ADVANCING A REGIONAL HYDROLOGIC ENSEMBLE PREDICTION SYSTEM

A Dissertation in
Civil Engineering
by
Sanjib Sharma

© 2019 Sanjib Sharma

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2019
The dissertation of Sanjib Sharma was reviewed and approved* by the following:

Alfonso Mejia  
Associate Professor  
Department of Civil and Environmental Engineering  
Dissertation Advisor  
Chair of Committee

Peggy Johnson  
Professor of Civil Engineering

Xiaofeng Liu  
Assistant Professor  
Department of Civil and Environmental Engineering

Steven Greybush  
Assistant Professor  
Department of Meteorology

Patrick Fox  
Professor and Department Head  
Department of Civil and Environmental Engineering

*Signatures are on file in the Graduate School
ABSTRACT

There is great potential for using ensemble weather forecasts to improve hydrological predictions across spatial and temporal scales. To realize this potential, research is needed to formulate, assemble, and assess a full (i.e. accounting for the complete hydrometeorological forecasting chain) ensemble hydrological system and test how different system components can best contribute to improving hydrological predictions. The primary goal of this Ph.D. research is to fundamentally advance a reliable and robust Regional Hydrologic Ensemble Prediction System (RHEPS) by integrating new system components and implementing novel statistical techniques within a verifiable scientific and experimental setting. The proposed forecasting framework should facilitate understanding and quantifying the uncertainty and implications of ensemble weather forecasts on regional hydrological predictions.

To meet my research goal, the following four distinct research objectives are carried out:

Objective 1 (O1) - to perform a comprehensive verification analysis of ensemble precipitation forecasts from three different weather forecasting systems or guidance across the eastern U.S., including the National Oceanic and Atmospheric Administration’s (NOAA’s) National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System Reforecast version 2 (GEFSRv2), Short Range Ensemble Forecast (SREF) and the NCEP’s Weather Prediction Center probabilistic quantitative precipitation forecasts (WPC-PQPFs).

Objective 2 (O2) - to investigate the interactions between a weather preprocessor and a hydrologic postprocessor in ensemble streamflow forecasting. The terms pre-processing and post-processing indicate the implementation of advanced statistical models and tools that rely on reforecast or hindcast information in order to improve forecast skill and reliability prior to issuing the actual forecast.

Objective 3 (O3) - to determine whether the skill of hydrological multimodel forecasts is
significantly larger than that of a single model, and whether the observed skill improvement is
dominated by model diversity or reduction of noise associated with the ensemble size. **Objective**

4 (O4) - to evaluate the ability of NOAA’s National Centers for Environmental Prediction
(NCEP) Climate Forecast System version 2 (CFSv2) to result in skillful streamflow and water
quality predictions.

With O1, the weather forecasting system to use with O2 and O3 is selected. The verification
result for O1 demonstrate that WPC-PQPFs tend to be superior, in terms of the forecast skill and
reliability, to both the GEFSRv2 and SREF across the eastern U.S. However, GEFSRv2 is used
to complete O2 and O3 since these data are available for a longer time period and expand longer
lead times. With O2, two different components of the RHEPS, i.e., the weather preprocessor and
hydrological postprocessor are tested. This objective is significant because it will clarify the
conditions and the degree to which the combined implementation of preprocessing and
postprocessing can contribute to enhance hydrological predictions. With O2, it is concluded that
implementing both preprocessing and postprocessing ensures the most skill improvements, but
postprocessing alone can often be a competitive alternative. With O3, it is examined if the skill
improvement of hydrological multimodel forecasts is dominated by the reduction of noise
associated with ensemble size, or by model diversity (i.e., additional information provided by the
different models). The analysis indicates that any skill improvement of multimodel forecasts are
largely dominated by model diversity and that increasing the ensemble size has only a small
influence.

Finally, with O4, a new dynamical-statistical approach is built and implemented to generate
S2S water quantity (streamflow) and quality (nutrients and suspended sediments) predictions.
This hybrid approach is more cost effective and computationally efficient than implementing a
process-based water quality model, which makes it readily implementable in an operational forecasting setting. With O4, it is concluded that the dynamical CFSv2 forecasts, when combined with quantile regression, can generate skillful streamflow, nutrient load, and suspended sediment load forecasts at lead times of 1 to 3 months. Overall, the findings from this research demonstrate several strategies for enhancing hydrological forecasting and, ultimately, providing information that could be used to issue better forecasting products to the public.
# TABLE OF CONTENTS

List of Figures ......................................................................................................................... viii

List of Tables ........................................................................................................................... xii

Acknowledgements.................................................................................................................. xiii

Chapter 1 Introduction ........................................................................................................ 1

1.1 Literature review ........................................................................................................ 1
1.2 Research objectives and hypothesis ........................................................................... 7
1.3 Organization of chapters ............................................................................................ 9
1.4 Chapter 1 references................................................................................................... 10

Chapter 2 Spatial verification of meteorological forecasts ................................................ 16

2.1 Background and literature review .............................................................................. 16
2.2 Data and methods ....................................................................................................... 20
  2.2.1 Study area ........................................................................................................ 20
  2.2.2 Forecast products ............................................................................................. 23
2.3 Verification of short-range forecasts .......................................................................... 28
2.4 Verification of short-to medium-range GEFSRv2 ..................................................... 42
2.5 Discussion .................................................................................................................. 49
2.6 Summary and conclusions .......................................................................................... 51
2.7 Chapter 2 references................................................................................................... 53

Chapter 3 Regional hydrological ensemble prediction system ........................................ 63

3.1 Background and literature review .............................................................................. 63
3.2 Study area ................................................................................................................... 67
3.3 Approach .................................................................................................................... 71
  3.3.1 Hydrometeorological observations .................................................................. 71
  3.3.2 Meteorological forecasts ................................................................................. 72
  3.3.3 Distributed hydrological model ....................................................................... 72
  3.3.4 Statistical weather preprocessor ...................................................................... 74
  3.3.5 Statistical streamflow postprocessor ............................................................... 77
3.4 Forecast experiments and verification........................................................................ 81
3.5 Results and discussion............................................................................................... 83
  3.5.1 Performance of distributed hydrological model.............................................. 83
  3.5.2 Verification of the raw and preprocessed ensemble precipitation forecasts.... 85
  3.5.3 Selection of streamflow postprocessor ............................................................ 88
  3.5.4 Verification of the ensemble streamflow forecasts ......................................... 91
3.6 Summary and conclusions .......................................................................................... 100
3.7 Chapter 3 references................................................................................................... 102

Chapter 4 Hydrological multimodel streamflow forecasting ........................................ 113

4.1 Background and literature review .............................................................................. 113
4.2 Methodology .............................................................................................................. 117
  4.2.1 Statistical multimodel postprocessor ............................................................... 117
LIST OF FIGURES

Figure 2-1. 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. .......................... 22

Figure 2-2. 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 (years 2012-2013). 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 spatial aggregation scales. .................. 29

Figure 2-3. RME of the mean ensemble forecast versus the forecast lead time for the eastern regions (years 2012-2013). 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 spatial aggregation scales. ........................................ 31

Figure 2-4. BSS versus the forecast lead time for the eastern regions (years 2012-2013). 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 spatial aggregation scale of 500 x 500 km2, and both light to moderate precipitation events (Pr=0.5) as well as moderate to heavy precipitation events (Pr=0.9). ........................................... 34

Figure 2-5. Mean CRPSS versus the forecast lead time for the eastern regions (years 2012-2013). 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 spatial aggregation scales. ............................................... 36

Figure 2-6. Reliability diagrams for precipitation forecasts from the (a-c) GEFSRv2, (d-f) SREF, and (g-i) WPC-PQPF for forecast lead times of 1 (19-24h), 2 (43-48 h) and 3 (67-72 h) days for the eastern regions (years 2012-2013). The reliability diagrams are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and 500 x 500 km2 spatial aggregation scales. The insets show the sample size in logarithmic scale of the different forecast probability bins. ................................................................. 39

Figure 2-7. 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 (years 2012-2013). The symbols (black triangles for SREF, black diamonds for WPC-PQPF, and gray circles for GEFSRv2) 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 km2 spatial aggregation scales. The diagonal line represents the ROC curve associated with sampled climatology. .......................... 41

Figure 2-8. (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 (years 2004-2013, except the SER which is for years 2006-2013). The plots are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and spatial aggregation scales of (a) 100 x 100, (b) 300 x 300, and (c) 500 x 500 km2. .......................... 43
Figure 2-9. Monthly BSS for the GEFSRv2 precipitation forecasts versus the calendar month for the eastern regions (years 2004-2013, except the SER which is for years 2006-2013). 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 km² spatial aggregation scales.

Figure 2-10. Mean CRPSS for the GEFSRv2 precipitation forecasts versus the forecast lead time for the eastern regions (years 2004-2013, except the SER which is for years 2006-2013). The plots are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and spatial aggregation scales of (a) 100 x 100, (b) 300 x 300, and (c) 500 x 500 km².

Figure 2-11. Reliability diagrams for precipitation forecasts from the GEFSRv2 for spatial aggregation scales of (a-c) 100 x 100, (d-f) 300 x 300, and (g-i) 500 x 500 km² for the eastern regions (years 2004-2013, except the SER which is for years 2006-2013). 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.

Figure 3-1. Map illustrating the location of the four selected river basins in the U.S. middle Atlantic region.

Figure 3-2. Performance statistics for the uncalibrated and calibrated simulation runs for the entire period of analysis (years 2004-2012): (a) Rm, (b) NSE, and (c) PB.

Figure 3-3. 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 3-4. 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 3-5. Continuous ranked probability skill score (CRPSS) 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 3-6. Decomposition of the CRPS into CRPS potential (CRPSpot) and CRPS reliability (CRPSrel) 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 3-7. 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).

Figure 4-1. 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.

Figure 4-2. 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 study area in the US.

Figure 4-3. Performance of the hydrological models in simulation mode over the entire period of analysis (2004-2009, May-October): a) Pearson’s correlation coefficient, R, b) Nash-Sutcliffe efficiency, NSE, and c) percent bias, PB, between the daily simulated and observed flows.

Figure 4-4. 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.

Figure 4-5. 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.

Figure 4-6. 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 raw 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 raw members from each model). The standard deviation of the CMI values varies from 0.02 to 0.06.

