta_{Iter}\right) - l\left(\theta_{Iter-1}\right)\right)\right|\right) < tol$, in this case tol=10^{-6}.

Supporting Information: Description of the Hydrological Models Used to Generate the Multimodel Ensemble Streamflow Forecasts

API

API is the current operational rainfall-runoff model used by the NOAA's Middle Atlantic River Forecast Center (MARFC) to generate daily streamflow forecasts. The API model was developed by NOAA (Nemec & Sittner, 1982; Sittner et al., 1969). It is a conceptual, spatially lumped hydrological model (Moreda et al., 2006). The model uses a graphical technique that consists of four quadrants to compute surface runoff and baseflow. The first quadrant accounts for the seasonal relationship between API and current soil moisture conditions referred to as the Antecedent Index (AI). The second quadrant adjusts the AI value from the first quadrant to account for the effects of soil moisture. The third quadrant computes incremental surface runoff based on surface and overall soil moisture conditions. Finally, the fourth quadrant computes the portion of precipitation that enters groundwater storage and becomes baseflow runoff.

The API streamflow simulations and ensemble forecasts used in this study were generated by the MARFC. For calibration purposes, the API was forced with mean areal precipitation and temperature estimates from gauge observations. To run API, the model parameters associated with surface runoff, vegetation, snow, and frozen ground, among others, were determined via calibration. The MARFC calibrated the model manually. The overall performance of the API

simulation runs was satisfactory during the verification period of 2004-2009 (May to October). The *R* values ranged from 0.78 to 0.95 and the *NSE* values from 0.61 to 0.89. Other than CINN6, the performance of API tended to be similar for the selected basins. The smallest basin, CINN6, showed the lowest performance with *R* and *NSE* values of 0.78 and 0.61, respectively. Additional details about the model can be found elsewhere (Moreda et al., 2006). The API ensemble streamflow forecasts were generated by forcing the calibrated model with GEFSRv2 ensemble precipitation and near-surface air temperature forecasts.

HL-RDHM

HL-RDHM is a conceptual, distributed hydrological model developed by NOAA (Koren et al., 2004). Within HL-RDHM, we implemented the heat transfer version of the Sacramento Soil Moisture Accounting model (SAC-HT) to represent rainfall-runoff generation, and the SNOW-17 submodel to represent snow accumulation and melt (Koren et al., 2004). The hillslope runoff, generated at each grid cell by SAC-HT and SNOW-17, was routed to the stream network using a nonlinear kinematic wave algorithm. A similar algorithm was used to route flows along the stream network (Koren et al., 2004). We ran HL-RDHM in fully distributed mode at a spatial resolution of 2 x 2 km^2. To perform the simulation runs with HL-RDHM, the model was forced with gridded precipitation (MPEs) and near-surface air temperature observations provided by the MARFC. As was the case with API, GEFSRv2 ensemble precipitation and near-surface air temperature forecasts were used to force HL-RDHM and generate the ensemble streamflow forecasts. Further information about the HL-RDHM model can be found elsewhere (Koren et al., 2004; Siddique & Mejia, 2017).

To calibrate HL-RDHM, a-priori parameter estimates based on previous studies (Anderson et al., 2006; Reed et al., 2004) were first manually adjusted. Once the manual changes did not yield noticeable improvements in model performance, the parameter values were tuned up using the Stepwise Line Search (SLS) approach (Kuzmin et al., 2008; Kuzmin, 2009). The square root of the mean square errors (i.e., the difference between observed and simulated flows) was used as the objective function in SLS. We adjusted 10 out the 17 SAC-HT parameters associated with each model grid cell. The most sensitive parameters were found to be the upper and lower soil zones transport and storage parameters, as well as the stream routing parameters. Note that when calibrating HL-RDHM we adjusted the parameter fields rather than the actual parameter values at each grid cell using a multiplier approach (Kuzmin et al., 2008; Kuzmin, 2009). We used 3 years (2003-2005) of streamflow data to calibrate HL-RDHM. The model simulations were performed for the period 2004-2009 (May to October), with the year 2003 used as warm-up. Overall, the performance of the HL-RDHM simulation runs over the period 2004-2009 was satisfactory. The performance was similar for the selected basins, with *R* values ranging between 0.87 and 0.93, and *NSE* values between 0.74 and 0.86.