Figure 5-1. CRPSS (relative to sampled climatology) of the CFSv2 raw and preprocessed ensemble (a)-(b) near-surface temperature and (c)-(d) precipitation forecasts versus the biweekly and monthly forecast lead times. The 90% confidence intervals, determined using the block bootstrap technique, are shown with shaded colors.
Figure 5-2. CRPSS (relative to sampled climatology) of the raw and postprocessed ensemble streamflow forecasts versus the biweekly forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. The 90% confidence intervals, determined using the block bootstrap technique, are shown with shaded colors. .................................................................188

Figure 5-3. Correlation coefficient and CRPSS (relative to sampled climatology) of the (a)-(d) total nitrogen, (b)-(e) total phosphorous, and (c)-(f) total suspended sediment load forecasts versus the monthly forecast lead time for all the selected locations. .................................190

Figure 5-4. CRPSS (relative to sampled climatology) of the streamflow, total nitrogen load, total phosphorous load, and total suspended sediment load forecasts at the month 1 lead time versus the basin area for the (a) entire year, (b) cool season, and (c) warm season. .......192

Figure 5-5. BSS (relative to sampled climatology) of the raw ensemble streamflow forecasts versus the biweekly forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. Results are shown for both low flows (flows with nonexceedance probability, Pr, less than 0.05) and high flows (flows with exceedance probability less than 0.05 or Pr>0.95). ........................................194

Figure 5-6. BSS (relative to sampled climatology) of the total nitrogen (TN), total phosphorous (TP) and total suspended sediment (TSS) load forecasts versus the monthly forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. Results are shown for both low loads (loads with nonexceedance probability, Pr, less than 0.05) and high loads (loads with exceedance probability less than 0.05 or Pr>0.95). ........................................196

Figure 5-7. CRPSS (relative to both sampled climatological flow observations and simulations) of the raw ensemble streamflow forecasts versus the biweekly forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. The 90% confidence intervals, determined using the block bootstrap technique, are shown with shaded colors. .................................................................198

Figure 5-8. CRPSS (relative to both sampled climatological load observations and simulations) of total nitrogen (TN), total phosphorous (TP) and total suspended sediment (TSS) load forecasts versus forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. ........................................199
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2-1</td>
<td>Summary and main characteristics of the datasets used in this study.</td>
<td>24</td>
</tr>
<tr>
<td>Table 3-1</td>
<td>Main characteristics of the four study basins.</td>
<td>70</td>
</tr>
<tr>
<td>Table 3-2</td>
<td>Summary and description of the verification scenarios.</td>
<td>82</td>
</tr>
<tr>
<td>Table 4-1</td>
<td>Characteristics of the selected gage locations.</td>
<td>131</td>
</tr>
<tr>
<td>Table 4-2</td>
<td>Summary and main characteristics of the datasets used in this study.</td>
<td>133</td>
</tr>
<tr>
<td>Table 5-1</td>
<td>Main characteristics of the study basins.</td>
<td>180</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Alfonso Mejia for his valuable guidance, scholarly inputs and consistent encouragement throughout my PhD endeavor. His valuable advices and constant feedback helped me throughout the research and writing of this dissertation. Dr. Mejia has always made himself available to clarify my doubts despite his busy schedule and I consider it as a great opportunity to do my PhD under his guidance and to learn from his research expertise.

I would like to thank the rest of my dissertation committee: Dr. Peggy Johnson, Dr. Xiaofeng Liu, and Dr. Steven Greybush for their insightful comments and encouragement. My sincere thanks also goes to Dr. Seann Reed of the National Weather Service (NWS) for providing insight about operational forecasting needs and problems. I would also like to thank my colleagues Dr. Ridwan Siddique and Michael Gomez for their collaboration in various projects related to this dissertation.

I am deeply thankful to my parents, my wife and my sister, who encouraged and helped me at every stage of my personal and academic life. This work would not have been possible without their love and encouragement.
CHAPTER 1
Introduction

A central hypothesis in this dissertation is that the quality of hydrologic prediction depends critically not only on the complexity of hydrological modelling, but on a range of key aspects of the forecast process chain, including meteorological forecast datasets, novel observations, hydrologic models, statistical methods and other tools. To test this hypothesis, this dissertation aims at developing and advancing a robust, regional hydrologic ensemble prediction system (RHEPS) through the integration of different system components, including weather/climate ensemble forecast, statistical weather preprocessor, hydrological/land-surface models, statistical streamflow postprocessor, hydrometeorological observations, and a verification strategy. 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 scale of implementation of the RHEPS is from small (~$10^2$ km$^2$) to regional basin scales (~$10^4$ km$^2$), and from short-range (~1-3 days) to seasonal (~3 months) temporal scales.

1.1 Literature review

Hydrologic predictions are inherently uncertain, therefore the need arises for effective uncertainty quantification approaches (Gupta et al. 2005; Brown and Seo 2010; Mejia and Reed 2011; Siddique et al. 2015). A number of uncertainty approaches have been implemented in hydrologic modeling. These approaches, however, have been overwhelmingly developed to function in simulation mode, as opposed to forecasting mode, leaving aside most of the actual
predictive uncertainty which is of hydrometeorological origin. There is a strong need for uncertainty approaches that take into consideration the full hydrometeorological prediction chain, i.e., approaches that can track the propagation of uncertainty starting from the weather predictions and all the way to the generation of the actual streamflow forecasts. Additionally, doing this (i.e. accounting for the full hydrometeorological prediction chain) may actually help bring into a broader and more insightful context assessments and opinions about what should the future hydrological research and modeling needs be.

Hydrometeorological uncertainty can originate from different sources, including shortcomings in forcing data, model structure and parameters, as well as model initial conditions (Bourgin et al. 2014; Brown et al. 2014; Roulin and Vannitsem 2015). In both meteorological and hydrological research, ensemble prediction systems (EPS) have been widely recognized to improve forecast skill over deterministic systems by providing probabilistic forecasts and information necessary to quantify forecast uncertainty (Schaake et al. 2006, 2007; Cloke and Pappenberger 2009). EPS typically consist of meteorological ensembles used to force a hydrological model (Wood and Lettenmaier 2008), where the former are generated through multiple runs of the numerical weather prediction (NWP) model(s) employing different initial conditions, physical parameterizations, and/or model physics (Schaake et al. 2007; Cloke and Pappenberger 2009; Dance and Zou 2010). Despite progress made in ensemble weather forecasting, ensembles are normally biased and underdispersive (Vrugt et al. 2008), particularly for surface weather quantities such as precipitation, and cannot explicitly distinguish different sources of uncertainty (Brown et al. 2014).

It is critical to understand and characterize the quality of ensemble precipitation forecasts. This is because precipitation forecasts are i) a critical input to hydrological models (Larson and
Peck 1974; McMillan et al. 2011; Sorooshian et al. 2011), and ii) often the dominant source of uncertainty in streamflow predictions (Roulin 2006; Cloke and Pappenberger 2009; Schaake et al., 2007; Yu et al. 2016). The majority of precipitation verification studies have been focused on the evaluation of precipitation ensembles from a particular forecasting system and/or a specific region (Baxter et al. 2014; Brown et al. 2012; Charles and Colle 2009; Hamill et al. 2013; Novak et al. 2014; Siddique et al. 2015; Stensrud and Yussouf 2007; Sukovich et al. 2014; Yuan et al. 2005). Hamill et al. (2013) and Baxter et al. (2014) recently verified precipitation forecasts from a single forecasting system, i.e. GEFSRv2. Likewise, Brown et al. (2012) verified the Short Range Ensemble Forecast (SREF) for selected regions across the U.S. Novak et al. (2014) and Sukovich et al. (2014) focused their verification on deterministic Weather Prediction Center quantitative precipitation forecasts (WPC-QPFs). Verification studies, however, that compare the quality of precipitation forecasts from different forecasting systems are quite limited (Siddique et al. 2015). Recently, Siddique et al. (2015) compared the relative quality of the GEFSRv2 and SREF but only within the domain of middle Atlantic region (MAR). None of the verification studies to date have verified and compared the quality of precipitation ensembles within multiple geographic regions while accounting for different spatial scales. Also, previous verification studies have tended to emphasize the skill and reliability of forecasts, thus ignoring other attributes of forecast performance (e.g., sharpness and discrimination) (Brown et al. 2012; Novak et al. 2014; Baxter et al. 2014). This dissertation overcome the outlined limitations of past verification studies.

Statistical preprocessors are often applied to remove systematic forecast biases and improve the forecast spread in weather ensembles (Sloughter et al., 2007, Bremnes, 2004, Wilks, 2009, Messner et al. 2014). Biases in the streamflow forecasts are handled through statistical
postprocessors (Zhao et al. 2011, Regonda et al. 2013, Reggiani et al. 2009, Brown and Seo 2010). Note that preprocessing and postprocessing is used to indicate the implementation of advanced statistical models and tools that rely on reforecast or hindcast information to improve forecast skill and reliability prior to issuing the actual forecasts. However, only a few studies have been carried out to investigate the interaction between preprocessing and postprocessing when addressing the total predictive uncertainty and improving the overall quality of ensemble streamflow forecasts. The few studies that have been completed indicate that postprocessing is more efficient than preprocessing alone (Kang et al. 2010; Roulin and Vannitsem, 2015). This is because improvements to the forcing data tend to lose their effect when propagated through the hydrological model, i.e. they do not tend to produce an equivalent improvement in the streamflow forecasts (Zalachori et al. 2012; Verkade et al. 2013). Although weather preprocessing can be complex and inefficient at improving streamflow forecasts, different streamflow postprocessors assume unbiased forcing (Zhao et al. 2011) and hydrological models can be sensitive to forcing biases. In some situations, it has been shown that the proper combination of preprocessing and postprocessing can result in improved skill, e.g., in seasonal forecasting (Yuan and Wood 2012). This raises the following key question: Under what circumstances does it make sense to combine weather preprocessing with streamflow postprocessing? Answering this question is crucial for improving streamflow forecasting in the most efficient way possible, considering that the time it takes to generate the forecasts is of outmost concern.

Model structural uncertainties are inevitable in hydrologic forecasting (Vrugt & Robinson, 2007). One approach in reducing model error is by optimally combining multiple hydrologic models to generate improved hydrologic predictions (Ajami et al., 2007). Most of the established
operational systems across the globe for short- to medium-range weather forecasting are multimodel, multiphysics ensemble systems (Hamill et al., 2013). In contrast, hydrological multimodel ensemble prediction systems have not been widely implemented and remain an underexplored area of research. Multimodel forecasting consists of using the outputs from several models to make and improve predictions about future events (Krishnamurti, 2003). Although hydrological multimodel approaches have been investigated before (Ajami 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 about near-real-time hydrological forecasting conditions. However, at medium-range timescales (≥ 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). 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 (i.e., additional information provided by the different models) or are due the addition of new ensemble members (i.e., increasing ensemble size). Identifying the sources of skill improvements is critical to improve the operational streamflow forecasting capabilities because generating many ensemble members in real time is often not feasible or realistic, and may not be as effective if skill enhancements are dominated by model diversity.

In the context of subseasonal to seasonal (S2S) timescales, there has remained a major weather-climate prediction gap, since it encompasses the time frame where most of the information from atmospheric initial conditions is lost and the timeframe is too short to be strongly influenced by climate modes of variability (National Academics of Sciences,
Engineering and Medicine 2016). However, recent improvements (e.g., improved observational datasets, spatial resolution, model physics, initial conditions, and assimilation techniques) in coupled atmosphere-ocean general circulation models (GCMs) are providing new opportunities to potentially enhance hydrological forecasting at the S2S timescale (Mendoza et al 2017, Yuan and Wood 2012). The NOAA’s National Centers for Environmental Prediction (NCEP) Climate Forecast System version 2 (CFSv2) is a recently developed, state-of-the science GCM (Saha et al 2014) designed to enhance global S2S climate predictions. Several recent studies have also shown that precipitation is a dominant driver of monthly nutrient loads at the basin scale (Sinha and Michalak 2016, Bastola and Mishra 2015). However, the ability of CFSv2 forecast to support water quality prediction has not been examined yet. Proven ability of CFSv2 to generate water quality predictions in a cost and time efficient manner will help to enhance current operational forecasting systems with new prediction capabilities for water quality, which are currently lacking in the US (Salas et al 2018).

In summary, the following key areas are identified where research is critically needed to continue to advance hydrological predictions: i) implementation of more comprehensive verification studies for precipitation ensembles (Brown et al., 2012; Welles et al., 2007), iii) understanding the interactions between preprocessing and postprocessing under a wider range of conditions (Kang et al. 2010; Roulin and Vannitsem 2015; Yuan and Wood, 2012), and iv) exploring the sources of skill enhancements of hydrological multimodel forecast (Delsole et al., 2014), and iv) investigating the ability of climate predictions to result in both skillful water quantity and quality forecasts (Sinha and Michalak, 2016; Bastola &Mishra, 2015).
1.2 Research objectives and hypothesis

The proposed dissertation will accomplish the following distinct research objectives:

- **Objective 1 (O1):** Perform a comprehensive verification analysis of ensemble precipitation forecasts from three different forecast systems or guidance across the eastern U.S., including the GEFSRv2, SREF and the NCEP’s WPC-PQPFs.

- **Objective 2 (O2):** Investigate the interaction between a weather preprocessor and hydrologic postprocessor in ensemble streamflow forecasting.

- **Objective 3 (O3):** Understand the ability of model diversity and increasing ensemble size for improving the skill of streamflow forecasts at short- to medium-range timescales.

- **Objective 4 (O4):** Evaluate the ability of climate forecasts to support skillful prediction of water quantity (streamflow) and quality (nutrients and sediments) at S2S timescales.

The goal with O1 is to choose as wisely as possible the weather forecasting system to use for the next objectives of this research, which makes O1 of utmost importance. To guide the execution of O1, I will seek to address 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 regions in the eastern U.S., e.g., within River Forecast Centers (RFCs)? Does the spatial aggregation scale influence the quality of precipitation forecasts? Are the RFCs in the eastern U.S. likely to benefit from statistical preprocessing techniques?

The goal with O2 is to decide if both preprocessing and postprocessing are required for the RHEPS, or if only postprocessing may suffice. This decision is critical as it can contribute to making the forecasting process more efficient, which is relevant to making near real-time
information available as required by operational forecasting. Some of the questions motivating O2 are: What are the separate impacts of statistical preprocessing and postprocessing on hydrologic ensemble forecasts? Will the statistical preprocessor interact with the streamflow postprocessor to improve the quality and skill of ensemble hydrological forecasts over the forecast lead times?

The goal with O3 is to decide if the skill of hydrological multimodel forecast is significantly larger than that of single model forecast, and to decide if the observed improvements are dominated by model diversity or by increasing ensemble size. Some of the questions motivating O3 are: 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 the addition of new ensemble members (i.e., increasing ensemble size)? Answering the latter is relevant to operational forecasting because generating many ensemble members in real time is often not feasible or realistic, and may not be as effective if skill enhancements are dominated by model diversity.

The goal with O4 is to develop a novel dynamical-statistical framework to investigate the ability of S2S climate predictions to generate skillful water quantity and quality forecasts. The outcomes from this alternative approach could serve to guide and support the development of future operational water quality forecasting capabilities by identifying the most relevant timescales, driving factors, and relative sources of uncertainty. To guide the execution of O4, I will seek to address the following questions: How skillful are the CFSv2 streamflow forecasts? Can the CFSv2 streamflow forecasts be used to anticipate nutrients and sediment loads? What is the relative importance of different sources of uncertainty (climate and hydrological) in streamflow, and nutrient and sediment load forecasts? How does the skill of the water quantity
and quality forecasts varies with different forecasting conditions (e.g., lead time, season, and flow threshold) and watershed characteristics (e.g., basin size and land cover)?

The hypotheses associated with each of the proposed research objectives are as follows:

- **Hypothesis 1 (H1):** Human-generated forecast guidance (WPC-PQPF) perform better than global (GEFSRv2) and regional (SREF) ensemble prediction systems at the short-range lead times (1-3 days).
- **Hypothesis 2 (H2):** Any improvements (e.g. gains in forecast skill) in the precipitation forecasts propagate through the hydrological model and translate into improved streamflow forecasts.
- **Hypothesis 3 (H3):** Skill improvements through multimodel forecasts are dominated by model diversity or increased ensemble size.
- **Hypothesis 4 (H4):** Climate forecasts (CFSv2) can be used to predict water quantity and quality at S2S timescales.

### 1.3 Organization of chapters

The remaining chapters of this dissertation are organized and structured based on the research objectives as follows. Chapters 2, 3, 4 and 5 contain the methods, results and conclusions associated with research objectives 1, 2, 3 and 4 respectively. Chapter 6 presents key summary findings, concluding remarks, and recommendations for the future. Note that chapters 2, 3, 4 and 5 are structured to be relatively independent of each other. Each chapter is organized to have its own literature review, methods, study area, results, discussion, conclusions and reference subsection.
1.4 Chapter 1 references


Bastola S and Misra V 2015 Seasonal hydrological and nutrient loading forecasts for watersheds over the Southeastern United States Environ. Model. Softw. 73 90–102


Mendoza P A et al 2017 An intercomparison of approaches for improving operational seasonal streamflow forecasts Hydrology and Earth System Sciences 21 3915-3935


Salas F R et al 2018 Towards Real-Time Continental Scale Streamflow Simulation in Continuous and Discrete Space JAWRA Journal of the American Water Resources Association 54 7-27


Sinha E and Michalak A M 2016 Precipitation dominates interannual variability of riverine nitrogen loading across the continental United States Environmental science & technology 50 12874-12884


Chapter 2

Spatial verification of meteorological forecasts

2.1 Background and literature review

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). Precipitation forecasts are used to issue severe weather warnings (Messner et al. 2014); forecast floods and other hydrological 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). 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). However, as ensemble forecasting systems evolve, the need arises to monitor and verify the quality of the evolving forecasting systems (Brown and Seo 2010). Forecast verification serves as a systematic framework for assessing and comparing the quality of forecasting systems (Casati et al. 2008; Mason and Weigel 2009; Murphy and Winkler 1987; Rossa et al. 2008; Welles et al. 2007). It provides information meaningful to administrators, scientists, and forecast users (Jolliffe and Stephenson 2012) such as metrics and plots needed to understand forecasting errors and biases (Davis et al. 2006; Ebert et al. 2013; Murphy and Winkler 1987).

A number of verification studies have been performed based on the forecasting systems or guidance used in this study (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), which consist of the National Oceanic and Atmospheric Administration’s (NOAA’s) National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System Reforecast version 2 (GEFSRv2) (Hamill et al. 2013), Short Range Ensemble Forecast (SREF) (Du et al. 2014; 2015), and NCEP’s Weather Prediction Center probabilistic quantitative precipitation forecasts (WPC-PQPFs) (WPC 2016). Previously, 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. Specifically, the GEFSRv2 precipitation
forecasts showed improvements of 1 day additional lead time at the early lead times and nearly 2
days at the longer lead times. 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 relative to the sampled climatology 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. For example, they noted that the SREF exhibits larger conditional bias and weaker
correlation within the geographic domain of the Middle Atlantic River Forecast Center (MARFC) than in other River Forecast Centers (RFCs). Previous verification studies tend to be
focused on a single forecasting system, do not tend to account for the effect of spatial scale, and
do not tend to distinguish forecasts based on the geographic domain of the different RFCs. This
study seeks to overcome these limitations of previous verification studies.

More recently, Siddique et al. (2015) compared precipitation forecasts from the GEFSRv2
and SREF against multisensor precipitation estimates (MPEs) over the MARFC domain. They
found that generally the two systems show similar skill and reliability over the MARFC but
some differences in performance were also noted. For example, in discriminating between the
occurrence and non-occurrence of a given precipitation amount, they noted that the GEFSRv2
reveals better discriminating ability than the SREF. 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. For instance, they showed that the skill of extreme WPC-QPFs is higher in the cool season than warm one and that extreme WPC-QPFs are more skillful over the geographic domain of the Northeast River Forecast Center (NERFC) than in the Southeast River Forecast Center (SERFC). Although the verification study of Sukovich et al. (2014) provided insight into WPC-QPFs, it did not consider probabilistic information (i.e. WPC-PQPFs).

Our primary objective with this study is to verify and compare the ensemble precipitation forecasts from the GEFSv2, SREF, and WPC-PQPFs, within the geographic domain of each of the four eastern RFCs in the U.S. The verification of WPC-PQPFs is one of the unique aspects of this study. The four eastern RFCs are the MARFC, NERFC, Ohio River Forecast Center (OHRFC), and 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.

To verify the 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) which accounts for spatial scale by verifying areal-averaged precipitation across different spatial aggregation scales. This areal-averaged approach is meaningful in this
case because, partly, the motivation for performing this verification is to inform future hydrological forecasting, research strategies. The areal-averaged approach can be viewed as representative of the aggregative hydrological response of a river basin to the precipitation forcing. This is another unique aspect of this study since previous studies do not tend to account for the effect of spatial scale on forecast verification. Further, we extend here the verification analysis of the GEFSRv2 and SREF performed by Siddique et al. (2015) for the MARFC domain to include WPC-PQPFs and all the RFCs in the eastern U.S. With this, 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?

2.2 Data and methods

2.2.1 Study area

The verification analysis is performed separately in each of the four RFCs considered, including the MARFC, NERFC, OHRFC, and SERFC, hereafter referred to as the middle Atlantic region (MAR), northeast region (NER), Ohio region (OHR), and southeast region (SER), respectively. Figure 1 illustrates the RFCs. These regions contain a massive and complex network of build infrastructure, which makes severe weather and flooding hazards particularly relevant (Hayhoe et al. 2006; Neff et al. 2000; Polsky et al. 2000; Siddique et al. 2015). Additionally, they comprise several major U.S. cities and river basins, including the Delaware,
Ohio, Potomac, and Susquehanna river basins (O’Donnell et al. 2000; Voisin et al. 2011). They are relatively wet and have a variety of geographic features that make weather forecasting particularly challenging such as an extensive coastal line, the Appalachian Mountains, and Great Lakes, among others (Baxter et al. 2014; Colle et al. 2003; Nam and Baigorria 2015; Yilmaz et al. 2005). The combination of physical features that contribute to landscape and boundary complexity make the hydrometeorological behavior of these regions diverse.
Figure 2-1. 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.
2.2.2 Forecast products

For the verification of the precipitation ensembles, we select outputs from the GEFSRv2, SREF, and WPC-PQPFs. 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 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.