WRF-Hydro

WRF-Hydro, the hydrological package extension to the WRF model, is a fully parallelized, community modeling architecture. Here we used WRF-Hydro version 3.0 in uncoupled mode. WRF-Hydro was configured to use the land surface model Noah with multi-parameterization (Noah-MP) options to represent surface and subsurface hydrological processes. We used the baseflow bucket model to represent baseflow to the stream network and a fully-unsteady, explicit, finite difference, diffusive wave formulation to route surface flows. Additionally, we used the Geographic Information System (ArcGIS version 10.3) Preprocessing Tool version 4.0

(Gochis et al., 2015) to create the data layers required by WRF-Hydro to model terrestrial overland flow, subsurface flow, and the channel routing process. The 1 arc-second (30 m) version 2 of the National Hydrography Dataset (NHDPlusV2) was used as a raster input to the Preprocessing Tool to delineate different water features. WRF-Hydro was ran at a spatial resolution of 1 x 1 km^2 to generate both streamflow simulations and forecasts. As was the case with API and HL-RDHM, GEFSRv2 data were used as forcing to generate the WRF-Hydro ensemble streamflow forecasts. In the case WRF-Hydro, the GEFSRv2 data consisted of ensemble precipitation, near-surface air temperature, specific humidity, surface pressure, downward short and long wave radiation, and u-v components of wind speed.

Two years (2004-2005; May-October) of streamflow data were used to calibrate WRF-Hydro. The first year (2004; January-April) was used to warm-up the model. A shorter warm-up period than HL-RDHM was selected to ameliorate computational demand. To perform the WRF-Hydro simulation runs, we forced the model with gridded precipitation, MPEs, and for the remainder forcing variables (i.e., near-surface air temperature, specific humidity, surface pressure, downward long and short wave radiation, and u-v components of wind speed) NLDAS-2 data were used. In order to minimize the number of model runs during calibration, we implemented a stepwise manual adjustment approach (Yucel et al., 2015), i.e., once a parameter value was calibrated its value was kept fixed during the calibration of subsequent parameters. We adjusted eight different parameters associated with soil transport, surface runoff, as well as groundwater and channelized flows. Out of all the adjusted parameters, the most sensitive parameters were the pore size distribution index (BEXP), saturated hydraulic conductivity (DKSAT), surface runoff parameter (REFKDT), surface retention depth scaling parameter (RETDEPRT), and the channel Manning roughness coefficient (MANN). After the manual calibration, the most sensitive parameter values were fine-tuned using an optimization algorithm, namely dynamically dimension search (DDS) (Tolson & Shoemaker, 2007). In the DDS algorithm, we used *NSE* as the objective function. The simulation performance of the WRF-Hydro over the entire analysis period of 2004-2009 (May-October) was reasonable. The *R* and *NSE* values were in the range 0.71-0.75 and 0.51-0.56, respectively, for the selected basins. These performance statistic values compare well with results from previous studies (Givati et al., 2016; Kerandi et al., 2017; Naabil et al., 2017; Salas et al., 2018; Silver et al., 2017; Yucel et al., 2015).

References

Abdi, H. (2007). Part (semi partial) and partial regression coefficients. *Encyclopedia of measurement and statistics*, 736-740.

Anderson, R. M., Koren, V. I., & Reed, S. M. (2006). Using SSURGO data to improve Sacramento Model a priori parameter estimates. *Journal of Hydrology*, 320, 103–116.

Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research, 43*(1), W01403, doi:10.1029/2005WR004745.

Bastola, S., Misra, V., & Li, H. (2013). Seasonal hydrological forecasts for watersheds over the southeastern United States for the boreal summer and fall seasons. *Earth Interactions, 17*(25), 1-22.

Becker, E., den Dool, H. v., & Zhang, Q. (2014). Predictability and forecast skill in NMME. *Journal of Climate, 27*(15), 5891-5906.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute, 4*(510), 126.

Bohn, T. J., Sonessa, M. Y., & Lettenmaier, D. P. (2010). Seasonal Hydrologic Forecasting: Do Multimodel Ensemble Averages Always Yield Improvements in Forecast Skill? *Journal of Hydrometeorology, 11*(6), 1358-1372. doi:10.1175/2010JHM1267.1.

Bosart, L. F. (1975). SUNYA experimental results in forecasting daily temperature and precipitation. *Monthly Weather Review, 103*(11), 1013-1020.

Brown, J. D., He, M., Regonda, S., Wu, L., Lee, H., & Seo, D.-J. (2014). Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *Journal of Hydrology, 519*, 2847-2868.

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y., & Wei, M. (2005). A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Monthly Weather Review, 133*(5), 1076-1097. doi:10.1175/MWR2905.1.

Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. *Elements of information theory, 2*, 1-55.

Davolio, S., Miglietta, M. M., Diomede, T., Marsigli, C., Morgillo, A., & Moscatello, A. (2008). A meteo-hydrological prediction system based on a multi-model approach for precipitation forecasting. *Natural Hazards and Earth System Science, 8*(1), 143-159.

DelSole, T. (2007). A Bayesian Framework for Multimodel Regression. *Journal of Climate, 20*(12), 2810-2826. doi:10.1175/JCLI4179.1.

DelSole, T., Nattala, J., & Tippett, M. K. (2014). Skill improvement from increased ensemble size and model diversity. *Geophysical Research Letters, 41*(20), 7331-7342. doi:10.1002/2014GL060133.

DelSole, T., Yang, X., & Tippett, M. K. (2013). Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quarterly Journal of the Royal Meteorological Society, 139*(670), 176-183. doi:10.1002/qj.1961.

Du, J., DiMego, G., Tracton, S., & Zhou, B. (2003). NCEP Short-Range Ensemble Forecasting (SREF) System: Multi-IC, Multi-model and Multi-physics approach. *Research Activities in Atmospheric and Oceanic Modelling*, CoteJ (ed). Report 33, CAS/JSC Working Group

Numerical Experimentation (WGNE), WMO/TD-No. 1161: WMO, Geneva, Switzerland 5.09–5.10.

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources, 30*(5), 1371-1386. doi:http://dx.doi.org/10.1016/j.advwatres.2006.11.014.

Fraedrich, K., & Leslie, L. M. (1987). Combining Predictive Schemes in Short-Term Forecasting. *Monthly Weather Review, 115*(8), 1640-1644. doi:10.1175/1520-0493(1987)115<1640:CPSIST>2.0.CO;2.

Fraedrich, K., & Smith, N. R. (1989). Combining Predictive Schemes in Long-Range Forecasting. *Journal of Climate, 2*(3), 291-294. doi:10.1175/1520-0442(1989)002<0291:CPSILR>2.0.CO;2.

Fraley, C., Raftery, A. E., & Gneiting, T. (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review, 138*(1), 190-202.

Fritsch, J. M., Hilliker, J., Ross, J., & Vislocky, R. L. (2000). Model Consensus. *Weather and Forecasting, 15*(5), 571-582. doi:10.1175/1520-0434(2000)015<0571:MC>2.0.CO;2

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., & Butts, M. B. (2004). Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *Journal of Hydrology, 298*(1–4), 222-241. doi:http://dx.doi.org/10.1016/j.jhydrol.2004.03.037.

Gitro, C. M., Evans, M. S., & Grumm, R. H. (2014). Two Major Heavy Rain/Flood Events in the Mid-Atlantic: June 2006 and September 2011. *Journal of Operational Meteorology, 2*(13), 152–168. https://doi.org/10.15191/nwajom.2014.0213.

Givati, A., Gochis, D., Rummler, T., & Kunstmann, H. (2016). Comparing one-way and two-way coupled hydrometeorological forecasting systems for flood forecasting in the Mediterranean region. *Hydrology, 3*(2), 19. http://dx.doi.org/10.3390/hydrology3020019.

Gochis, D., Yu, W., & Yates, D. (2015). The WRF-Hydro Model Technical Description and User's Guide, Version 3.0, NCAR Technical Document, 120 pp , NCAR, Boulder, Colorado.

Gyakum, J. R. (1986). Experiments in temperature and precipitation forecasting for Illinois. *Weather and Forecasting, 1*(1), 77-88.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., & Palmer, T. (2012). Comparing TIGGE multimodel forecasts with reforecast- calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society, 138*(668), 1814-1827.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., et al. (2013). NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bulletin of the American Meteorological Society, 94*(10), 1553-1565. doi:10.1175/BAMS-D-12-00014.1.