i) 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° Gaussian grid spacing) and T190L42 (~0.67° Gaussian grid spacing) resolutions for the first and second 8 days, respectively. The 11-member reforecasts are initiated only once daily at 00 Coordinated Universal Time (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.
Table 2-1. Summary and main characteristics of the datasets used in the study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Horizontal Resolution (km²)</th>
<th>Number of models</th>
<th>Number of forecast cycles (per day)</th>
<th>Number of ensemble members</th>
<th>Lead time (hours)</th>
<th>Period of analysis (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEFSRv2</td>
<td>~55 x 55 ( (0.5^\circ \times 0.5^\circ) )</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>1-192</td>
<td>2004-2013</td>
</tr>
<tr>
<td></td>
<td>~73 x 73 ( (0.67^\circ \times 0.67^\circ) )</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>193-384</td>
<td>2004-2013</td>
</tr>
<tr>
<td>SREF</td>
<td>~32 x 32</td>
<td>3/4</td>
<td>4</td>
<td>21</td>
<td>1-87</td>
<td>2012</td>
</tr>
<tr>
<td></td>
<td>~16 x 16</td>
<td>3/4</td>
<td>4</td>
<td>21</td>
<td>1-87</td>
<td>2012-2013</td>
</tr>
<tr>
<td>WPC-PQPF</td>
<td>~32 x 32</td>
<td>-</td>
<td>2</td>
<td>7</td>
<td>1-72</td>
<td>2012-2013</td>
</tr>
<tr>
<td>MPEs</td>
<td>~4 x 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2004-2013</td>
</tr>
</tbody>
</table>
ii) 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 UTC (Du et al. 2014; 2015). 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. 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.

iii) WPC-PQPFs

The WPC-PQPFs are derived, for lead times of 1 to 3 days and at 32 km horizontal resolution, by incorporating forecast uncertainty information into 6-hour deterministic WPC-QPFs (WPC 2016; Novak et al. 2014). 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.

iv) MPEs

We use MPEs as the observed precipitation data 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, gridded precipitation observations at 4 x 4 km² 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 MPEs over the period of 2004-2013 (Table 1).

v) Verification strategy

A wide variety of verification metrics are used in this study, including both deterministic and probabilistic measures. Specifically, the following 6 different verification metrics are considered: 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).

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, the verification analysis is conducted using the EVS developed by Brown et al. (2010). The EVS is a modular and flexible software tool
specifically designed to facilitate the diagnostic verification of ensemble forecasts of hydrometeorological (e.g., precipitation and temperature) and hydrological variables (e.g., streamflow and water levels) at discrete locations (Brown et al. 2010). It has been used in several verification studies (Brown et al. 2012; Siddique et al. 2015). We use the EVS to accomplish the following tasks: i) pair observations and ensemble forecast values, ii) compute different verification statistics, iii) perform temporal aggregation and data stratification, and iv) produce graphical and numerical outputs for the different verification statistics.

When verifying the selected forecast products, the forecasts and observed datasets are conditioned upon different variables (e.g., forecast lead time, seasonality, precipitation threshold, and spatial aggregation scale) to account for various relevant scenarios. The verification is conducted using 6 hourly precipitation accumulations and mostly for moderate to heavy precipitation amounts. For this, precipitation amounts greater than that implied by a nonexceedance probability, in the sampled climatological probability distribution, of 0.9 are selected. This corresponds to precipitation amounts greater than ~4 mm for the 6 hourly accumulations. To account for the effect of the spatial aggregation scale, areal-averaged precipitation is verified, as opposed to individual grid cells, across different area sizes. Three different area sizes are selected: small (100 x 100 km²), intermediate (300 x 300 km²), and large (500 x 500 km²). To account for sampling variability, three or more area extents of the same size (i.e. small, intermediate, or large) are selected within the boundary of the different RFCs. For a particular area size, each extent is treated as a separate verification unit. Thus, the verification metrics are computed using the area sizes (i.e. small, intermediate, and large), not the individual grid cells of the forecasting systems.
To perform the verification analysis, two main case studies are considered. In the first case study, 6 hourly precipitation accumulations from the GEFSRv2, SREF, and WPC-PQPFs are verified, over their common period of two years (2012-2013), for forecast lead times of 1 to 3 days. In the second case study, 6 hourly precipitation accumulations from the GEFSRv2 alone are verified, 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. Note that the GEFSRv2, SREF, and WPC-PQPF products account for different number of cycles and members (Table 1), which can have an important influence on forecast quality; analyzing this further is, nonetheless, beyond the scope of this study.

2.3 Verification of short-range forecasts

i) Correlation coefficient and RME

The correlation coefficient and RME are used as the deterministic metrics of forecast quality. Fig. 2 shows, for the three different spatial scales 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 spatial scale increases. For instance, regarding the latter, the values of the correlation coefficient for the GEFSRv2 tend to be larger in Fig. 2c (large spatial scale), across regions and forecast lead times, than in Fig. 2a (small spatial scale). This behavior is similar for the SREF and WPC-PQPF in Fig. 2.
Figure 2-2. 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 (years 2012-2013). 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 spatial aggregation scales.
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). The diurnal cycle is investigated 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 is 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 (Roebber et al. 1996).

To examine the bias associated with the mean ensemble forecast, Fig. 3 plots 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). All three forecasting systems or guidance exhibit underforecasting bias for moderate to heavy precipitation events, across forecast lead times and spatial scales (Fig. 3). The bias increases, in most cases, with the forecast lead time and decreases some as the spatial scale increases. Comparing the three forecasting systems or guidance against each other, the WPC-PQPF seems to have the least bias of the three, at least at the early lead times.
Figure 2-3. RME of the mean ensemble forecast versus the forecast lead time for the eastern regions (years 2012-2013). 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 spatial aggregation scales.
For example, the bias for the NER at a lead time of 12 hours and the largest spatial scale 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. This is assessed further in the next subsection by comparing the forecast skill in the cool and warm season.

ii) Skill scores: BSS and CRPSS

To investigate the probabilistic attributes of the selected forecasting systems and guidance, the BSS and CRPSS associated with the precipitation forecasts and observations pairs are examined. When computing the BSS and CRPSS, the sampled climatology is used as the reference system. In Fig. 4, the BSS is shown 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$^2$ spatial scale, and both light to moderate precipitation events (Pr=0.5 or 6 hourly accumulations greater than ~1 mm) as well as moderate to heavy precipitation events (Pr=0.9 or 6 hourly accumulations greater than ~4 mm). 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 across all the regions and it is
generally higher in the cool season than in the warm one for the moderate to heavy precipitation events. 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).
Figure 2-4. BSS versus the forecast lead time for the eastern regions (years 2012-2013). 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 spatial aggregation scale of 500 x 500 km², and both light to moderate precipitation events (Pr=0.5) as well as moderate to heavy precipitation events (Pr=0.9).
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). For example, during the cool season, the BSS for the MAR at a lead time of 24 hours and moderate to heavy precipitation events is ~0.2 for the SREF (Fig. 4f) and ~0.5 for the WPC-PQPF (Fig. 4j). 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, among other potential factors, to the influence of the jet stream and extratropical cyclones on precipitation in this most poleward of the study domains. Note that the lack of a diurnal cycle in the precipitation forecasts from the SREF and WPC-PQPF is likely due to the fact that these systems issue or release forecasts at least twice a day.
Figure 2-5. Mean CRPSS versus the forecast lead time for the eastern regions (years 2012-2013). 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 spatial aggregation scales.
In Fig. 5, the CRPSS (relative to sampled climatology) is plotted 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 spatial scales. In general, the CRPSS decreases with increasing forecast lead time but increases with increasing spatial scale, independently of the region. The CRPSS tends to increase from the GEFSRv2 to the SREF, except in the MAR, and from the SREF to the WPC-PQPF in all of the eastern regions. For instance, the skill for the SER at a lead time of 12 hours and the intermediate spatial scale (300 x 300 km$^2$) is ~0.4 for the GEFSRv2 (Fig. 5b), ~0.5 for the SREF (Fig. 5e), and ~0.6 for the WPC-PQPF (Fig. 5h). 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 spatial scale (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, 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).

iii) Reliability

The reliability diagrams for the GEFSRv2, SREF, and WPC-PQPF and the four eastern regions are shown in Fig. 6. The reliability diagrams are based on moderate to heavy precipitation events and the large spatial scale (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-PQPF is underconfident at high forecast probabilities, e.g., probabilities greater than 0.7 (Figs. 6g-i). Nevertheless, the WPC-PQPF seems 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.
Figure 2-6. 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 (years 2012-2013). The reliability diagrams are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and 500 x 500 km² spatial aggregation scales. The insets show the sample size in logarithmic scale of the different forecast probability bins.
iv) ROC curves

To examine the ability of the different forecasting systems and guidance to discriminate between occurrences versus non-occurrences of a precipitation event, Fig. 7 plots the ROC curve for each region. 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. To determine the ROC curves, 6 hourly accumulations, 500 x 500 km² spatial scale, and moderate to heavy precipitation events are considered.

The ROC curves for the MAR, NER, OHR, and SER are shown in Figs. 7a-d, respectively. Overall, the GEFSRv2 and WPC-PQPF show better discrimination ability than the SREF across regions, although these differences are mostly small (e.g., Fig. 7d) and likely insignificant. 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.
Figure 2-7. 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 (years 2012-2013). The symbols (black triangles for SREF, black diamonds for WPC-PQPF, and gray circles for GEFSRv2) 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 km² spatial aggregation scales. The diagonal line represents the ROC curve associated with sampled climatology.
2.4 Verification of short- to medium-range GEFSRv2 forecasts

i) Correlation and RME

Attention is now focused on short- to medium-range precipitation forecasts (days 1-16) from the GEFSRv2, where forecasts for the period of 2004-2013 are considered. In Figs. 8a-c, the correlation coefficient between the GEFSRv2 mean ensemble forecast and the corresponding observed precipitation values is shown as a function of the forecast lead time for small, intermediate, and large spatial scales, respectively. The correlation coefficient declines with increasing forecast lead time and increases slightly with the spatial scale. 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. The predictability, however, can be already quite low for shorter forecast lead times. For example, the correlation coefficient is around 0.2 at a lead time of 4 days in Figs. 8a-b.
Figure 2-8. (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 (years 2004-2013, except the SER which is for years 2006-2013). The plots are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events (Pr=0.9), and spatial aggregation scales of (a) 100 x 100, (b) 300 x 300, and (c) 500 x 500 km².
Figs. 8d-f plot the RME of the GEFSRv2 mean ensemble forecast against the forecast lead time for small, intermediate, and large spatial scales, 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 spatial scale. 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. 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). A similar conclusion is reached by contrasting the RME values in Figs. 8d-f against those in Figs. 3a-c.

ii) Skill scores: BSS and CRPSS

Figs. 9a-c show 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 km² spatial scale, 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.
Figure 2-9. Monthly BSS for the GEFSRv2 precipitation forecasts versus the calendar month for the eastern regions (years 2004-2013, except the SER which is for years 2006-2013). 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 km² spatial aggregation scales.
Figure 2-10. Mean CRPSS for the GEFSRv2 precipitation forecasts versus the forecast lead time for the eastern regions (years 2004-2013, except the SER which is for years 2006-2013). The plots are based on 6 hourly precipitation accumulations, moderate to heavy precipitation events ($Pr=0.9$), and spatial aggregation scales of (a) 100 x 100, (b) 300 x 300, and (c) 500 x 500 km$^2$. 
In terms of the CRPSS, the skill decreases with increasing forecast lead time, as expected, and increases somewhat with increasing spatial scale across regions (Fig. 10). In Fig. 10, the CRPSS is computed 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 spatial scales (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.

iii) 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 and across spatial scales (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 spatial scales. 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. 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 additional statistical postprocessing to improve the raw ensemble forecasts from the GEFSRv2.
Figure 2-11. Reliability diagrams for precipitation forecasts from the GEFSRv2 for spatial aggregation scales of (a-c) 100 x 100, (d-f) 300 x 300, and (g-i) 500 x 500 km$^2$ for the eastern regions (years 2004-2013, except the SER which is for years 2006-2013). 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.
2.5 Discussion

This study extends the verification analysis of the GEFSRv2 and SREF recently performed by Siddique et al. (2015) for the MAR to include WPC-PQPF guidance and three additional geographic regions, which together with the MAR encompass the entire eastern U.S. The verification results obtained here for the GEFSRv2 and SREF are similar to those reported by Siddique et al. (2015). This is not surprising since the experimental setting is similar for both studies. Both studies indicate that, over the MAR, there is a tendency for the CRPSS to increase from the GEFSRv2 to the SREF, and the GEFSRv2 shows better discrimination ability than the SREF. Additionally, the studies demonstrate that both forecasting systems are more skillful in the cool season than the warm one for moderate to heavy precipitation events, and the skill increases with the spatial scale, which is particularly noticeable at the early lead times. The seasonal trend in the forecast skill of moderate to heavy precipitation events is visible in all the regions and forecasting systems or guidance considered. This trend has been identified in several other related studies (Baxter et al. 2014; Brown et al. 2012; Fritsch and Carbone 2004; Ralph et al. 2005; Sukovich et al. 2014).

In particular, Baxter et al. (2014) verified precipitation forecasts from the GEFSRv2 over the southeastern U.S. Their results compare well with the results in this study, even though they used different metrics, observed data, and period of analysis. Over a longer period of analysis (1985-2013) than the one used here, they showed that GEFSRv2 precipitation forecasts for the SER are more skillful in the cool season than warm one, and tend to exhibit an underforecasting bias for moderate precipitation events. This is qualitatively similar to the results obtained here for the GEFSRv2 (e.g., Figs. 3a-c, and Figs. 4b and 4d). With no other published verification
studies for WPC-PQPFs that we know of, the verification of extreme WPC-QPFs conducted by Sukovich et al. (2014) can be used here for comparison purposes. Sukovich et al. (2014) found that the skill is higher in the NER and lower in the SER, compared to other eastern U.S. regions. This is consistent with our findings for WPC-PQPFs based on moderate to heavy precipitation events (Fig. 5). In regards to the SREF, Brown et al. (2012) verified precipitation forecasts for different regions in the U.S., including the MAR but not the other eastern U.S. regions. The results of Brown et al. (2012) for the MAR compare well with the results from this study. For example, Brown et al. (2012) obtained that the CRPSS over the MAR is approximately equal to 0.38 for moderate precipitation events at a lead time of 1 day, while under similar conditions (moderate to heavy precipitation events) the CRPSS is here equal to ~0.35. Indeed, the comparisons are similar for other metrics as well, e.g., the relative mean error and correlation coefficient.

Although a number of studies have examined and discussed the value of human-generated forecasts for different weather variables and forecasting conditions (Charba et al. 2003; Homer et al. 2006; Novak et al. 2014), the ability of forecasters to add quality to probabilistic forecasts remains an opened question and a topic of discussion. This verification study indicates that the human-influenced WPC-PQPF guidance is generally superior to the GEFSRv2 and SREF. It is interesting to observe here that the WPC-PQPFs show the least bias and the highest skill scores among the three forecasting sources verified. Moreover, this is the case across all the different RFCs considered. Further research is warranted to better understand this and isolate the role played by human influence on improving the forecasts and verification results. Besides improved forecasts, this could also provide valuable diagnostic information about potential modifications to the forecasting systems and models.
2.6 Summary and conclusions

In this study, we verified the quality of ensemble precipitation forecasts from the GEFSRv2, SREF, and WPC-PQPF since precipitation is a key forcing of interest in many weather-related applications. We selected these three forecasting systems or guidance because they are operational, have multiyear data available, and/or represent conditions of interest to forecasters. Our verification was based on selected metrics (see the Appendix) conditioned upon the precipitation threshold, forecast lead time, seasonality, and spatial scale. 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 main finding of this verification study can be summarized as follows:

- Precipitation forecast bias increases with the forecast lead time and decreases some as the spatial scale increases. However, all the forecasts exhibit a strong underforecasting bias. For short range forecast (days 1-3), the general tendency is for the SREF and WPC-PQPF to show less bias in the NER and more bias in the SER than in the other eastern regions. Overall, the WPC-PQPF shows the least bias of the three forecasting systems or guidance among all the eastern RFCs. The GEFSRv2 analysis for short to medium range forecasts (days 1-16) indicates that the bias does not vary greatly among the regions.

- Variations in the skill among the regions are visible in all three forecasting system or guidance. The skill of the forecasts is appreciably better in the cool season than in the warm one and is particularly larger for moderate to heavy precipitation events. Relative to the other regions considered, the MAR shows unusually low skill with the SREF and a noticeable skill
gain with the WPC-PQPF, relative to both the GEFSRv2 and SREF. For the regions considered, according to the short- to medium-range GEFSRv2 outputs, the precipitation forecasts tend to have some useful skill up to approximately day 7, beyond that the skill is similar to that from sampled climatology. Also in relation to the GEFSRv2, the month of July seems to generally have the lowest skill at forecast lead time of 1 to 16 days.

- Relative to the other forecasting systems or guidance, the GEFSRv2 precipitation forecasts show the most variability across forecast lead times and are marked by a diurnal cycle. Among the regions considered, the SER shows a stronger daily oscillation.

- The analysis based on reliability diagrams indicates that forecasts tend, for the most part, to be overconfident. For the MAR, the GEFSRv2 and SREF forecasts are slightly more unreliable than other RFCs at forecast lead times of 1-3 days.

- The GEFSRv2 and WPC-PQPFs show better ability to discriminate between occurrences versus non-occurrences of a moderate to heavy precipitation event than the SREF across eastern RFCs, but with the MAR, the GEFSRv2 is somewhat better at discriminating the precipitation events than the WPC-PQPF.

- The WPC-PQPFs tend to be superior in the correlation coefficient, relative mean error, reliability, and forecast skill scores relative to both the GEFSRv2 and SREF, but the performance varies with the region and forecast lead time.

- For the regions considered, according to the short- to medium-range GEFSRv2 outputs (days 1-16), the precipitation forecasts tend to have skill similar to that from sampled climatology after ~ day 7; 0- to 7-day forecast bias grows with the forecast lead time, and the forecast tend, for the most part, to be overconfident.
Moreover, by evaluating the influence of the forecasting systems or guidance used in this study on flood forecasts, information directly relevant to RFCs could be generated. For this, the raw and postprocessed precipitation forecasts from the GEFS, SREF, and WPC-PQPF could be used to force different hydrological models. This will ultimately produce verifiable streamflow forecasts and provide a context for evaluating if modifications to the weather forecasting systems and guidance can, and by how much, improve flood forecasts. As was the case here with the precipitation forecasts, the latter should be done for a wide range of conditions including different streamflow thresholds, seasons, and spatial scales. Also, conditions unique to hydrological predictions could be considered such as the effect of land cover and engineered infrastructure (e.g., reservoirs) on flood forecast quality.

2.7 Chapter 2 references


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.


Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, & D. R. Novak, 2014:


Chapter 3

Regional hydrological ensemble prediction system

3.1 Background and literature review

Both climate variability and climate 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 (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 streamflow 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 streamflow 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 streamflow 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, to meet specific regional needs, and/or for 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 streamflow 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 (Shi et al., 2008; Brown and Seo, 2010; Madadgar et al., 2014; Ye 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; Abaza et al., 2017). 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 (Renard et al., 2010).

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) (Zalachori et al., 2012; Roulin and Vannitsem, 2015; 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 (Yuan and Wood, 2012), 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 streamflow 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 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, flow 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)?

### 3.2 Study area

The North Branch Susquehanna River (NBSR) 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 km². The NBSR basin is selected as flooding is an important regional concern. This region has a relatively 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). The climate in the upper MAR, where the NBSR basin is located, can be classified as warm, humid summers and snowy, cold winters with frozen precipitation (Polsky et al, 2000). During the cool season, a positive North Atlantic Oscillation phase generally results in increased precipitation amounts and occurrence of heavy snow (Durkee et al., 2007). Thus, flooding in the cool season is dominated by heavy precipitation events accompanied by snowmelt runoff. In the summer season, convective thunderstorms with increased intensity may lead to greater variability in streamflow. In the NBSR basin, we select four different U.S. Geological Survey (USGS) daily gauge stations,
representing a system of nested subbasins, 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 km². Table 1 outlines some key characteristics of the study basins.
Figure 3-1. Map illustrating the location of the four selected river basins in the U.S. middle Atlantic region.
Table 3-1. Main characteristics of the four study basins.

<table>
<thead>
<tr>
<th>Location of outlet</th>
<th>Cincinnatus, New York</th>
<th>Chenango Forks, New York</th>
<th>Conklin, New York</th>
<th>Waverly, New York</th>
</tr>
</thead>
<tbody>
<tr>
<td>NWS id</td>
<td>CINN6</td>
<td>CNON6</td>
<td>CKLN6</td>
<td>WVYN6</td>
</tr>
<tr>
<td>USGS id</td>
<td>01510000</td>
<td>01512500</td>
<td>01503000</td>
<td>01515000</td>
</tr>
<tr>
<td>Area [km²]</td>
<td>381</td>
<td>3841</td>
<td>5781</td>
<td>12362</td>
</tr>
<tr>
<td>Latitude</td>
<td>42°32’28”</td>
<td>42°13’05”</td>
<td>42°02’07”</td>
<td>41°59’05”</td>
</tr>
<tr>
<td>Longitude</td>
<td>75°53’59”</td>
<td>75°50’54”</td>
<td>75°48’11”</td>
<td>76°30’04”</td>
</tr>
<tr>
<td>Minimum daily flow* [m³/s]</td>
<td>0.31 (0.11)</td>
<td>4.05 (2.49)</td>
<td>6.80 (5.32)</td>
<td>13.08 (6.71)</td>
</tr>
<tr>
<td>Maximum daily flow* [m³/s]</td>
<td>172.73 (273.54)</td>
<td>1248.77 (1401.68)</td>
<td>2041.64 (2174.734)</td>
<td>4417.42 (4417.42)</td>
</tr>
<tr>
<td>Mean daily flow* [m³/s]</td>
<td>8.89 (9.17)</td>
<td>82.36 (81.66)</td>
<td>122.93 (121.99)</td>
<td>277.35 (215.01)</td>
</tr>
<tr>
<td>Climatological flow (Pr=0.95)** [m³/s]</td>
<td>29.45</td>
<td>266.18</td>
<td>382.28</td>
<td>843.84</td>
</tr>
</tbody>
</table>

*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.
3.3 Approach

In this section, we describe the different components of the RHEPS, including the hydrometeorological observations, weather forecasts, preprocessor, postprocessors, hydrological model, and the forecasting experiments and verification strategy.