Hamill, T. M., & Colucci, S. J. (1997). Verification of Eta–RSM Short-Range Ensemble Forecasts. *Monthly Weather Review, 125*(6), 1312-1327. doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting, 15*(5), 559-570.

Hopson, T. M., & Webster, P. J. (2010). A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *Journal of Hydrometeorology, 11*(3), 618-641.

Hsu, K.-l., Moradkhani, H., & Sorooshian, S. (2009). A sequential Bayesian approach for hydrologic model selection and prediction. *Water Resources Research, 45*(12), n/a-n/a. doi:10.1029/2008WR006824.

Jessup, S. M., & DeGaetano, A. T. (2008). A Statistical Comparison of the Properties of Flash Flooding and Nonflooding Precipitation Events in Portions of New York and Pennsylvania. *Weather and Forecasting, 23*(1), 114-130. doi:10.1175/2007waf2006066.1.

Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*: Academic press, London.

Kerandi, N., Arnault, J., Laux, P., Wagner, S., Kitheka, J., & Kunstmann, H. (2017). Joint atmospheric-terrestrial water balances for East Africa: a WRF-Hydro case study for the upper Tana River basin. *Theoretical and Applied Climatology*, 1-19. . http://dx.doi.org/10.1007/s00704-017-2050-8.

Kinney, J. B., & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences, 111*(9), 3354-3359.

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., et al. (2013). The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society, 95*(4), 585-601. doi:10.1175/BAMS-D-12-00050.1.

Koenker, R. (2005). Quantile regression, Cambridge University Press, Cambridge, 38, https://doi.org/10.1017/CBO9780511754098.

Kollet, S., Sulis, M., Maxwell, R. M., Paniconi, C., Putti, M., Bertoldi, G., et al. (2017). The integrated hydrologic model intercomparison project, IH- MIP2: A second set of benchmark results to diagnose integrated hydrology and feedbacks. *Water Resources Research, 53*(1), 867-890.

Koren, V., Reed, S., Smith, M., Zhang, Z., & Seo, D.-J. (2004). Hydrology laboratory research modeling system (HL-RMS) of the US national weather service. *Journal of Hydrology, 291*(3–4), 297-318. doi:http://dx.doi.org/10.1016/j.jhydrol.2003.12.039.

Krishnamurti, T. N. (2003). Methods, systems and computer program products for generating weather forecasts from a multi-model superensemble. U.S. Patent 6535817 B1, filed 13 November 2000, issued 18 Mar 2003.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., et al. (1999). Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble. *Science, 285*(5433), 1548.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., et al. (2000). Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate, 13*(23), 4196-4216.

Krzysztofowicz, R. (1997). Transformation and normalization of variates with specified distributions. *Journal of Hydrology, 197*(1), 286-292.

Kuzmin, V. (2009). Algorithms of automatic calibration of multi-parameter models used in operational systems of flash flood forecasting. *Russian Meteorology and Hydrology*, 34, 473–481.

Kuzmin, V., Seo, D.-J., & Koren, V. (2008). Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search. *Journal of Hydrology*, 353, 109–128.

Liang, Z., Wang, D., Guo, Y., Zhang, Y., & Dai, R. (2013). Application of Bayesian Model Averaging Approach to Multimodel Ensemble Hydrologic Forecasting. *Journal of Hydrologic Engineering, 18*(11), 1426-1436. doi:10.1061/(ASCE)HE.1943-5584.0000493.

Madadgar, S., & Moradkhani, H. (2014). Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. *Water Resources Research, 50*(12), 9586-9603.

Meyer, P. E. (2008). Information-theoretic variable selection and network inference from microarray data. *Ph. D. Thesis. Université Libre de Bruxelles*.

Miller, G. A. (1955). Note on the bias of information estimates. *Information theory in psychology: Problems and methods, 2*(95), 100.

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). The multi- institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research: Atmospheres, 109*(D7).

Moore, B. J., Mahoney, K. M., Sukovich, E. M., Cifelli, R., & Hamill, T. M. (2014). Climatology and Environmental Characteristics of Extreme Precipitation Events in the Southeastern United States. *Monthly Weather Review, 143*(3), 718-741. doi:10.1175/MWR-D-14-00065.1.