3.3.1 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 km² resolution were provided by the NOAA’s Middle Atlantic River Forecast Center (MARFC) (Siddique and Mejia 2017). 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 km² cells, which are produced by combining multiple radar estimates and rain gauge measurements. The gridded near-surface air temperature data at 4 x 4 km² resolution were developed by the MARFC 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.
3.3.2 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° Gaussian grid spacing or ~55 km) and T190L42 (~0.67° 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). The GEFSRv2 data are bilinearly interpolated onto the 4 x 4 km² grid cell resolution of the HL-RDHM model.

3.3.3 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. Each grid 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. Further, the SNOW-17 module allows each cell to 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 (Koren et al., 2004; Smith et al., 2012a). In this study, we run HL-RDHM using a 2-km horizontal resolution. 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; Rafieeinhasab 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:

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

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

where \( q_i \) and \( s_i \) denote the daily observed and simulated flows at time \( i \), respectively; \( \Omega \) 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).

### 3.3.4 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, such as Bayesian Model Averaging (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

\[ P(y \leq z \mid x) = \Lambda[\omega(z) - \delta(x)], \]  

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 \( \varphi \) to control the dispersion of the logistic predictive distribution,

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

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

\[ \delta(x) = a_0 + a_1 x, \quad \text{and} \]

\[ \eta(\varphi) = b_0 + b_1 \varphi, \]

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 \( f_i \); \( K \) is the total number of ensemble members; and \( \varphi \) 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 $\pi_i$ of the $i$th 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, $\pi_i$ can be expressed as (Messner et al., 2014a)

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

(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 Schaeke 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 Schaeke 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.

3.3.5 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. For this, the 6-hourly precipitation accumulations are used to force the HL-RDHM and generate 6-hourly flows. Note that we use 6-hourly accumulations since this is the resolution of the GEFSRv2 data after day 4 and this is a temporal resolution often used in operational forecasting in the U.S. Since the observed flow data are mean daily, we compute the mean daily flow forecast from the
6-hourly flows. The postprocessor is then applied to the mean daily values from day 1 to 7. ARX (1,1) and QR postprocessors are discussed below:

**i) 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:

\[
q_{i+1}^T = (1 - c_i) q_i^T + c_i f_{i+1}^T + \xi_{i+1},
\]

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 \( \xi \) is the residual error term. In Eq. (7), assuming that there is significant correlation between \( \xi_{i+1} \) and \( \xi_i \), \( \xi_{i+1} \) can be calculated as:

\[
\xi_{i+1} = \frac{\sigma_{\xi_{i+1}}}{\sigma_{\xi_i}} \rho(\xi_{i+1}, \xi_i) \xi_i + \vartheta_{i+1},
\]

where \( \sigma_{\xi_i} \) and \( \sigma_{\xi_{i+1}} \) are the standard deviation of \( \xi_i \) and \( \xi_{i+1} \), respectively; \( \rho(\xi_{i+1}, \xi_i) \) is the serial correlation between \( \xi_{i+1} \) and \( \xi_i \); and \( \vartheta_{i+1} \) is a random Gaussian error generated from \( \mathbb{N}(0, \sigma_{\vartheta_{i+1}}^2) \).

To estimate \( \sigma_{\vartheta_{i+1}}^2 \) in Eq. (8), the following equation is used:

\[
\sigma_{\vartheta_{i+1}}^2 = [1 - \rho^2(\xi_{i+1}, \xi_i)] \sigma_{\xi_i}^2.
\]
To implement Eq. (7), ten equally spaced values of $c_{i+1}$ are selected from 0.1 to 0.9. For each value of $c_{i+1}$, $\sigma^2_{\hat{\vartheta}_{i+1}}$ is determined from Eq. (9) using the training data to determine the other variables in Eq. (9). Then, $\hat{\vartheta}_{i+1}$ is generated from $\mathcal{N}(0, \sigma^2_{\hat{\vartheta}_{i+1}})$ and $\xi_{i+1}$ 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 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).

**ii) 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 space as well as 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

$$
\epsilon_t' = d_t + e_t \bar{f},
$$

(10)

where $\epsilon_t'$ is the error estimate at quantile interval $\tau$; $\bar{f}$ is the ensemble mean; and $d_t$ and $e_t$ are the linear regression coefficients at $\tau$. The coefficients are determined by minimizing the sum of the residuals based on the training data as follows:

$$
\min \sum_{i=1}^{N} w_{\tau}[\epsilon_{\tau,i} - \epsilon_{\tau,i}(i, \bar{f})],
$$

(11)

$\epsilon_{\tau,i}$ and $\bar{f}_i$ are the $i^{th}$ paired samples from a total of $N$ samples; $\epsilon_{\tau,i}$ is computed as the observed flow minus the forecasted one, $q_t - f_t$; and $w_{\tau}$ is the weighting function for the $\tau^{th}$ quantile defined as:

$$
w'_{\tau}(\zeta_i) = \begin{cases} (\tau - 1)\zeta_i & \text{if } \zeta_i \leq 0 \\ \tau \zeta_i & \text{if } \zeta_i > 0 \end{cases}.
$$

(12)

$\zeta_i$ is the residual term defined as the difference between $\epsilon_{\tau,i}$ and $\epsilon_{\tau,i}(i, \bar{f})$ for the quantile $\tau$. The minimization in Eq. (11) is solved using linear programming (Koenker, 2005).

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

$$
f_t = \bar{f} + \epsilon_t'.
$$

(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: 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). 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) scenario verify the raw and preprocessed ensemble precipitation forecasts, respectively. Scenarios 3 (S3), 4 (S4) and 5 (S5) verify the raw, preprocessed, and postprocessed ensemble streamflow forecasts, respectively. The last scenario, S6, verifies the combined preprocessed and postprocessed ensemble streamflow 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 streamflow forecast scenarios, S3-S6, are verified against mean daily streamflow observations from the USGS. The quality of the streamflow forecasts is evaluated conditionally upon forecast lead time, season (cool and warm), and flow threshold.
Table 3-2. Summary and description of the verification scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Verification of the raw ensemble precipitation forecasts from the GEFSRv2</td>
</tr>
<tr>
<td>S2</td>
<td>Verification of the preprocessed ensemble precipitation forecasts from the GEFSRv2: GEFSRv2+HCLR</td>
</tr>
<tr>
<td>S3</td>
<td>Verification of the raw ensemble flood forecasts: GEFSRv2+HL-RDHM</td>
</tr>
<tr>
<td>S4</td>
<td>Verification of the preprocessed ensemble flood forecasts: GEFSRv2+HCLR+HL-RDHM</td>
</tr>
<tr>
<td>S5</td>
<td>Verification of the postprocessed ensemble flood forecasts: GEFSRv2+HL-RDHM+QR</td>
</tr>
<tr>
<td>S6</td>
<td>Verification of the preprocessed and postprocessed ensemble flood forecasts: GEFSRv2+HCLR+HL-RDHM+QR</td>
</tr>
</tbody>
</table>
3.5 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 streamflow forecasting scenarios is shown in the fourth subsection (forecasting scenarios S3-S6).

3.5.1 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 observed precipitation and near surface temperature data. 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.
Figure 3-2. 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.
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), shows slightly higher performance 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).

3.5.2 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.
Figure 3-3. 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.
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 ~0.17 to 0.5 and from ~0.0 to 0.4 for the preprocessed and raw forecasts, respectively; while in the cool season (Figs. 3e-h) the CRPSS values vary overall in the range from ~0.2 to 0.6 and from ~0.1 to 0.6 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 (Figs. 3e-h).

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 streamflow ensembles from forcing the HL-RDHM with the preprocessed precipitation, as opposed to using the raw forecast forcing. 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.

### 3.5.3 Selection of the streamflow postprocessor

The ability of the ARX(1,1) and QR postprocessors to improve ensemble streamflow forecasts is investigated here. The postprocessors are applied to the raw streamflow ensembles at each forecast lead time from day 1 to 7. To examine the skill of the postprocessed streamflow forecasts, Fig. 4 displays the CRPSS (relative to the raw ensemble streamflow forecasts) versus the forecast lead time for all the selected basins, for both warm (Figs. 4a-d) and cool (Figs. 4e-h) seasons. In the cool season (Figs. 4e-h), the tendency is for both postprocessing techniques to demonstrate improved forecast skill across all the basins and lead times. The skill can improve as much as 40% at the later lead times (Fig. 4f). The skill improvements, however, from the ARX(1,1) postprocessor are not as consistent for the warm season (Figs. 4a-d), displaying negative skill values for some of the lead times in all the basins. The latter underscores an inability of the ARX(1,1) postprocessor to enhance the raw streamflow ensembles for the warm season. In some cases (Figs. 4b and 4e-f), the skill of the postprocessors shows an increasing trend with the lead time. This is the case since the skill is here measured relative to the raw
streamflow forecasts, which is done to better isolate the effect of the postprocessors on the streamflow forecasts.
Figure 3-4. 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.
The gains in skill from QR vary from ~0% (Fig. 4b at the day 1 lead time) to ~40% (Fig. 4f at lead times > 4 days) depending upon the season and lead time. The gains from ARX(1,1), on the other hand, vary from ~0% (Fig. 4g at the day 1 lead time) to a much lower level of ~28% (Fig. 4f at the day 4 lead time) during the cool season, while there are little to no gains in the warm season. In the cool season (Figs. 4e-h), both postprocessors exhibit somewhat similar performance at different lead times, with the exception of Fig. 4h, but in the warm season QR tends to consistently perform better than ARX(1,1). The overall trend in Fig. 4 is for QR to mostly 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, except WVYN6 in Fig. 4h, most of the lead times and for both seasons.

As discussed and demonstrated in Fig. 4, QR performs better than ARX(1,1). We also computed reliability diagrams, as determined by Sharma et al., (2017), for the two postprocessors (plots not shown) and found that QR tends to display better reliability than ARX(1,1) across lead times, basins, and seasons. Therefore, we select QR as the statistical streamflow postprocessor to examine the interplay between preprocessing and postprocessing in the RHEPS.

### 3.5.4 Verification of the ensemble streamflow forecasts

In this subsection, we examine the effects of different statistical processing scenarios on the ensemble streamflow forecasts from the RHEPS. The forecasting scenarios considered here are S3-S6 (Table 2 defines the scenarios). To facilitate presenting the verification results, this subsection is divided into the following three parts: CRPSS, CRPS decomposition, and BSS.
i) CRSS

The skill of the ensemble streamflow forecasts for S3-S6 is assessed using the CRPSS relative to the sampled climatology (Fig. 5). The CRPSS in Fig. 5 is shown as a function of the forecast lead time for all the basins, and the warm (Fig. 5a-d) and cool (Fig. 5e-h) seasons. The most salient feature of Fig. 5 is that the performance of the streamflow forecasts tends for the most part 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. Although there is some tendency for the large basins to show better forecast skill than the small ones, the scaling (i.e., the dependence of skill on the basin size) is rather mild and not consistent across the four basins.