Moreda, F., Koren, V., Zhang, Z., Reed, S., & Smith, M. (2006). Parameterization of distributed hydrological models: learning from the experiences of lumped modeling. *Journal of Hydrology, 320*(1), 218-237.

Naabil, E., Lamptey, B., Arnault, J., Olufayo, A., & Kunstmann, H. (2017). Water resources management using the WRF-Hydro modelling system: Case-study of the Tono dam in West Africa. *Journal of Hydrology: Regional Studies, 12*, 196-209.

Najafi, M. R., Moradkhani, H., & Jung, I. W. (2011). Assessing the uncertainties of hydrologic model selection in climate change impact studies. *Hydrological Processes, 25*(18), 2814-2826. doi:10.1002/hyp.8043.

Nelson, J. G. (1966). Man and geomorphic process in the Chemung River Valley, New York and Pennsylvania. *Annals of the Association of American Geographers, 56*(1), 24-32. doi:10.1111/j.1467-8306.1966.tb00541.x.

Nemec, J., & Sittner, W.T. (1982). Application of the continuous API catchment model in the Indus River forecasting system in Pakistan. *Applied modeling in catchment hydrology, Water Resources Publications, Highlands Ranch,* p 313–322.

Nohara, D., Kitoh, A., Hosaka, M., & Oki, T. (2006). Impact of Climate Change on River Discharge Projected by Multimodel Ensemble. *Journal of Hydrometeorology, 7*(5), 1076-1089. doi:10.1175/JHM531.1.

Palmer, T. N., Alessandri, A., Andersen, U., & Cantelaube, P. (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society, 85*(6), 853.

Prat, O. P., & Nelson, B. R. (2015). Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). *Hydrology and Earth System Sciences, 19*(4), 2037-2056.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review, 133*(5), 1155-1174. doi:10.1175/mwr2906.1.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*(437), 179-191.

Randrianasolo, A., Ramos, M. H., Thirel, G., Andréassian, V., & Martin, E. (2010). Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models. *Atmospheric Science Letters, 11*(2), 100-107.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., & Participants, D. (2004). Overall distributed model intercomparison project results. *Journal of Hydrology, 298*(1), 27-60.

Regonda, S. K., Rajagopalan, B., Clark, M., & Zagona, E. (2006). A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin. *Water Resources Research, 42*(9), n/a-n/a. doi:10.1029/2005WR004653.

Salas, F. R., Somos- Valenzuela, M. A., Dugger, A., Maidment, D. R., Gochis, D. J., David, C. H., et al. (2018). Towards Real- Time Continental Scale Streamflow Simulation in Continuous and Discrete Space. *JAWRA Journal of the American Water Resources Association, 54*(1), 7-27.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology, 2*(2), 191-201.

Sanders, F. (1973). Skill in forecasting daily temperature and precipitation: Some experimental results. *Bulletin of the American Meteorological Society, 54*(11), 1171-1178.

Sedghi, H., & Jonckheere, E. (2014). On the conditional mutual information in the Gaussian–Markov structured grids. In *Information and Control in Networks* (pp. 277-297): Springer.

Shamseldin, A. Y., & O'Connor, K. M. (1999). A real-time combination method for the outputs of different rainfall-runoff models. *Hydrological Sciences Journal, 44*(6), 895-912.

Shamseldin, A. Y., O'Connor, K. M., & Liang, G. C. (1997). Methods for combining the outputs of different rainfall–runoff models. *Journal of Hydrology, 197*(1), 203-229. doi:https://doi.org/10.1016/S0022-1694(96)03259-3.

Sharma, S., Siddique, R., Balderas, N., Fuentes, J. D., Reed, S., Ahnert, P., et al. (2017). Eastern U.S. Verification of Ensemble Precipitation Forecasts. *Weather and Forecasting, 32*(1), 117-139. doi:10.1175/waf-d-16-0094.1.

Sharma, S., Siddique, R., Reed, S., Ahnert, P., Mendoza, P., & Mejia, A. (2018). Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system. *Hydrology and Earth System Sciences, 22*(3), 1831.