In Fig. 5, 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. The gain in skill between S3 and S4 is generally small at the short lead times (< 3 days) but increases for the later lead times; this is somewhat more evident for the cool season than the warm one. This skill trend between S3 and S4 is not entirely surprising as we previously saw (Fig. 3) that differences between the raw and preprocessed precipitation ensembles are more significant at the later lead times. The skill in Fig. 5 then shows further improvements for both S5 and S6, relative to S4. Although S6 tends to outperform both S4 and S5 in Fig. 5, the differences in skill among these three scenarios are not as significant, their confidence intervals tend to overlap in most cases, with the exception of Fig. 5f where S4 underperforms relative to both S5 and S6. Fig. 5 shows that S6 is the preferred scenario in that it tends to more consistently improve the ensemble streamflow forecasts across basins, lead times and seasons than the other scenarios. It also shows that S5 may be slightly more effective than S4 in correcting streamflow forecast biases.
Figure 3-5. Continuous ranked probability skill score (CRPSS) 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.
There are also seasonal differences in the forecast skill among the scenarios. The skill of the streamflow forecasts tends to be slightly greater in the warm season (Figs. 5a-d) than in the cool one (Figs. 5e-h) across all the basins and lead times. In the warm season (Figs. 5a-d), all the scenarios tend to show similar skill, except CNON6 (Fig. 5b), with S5 and S6 only slightly outperforming S3 and S4. In the cool season (Figs. 5e-h), with the exception of CNON6 (Fig. 5f), the performance is similar among the scenarios for the short lead times but S3 tends to consistently underperform for the later lead times relative to S4-S6. There is also a skill reversal between the seasons when comparing the ensemble precipitation (Fig. 3) and streamflow (Fig. 5) forecasts. That is, the skill tends to be higher in the cool season than the warm one in Fig. 3, but this trend reverses in Fig. 5. The reason for this reversal is that in the cool season hydrological conditions are strongly influenced by snow dynamics, which can be challenging to represent with HL-RDHM, particularly when specific snow information or data are not available. In any case, this could be a valuable area for future research since it appears here to have a significant influence on the skill of the ensemble streamflow forecasts.

The underperformance of S4 in the CNON6 basin (Fig. 5f), relative to the other scenarios, is in part due to the unusually low skill of the raw ensemble streamflow forecasts of S3, so that even after preprocessing the skill improvement attained with S4 is not comparable to that associated with S5 and S6. This is also the case for CNON6 in the warm season (Fig. 5b). However, in addition, during the cool season it is likely that streamflows in CNON6 are affected by a reservoir just upstream from the main outlet of CNON6. The reservoir is operated for flood control purposes. The reservoir affects during the cool season low flows by maintaining them somewhat higher than in natural conditions. Since we do not account for reservoir operations in our hydrological modeling, it is likely that part of the benefits of postprocessing are in this case
to correct for this modeling bias. In fact, this is also reflected in the calibration results (e.g., in Fig. 2c), where the performance of CNON6 is somewhat lower than in the other basins. Interestingly, after postprocessing (S5 in Fig. 5f), the skill of CNON6 is as good as that of CINN6, even though at the day 1 lead time the skill for S3 is ~0.1 for CNON6 (Fig. 5f) and ~0.4 for CINN6 (Fig. 5e). Hence, the postprocessor seems capable to compensate some for the lesser performance of CNON6 in both calibration or after preprocessing in the cool season.

ii) CRS decomposition

Fig. 6 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. 6a-d) and cool (Figs. 6e-h) seasons. The components presented here are reliability (CRPS\text{rel}) and potential CRPS (CRPS\text{pot}) (Hersbach, 2000). CRPS\text{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\text{pot} represents the lowest possible CRPS that could be obtained if the forecasts were made perfectly reliable (i.e., CRPS\text{rel}=0). Note that the CRPS, CRPS\text{rel}, and CRPS\text{pot} are all negatively oriented, with perfect score of zero. Overall, as was the case with the CRPSS (Fig. 5), the CRPS decomposition reveals that forecast reliability tends mostly to progressively improve from S3 to S6.
Figure 3-6. Decomposition of the CRPS into CRPS potential (CRPSpot) and CRPS reliability (CRPSrel) 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.
Interestingly, improvements in forecast quality for S5 and S6, relative to the raw streamflow forecasts of S3, are mainly due to reductions in CRPS\textsubscript{rel} (i.e., by making the forecasts more reliable), whereas for S4 better forecast quality is achieved by reductions in both CRPS\textsubscript{rel} and CRPS\textsubscript{pot}. CRPS\textsubscript{pot} appears to play a bigger role in S4 than in the other scenarios, since in many cases in Fig. 6 the CRPS\textsubscript{pot} value for S4 is the lowest among all the scenarios. 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\textsubscript{pot} is sensitive to the spread (Messner et al., 2014a). In terms of the warm and cool seasons, the warm season tends to show a slightly lower CRPS than the cool one for all the scenarios. There are other, more nuanced differences between the two seasons. For example, S5 is more reliable than S4 in several cases in Fig. 6, such as for the day 1 lead time in the cool season. The CRPS decomposition demonstrates that the ensemble streamflow forecasts for S5 and S6 tend to be more reliable than for S3 and S4. It also shows that the forecasts from S5 and S6 tend to exhibit comparable reliability. However, the ensemble streamflow forecasts generated using both preprocessing and postprocessing, S6, ultimately result in lower CRPS than the other scenarios. The latter is seen across all the basins, lead times, and seasons, except in one case (Fig. 6d at the day 7 lead time).

iii) BSS

In our final verification comparison, the BSS of the ensemble streamflow forecasts for S5 (Figs. 7a-d) and S6 (Figs. 7e-h) are plotted against the non-exceedance probability associated with different streamflow 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. 7) and simulated (dashed lines in Fig. 7) flows. When the BSS is computed relative to observed flows, it considers the effect on forecast skill of both
meteorological and hydrological uncertainties. While the BSS relative to simulated flows is mainly affected by meteorological uncertainties. The difference between the two, i.e., the BSS relative to observed flows minus the BSS relative to simulated ones, provides an estimate of the effect of hydrological uncertainties on the skill of the streamflow 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 streamflow forecasts tends to decrease with lead time across the flow thresholds and basins. In contrast to the CRPSS (Fig. 5) where S6 tends for the majority of cases to slightly outperform S5, the BSS values for the different flow thresholds appear similar for S5 (Figs. 7a-d) and S6 (Figs. 7e-h). The only exception is CKLN6 (Figs. 7c and 7g) where S6 has better skill than S5 at the day 1 and 3 lead times, particularly at the highest flow thresholds considered. 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. 7a), 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. 7d).
Figure 3-7. 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).
The most notable feature in Fig. 7 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. 7, 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. 7a-b and 7e-f). For example, for a non-exceedance probability of 0.95 and at a day 1 lead time (Fig. 7b), the BSS value relative to the simulated and observed flows are ~0.79 and 0.38, respectively, suggesting a reduction of ~50% skill due to hydrological uncertainties.

3.6 Summary and conclusions

In this study, we used the RHEPS to investigate the effect of statistical processing on short-to medium-range ensemble streamflow forecasts. First, we assessed the raw precipitation forecasts from the GEFSRv2 (S1), and compared them with the preprocessed precipitation ensembles (S2). Then, streamflow 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 streamflow forecasts was done for the years 2004-2012, using four nested, gauge locations in the NBSR basin of the U.S. MAR. We found that the scenario involving both preprocessing and postprocessing consistently outperforms the other scenarios. In some cases, however, the differences between the scenario involving preprocessing and postprocessing, and the scenario with postprocessing alone are not as significant, suggesting for
those cases (e.g., warm season) that postprocessing alone can be effective in removing systematic biases. 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 streamflow forecasts in the cool season. In contrast, in the warm season, ARX(1,1) shows little or no gains in skill. Overall, for the majority of cases analyzed, the gains with QR tend to be greater than with ARX(1,1), specially during the warm season.
- In terms of the forecast skill (i.e., CRPSS), in the warm season the scenarios including only preprocessing and only postprocessing have a comparable performance to the more complex scenario consisting of both preprocessing and postprocessing. While in the cool season, the scenario involving both preprocessing and postprocessing consistently outperforms the other scenarios but the differences may not be as significant.
- 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 streamflow thresholds and the warm season. This assessment suggests that for high flow thresholds the similarities in skill between both scenarios, S5 and S6, become greater.
- Decomposing the CRPS into reliability and potential component, we observed that the scenario that combines preprocessing and postprocessing results in slightly lower CRPS than the other scenarios. We found that the scenario involving only postprocessing tends to demonstrate similar reliability to the scenario consisting of both preprocessing and postprocessing across most of the lead times, basins and seasons. We also found that in
several cases the postprocessing alone scenario displays improved reliability relative to the preprocessing alone scenario.

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.

### 3.7 Chapter 3 references


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.


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


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

Sharma, S., Siddique, R., Balderas, N., Fuentes, J. D., Reed, S., Ahnert, P., Shedd, R., Astifan,

Shi, X., Andrew, W. W., and Dennis, P. L.: How essential is hydrologic model calibration to

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.

precipitation forecasting using Bayesian model averaging, Monthly Weather Review, 135,

Smith, M. B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., Cui, Z., Mizukami, N.,
2: Motivation and design of the Oklahoma experiments, Journal of Hydrology, 418, 3-16,
2012a.


Chapter 4

Hydrological multimodel streamflow forecasts

4.1 Background and literature review

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 about near-real time hydrological forecasting conditions. However, at medium-range timescales (≥ 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 attempts 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 for the first time with hydrological multimodel forecasts.

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; Gneiting et al., 2005, 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 (Gomez et al., 2019; 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 the addition of new ensemble members (i.e., increasing ensemble size)? Answering the latter is relevant to operational forecasting because generating many ensemble members in real time is often not feasible or realistic, and may not be as effective if skill enhancements are dominated by model diversity.

4.2 Methodology

4.2.1 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 $\Delta$ be the forecasted variable, $D$ the training data, and $M = [M_1, M_2, \ldots, M_K]$ the independent predictions from a total of $K$ hydrological models, the pdf of the BMA probabilistic prediction of $\Delta$ can be expressed by the law of total probability as

$$P(\Delta | (M_1, M_2, \ldots, M_K)) = \sum_{k=1}^{K} P(\Delta | M_k)P(M_k | D),$$

(14)
where $P(\Delta|M_k)$ is the posterior distribution of $\Delta$ given the model prediction $M_k$, and $P(M_k|D)$ is the posterior probability of model $M_k$ being the best one given the training data $D$. $P(M_k|D)$ 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

$$\sum_{k=1}^{K} P(M_k|D) = 1.$$  \hspace{1cm} (15)

Thus, $P(M_k|D)$ 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

$$P(\Delta|\sigma_t, M_1^t, M_2^t, ..., M_k^t) = \sum_{k=1}^{K} w_k^t P(\Delta|\sigma_t, M_k^t).$$  \hspace{1cm} (16)

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 \( t \), \( f'_{k,t} \), are determined using

\[
f'_{k,t} = \overline{f}'_k + \hat{\xi}'_{k,t},
\]

where \( \overline{f}'_k \) is the ensemble mean forecast of model \( k \) at time \( t \), and \( \hat{\xi}'_{k,t} \) is the error estimate at the quantile interval \( \tau \) defined as

\[
\hat{\xi}'_{k,t} = a'_{k,t} + b'_{k,t} \overline{f}'_k.
\]

In equation (5), \( a'_{k,t} \) and \( b'_{k,t} \) are the regression parameters for model \( k \) and quantile interval \( \tau \) 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

\[
\arg \min_{\tau \in \mathbb{R}} \sum_{j=1}^{J} \Gamma'_\tau [\xi'_{\tau,j} - \hat{\xi}'_{\tau,j}(j, f'_j)].
\]

\( \xi'_{\tau,j} \) and \( f'_j \) are the \( j^{th} \) paired samples from a total of \( J \) samples; \( \xi'_{\tau,j} \) is computed as the observed flow minus the forecasted one at time \( t \); \( \Gamma'_\tau \) is the QR function for the \( \tau^{th} \) quantile at time \( t \) defined as

\[
\Gamma'_\tau (\Psi'_j) = \begin{cases} 
(\tau - 1)\Psi'_j & \text{if } \Psi'_j \geq 0 \\
\tau\Psi'_j & \text{if } \Psi'_j < 0 
\end{cases},
\]

and \( \Psi'_j \) is the residual term computed as the difference between \( \xi'_{\tau,j} \) and \( \hat{\xi}'_{\tau,j}(j, f'_j) \) 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). Note that the \( \tau \) values were chosen to cover the domain \([0, 1]\) sufficiently well, so that the lead time specific error estimate in equation (5) is a continuous distribution. Specifically, the number of \( \tau \) values were based on the
number of ensemble members required by a particular forecasting experiment and were chosen to vary uniformly between 0.06 and 0.96.