Siddique, R., & Mejia, A. (2017). Ensemble Streamflow Forecasting across the U.S. Mid-Atlantic Region with a Distributed Hydrological Model Forced by GEFS Reforecasts. *Journal of Hydrometeorology, 18*(7), 1905-1928. doi:10.1175/jhm-d-16-0243.1.

Siddique, R., Mejia, A., Brown, J., Reed, S., & Ahnert, P. (2015). Verification of precipitation forecasts from two numerical weather prediction models in the Middle Atlantic Region of the USA: A precursory analysis to hydrologic forecasting. *Journal of Hydrology, 529*, 1390-1406.

Silver, M., Karnieli, A., Ginat, H., Meiri, E., & Fredj, E. (2017). An innovative method for determining hydrological calibration parameters for the WRF-Hydro model in arid regions. *Environmental Modelling & Software, 91*, 47-69.

Sittner, W. T., Scauss, C. E., and Munro, J. C. (1969). Continuous hydrograph synthesis with an API-type hydrologic model. *Water Resources Reseasrch*, 5(5), 1007–1022.

Smith, M. B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., et al. (2012). The distributed model intercomparison project–Phase 2: Motivation and design of the Oklahoma experiments. *Journal of Hydrology, 418*, 3-16.

Steinschneider, S., Wi, S., & Brown, C. (2015). The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrological Processes, 29*(12), 2823-2839. doi:10.1002/hyp.10409.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., & Rogers, E. (1999). Using Ensembles for Short-Range Forecasting. *Monthly Weather Review, 127*(4), 433-446. doi:10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2.

Thirel, G., Regimbeau, F., Martin, E., Noilhan, J., & Habets, F. (2010). Short- and medium-range hydrological ensemble forecasts over France. *Atmospheric Science Letters, 11*(2), 72-77. doi:10.1002/asl.254.

Thirel, G., Rousset-Regimbeau, F., Martin, E., & Habets, F. (2008). On the impact of short-range meteorological forecasts for ensemble streamflow predictions. *Journal of hydrometeorology*, *9*(6), 1301-1317.

Thompson, P. D. (1977). How to Improve Accuracy by Combining Independent Forecasts. *Monthly Weather Review*, *105*(2), 228-229. doi:10.1175/1520-0493(1977)105<0228:HTIABC>2.0.CO;2.

Tolson, B. A., & Shoemaker, C.A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43, W01413, doi:10.1029/2005WR004723.

Toth, Z., & Kalnay, E. (1993). Ensemble Forecasting at NMC: The Generation of Perturbations. *Bulletin of the American Meteorological Society*, *74*(12), 2317-2330. doi:10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Velázquez, J. A., Anctil, F., Ramos, M. H., & Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. *Adv. Geosci., 29*, 33-42. doi:10.5194/adgeo-29-33-2011.

Vrugt, J. A., & Robinson, B. A. (2007). Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research, 43*(1), n/a-n/a. doi:10.1029/2005WR004838.

Wei, M., Toth, Z., Wobus, R., & Zhu, Y. (2008). Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A, 60*(1), 62-79.

Weigel, A. P., Liniger, M., & Appenzeller, C. (2008). Can multi- model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241-260.

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2009). Seasonal ensemble forecasts: are recalibrated single models better than multimodels? *Monthly Weather Review, 137*(4), 1460-1479.

Weisheimer, A., Doblas- Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Déqué, M., . . . Rogel, P. (2009). ENSEMBLES: A new multi- model ensemble for seasonal- to- annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical research letters, 36*(21).

Winkler, R. L., Murphy, A. H., & Katz, R. W. (1977). The consensus of subjective probability forecasts: Are two, three,... heads better than one? *In Preprints of the fifth conference on*

probability and statistics in atmospheric sciences (pp. 57–62). Boston: American Meteorological Society.

Xiong, L., Shamseldin, A. Y., & O'Connor, K. M. (2001). A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi–Sugeno fuzzy system. *Journal of Hydrology, 245*(1), 196-217. doi:https://doi.org/10.1016/S0022-1694(01)00349-3.

Yuan, X., & Wood, E. F. (2013). Multimodel seasonal forecasting of global drought onset. *Geophysical Research Letters, 40*(18), 4900-4905.

Yucel, I., Onen, A., Yilmaz, K., & Gochis, D. (2015). Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. *Journal of Hydrology, 523*, 49-66.