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

\[
f^\tau_{k,NQT} = G^{-1}(cdf(f^\tau_k)),
\]

where \(cdf(.)\) is the cdf of the bias-corrected forecasts from model \(k\) at time \(t\), \(f^\tau_k\); \(G\) is the standard normal distribution and \(G^{-1}\) its inverse; and \(f^\tau_{k,NQT}\) 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 \(w^\tau_k\) and variance \(\sigma^2_{k,\tau}\) of model \(k\) at the forecast lead time \(t\) are estimated using the log likelihood function. Note that \(\sigma^2_{k,\tau}\) is the variance associated with the Gaussian posterior distribution of model \(k\). Setting the parameter vector \(\theta = \{w^\tau_k, \sigma^2_{k,\tau}, k = 1, 2, ..., K\}\), the log likelihood function of \(\theta\) at the forecast lead time \(t\) is approximated as
\[ l(\theta) = \log\left( \sum_{k=1}^{K} w_k' g(\Delta'_{NQT} \mid f_{k,NQT}') \right), \]  

where \( g(.) \) denotes a Gaussian pdf, and \( \Delta'_{NQT} \) 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 \( \theta \) are determined using the expectation maximization (EM) optimization algorithm (Bilmes, 1998). The steps required to implement the EM algorithm are provided in Appendix. Finally, discrete ensembles are sampled from the postprocessed predictive distribution using the equidistant quantiles sampling approach (Schefzik et al., 2013).

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 streamflow forecasts were generated from GEFSRv2, while the multimodel forecasts were generated using the QR-BMA technique to optimally combine the single model forecasts. The single-model forecasts were postprocessed using QR following the same leave-one-out approach used with QR-BMA. Note that QR-BMA was applied here independently at each lead time, thus it is suitable for generating forecasts when predictions are needed for a single time.

### 4.2.2. Measures of forecast kill

i) **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 (i.e., the addition of new ensemble members). 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 $H(F)$ defined as

$$H(F) = -\int P(f) \ln P(f) \, df.$$  \hspace{1cm} (23)

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

$$H(O \mid F) = H(O, F) - H(F).$$  \hspace{1cm} (24)

With equations (10)-(11), the mutual information ($MI$) between the streamflow observations and the forecasts, $MI(O; F)$, is given by (Cover & Thomas, 1991)

$$MI(O; F) = H(O) + H(F) - H(O, F)$$

$$= \iint P(o, f) \log \left[ \frac{P(o, f)}{P(o)P(f)} \right] \, do \, df,$$  \hspace{1cm} (25)
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

$$CMI(O; F_2 | F_1) = MI(O; (F_1, F_2)) - MI(O; F_1),$$

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)

$$CMI(O; F_2 | F_1) = -\frac{1}{2} \log \left(1 - \rho_{O2|1}^2\right),$$

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)

$$\rho_{O2|1} = \frac{\rho_{O2} - \rho_{O1}\rho_{21}}{\sqrt{(1 - \rho_{O1}^2)(1 - \rho_{21}^2)}},$$
where $\rho_{o1}$ and $\rho_{o2}$ are the correlation skills of $F_1$ and $F_2$, respectively; and $\rho_{i2}$ 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 $\alpha$, and a deviation about the conditional mean, called the noise variable $\beta$. 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

$$
\rho_{o2 \alpha}^\text{noise} = \frac{\rho_{\alpha o}}{E_1} \frac{\sqrt{\text{SNR}}}{\sqrt{\text{SNR}} \left( \frac{1}{E_1} + \frac{1}{E_2} \right) + \frac{1}{E_1 E_2}} \sqrt{\text{SNR} \left( 1 - \rho_{\alpha o}^2 \right) + \frac{1}{E_1}},
$$

(29)

where the signal-to-noise ratio $\text{SNR}$ 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., $\text{SNR} \neq 0$), forecast skill exists ($\rho_{\alpha o} \neq 0$), and the ensemble sizes are finite. To the extent that forecast skill exceeds predictability skill,

$$
|\rho_{\alpha o}| \leq \sqrt{\frac{\text{SNR}}{\text{SNR} + 1}}.
$$

(30)

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 $SNR \to \infty$,

$$\rho_{\text{noise}}^{2} \leq \frac{E_{2}}{\sqrt{(E_{1} + E_{2})(E_{1} + 1)}}.$$  (31)

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.

To implement $CMI$, three different experiments were performed: 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 ($F_{1}$) combined with a 6-member ensemble from the same model ($F_{2}$). Note that this 6-member ensemble may be treated as proxy for adding members from hydrological models with very similar structures. This experiment was repeated for each of the models used. In the 9-m multimodel experiment, a 3-member single-model ensemble from one of the models ($F_{1}$) was combined with a 6-member multimodel ensemble obtained using the remaining two other models ($F_{2}$). This 6-member multimodel ensemble was generated as follows: i) 3 raw
members from each of the remaining two models were randomly selected, and ii) the selected members were combined using the QR-BMA postprocessor to generate a 6-member multimodel ensemble. Note that the number of ensemble members from each model are equal only in relation to the number of raw forecast members sampled from each model. Additionally, in both the 9-m single-model and 9-m multimodel experiments, the values of $E_1$ and $E_2$ in equation (18) are 3 and 6, respectively. The last experiment, 33-m multimodel, was the same as the 9-m multimodel experiment but using instead 33 members. That is, an 11-member single-model ensemble from one of the models ($F_1$) was combined with a 22-member multimodel ensemble obtained by postprocessing the remaining two other models ($F_2$). For the CMI experiments, raw single-model forecasts were used for $F_1$ to emulate basic operational conditions. The CMI values for the different experiments were computed by first randomly selecting raw ensemble members from each hydrological model. This process of randomly selecting raw forecasts from each model was repeated several times for each CMI value, so that the reported CMI value is the average from multiple realizations.

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.

**ii) 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

\[
CRPS = \int_{-\infty}^{\infty} [cdf(f) - \Pi(f - o)]^2 df,
\]

where

\[
\Pi(f) = \begin{cases} 
0 & \text{for } f < 0 \\
1 & \text{otherwise}
\end{cases}
\]

\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

\[
CRPSS = 1 - \frac{\overline{CRPS}_m}{\overline{CRPS}_r},
\]

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. Confidence intervals for the CRPSS were determined using the stationary block bootstrap technique (Politis and Romano, 1994).
Note that the CRPSS represents a quantitative measure of the overall forecast skill relative to the reference system (i.e., sampled climatology), whereas the CMI represents the skill improvement or enhancement provided by the multimodel forecasts. Thus, the CMI and CRPSS are not directly comparable against each other. Our proposed multimodel forecasting approach is summarized in Figure 1.

4.3 Experimental setup

4.3.1 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; 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 km². A summary of the main characteristics of the selected gauge locations is provided in Table 1.
Figure 4-1. 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.
**Figure 4-2.** 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 study area in the US.
Table 4-1. Characteristics of the selected gauged locations.

<table>
<thead>
<tr>
<th>Location of outlet</th>
<th>Cincinnatus, New York</th>
<th>Chenango Forks, New York</th>
<th>Conklin, New York</th>
<th>Waverly, New York</th>
</tr>
</thead>
<tbody>
<tr>
<td>NWS id</td>
<td>CINN6</td>
<td>CNON6</td>
<td>CKLN6</td>
<td>WVYN6</td>
</tr>
<tr>
<td>USGS id</td>
<td>01510000</td>
<td>01512500</td>
<td>01503000</td>
<td>01515000</td>
</tr>
<tr>
<td>Area [km²]</td>
<td>381</td>
<td>3841</td>
<td>5781</td>
<td>12362</td>
</tr>
<tr>
<td>Outlet latitude [North]</td>
<td>42°32′28″</td>
<td>42°13′05″</td>
<td>42°02′07″</td>
<td>41°59′05″</td>
</tr>
<tr>
<td>Outlet longitude [West]</td>
<td>75°53′59″</td>
<td>75°50′54″</td>
<td>75°48′11″</td>
<td>76°30′04″</td>
</tr>
<tr>
<td>Minimum daily flow [m³ s⁻¹]</td>
<td>0.31</td>
<td>4.05</td>
<td>6.80</td>
<td>13.08</td>
</tr>
<tr>
<td>Maximum daily flow [m³ s⁻¹]</td>
<td>172.73</td>
<td>1248.77</td>
<td>2041.64</td>
<td>4417.42</td>
</tr>
<tr>
<td>Mean daily flow [m³ s⁻¹]</td>
<td>8.89</td>
<td>82.36</td>
<td>122.93</td>
<td>277.35</td>
</tr>
</tbody>
</table>

*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.
4.3.2. Datasets

i) 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° 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. The GEFSRv2 data were bilinearly interpolated onto the regularly spaced grid required by the hydrological models. Table 2 summarizes key information about the GEFSRv2 dataset. Additional details about the GEFSRv2 can be found elsewhere (Hamill et al., 2013).
Table 4-2. Summary and Main Characteristics of the Datasets Used in this Study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Horizontal Resolution [km²]</th>
<th>Temporal Resolution [hour]</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meteoerological forecasts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEFSRv2</td>
<td>NCEP</td>
<td>~55 x 55 (0.5° x 0.5°)</td>
<td>3 (days 1-3) and 6 (days 4-7) hourly accumulations</td>
<td>Precipitation, near-surface temperature, specific humidity, surface pressure, downward short and long wave radiation, and u-v components of wind speed</td>
</tr>
</tbody>
</table>

Hydrometeorological observations

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Horizontal Resolution [km²]</th>
<th>Temporal Resolution</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLDAS-2</td>
<td>NASA</td>
<td>~13 x 13 (0.125° x 0.125°)</td>
<td>Hourly</td>
<td>Near-surface temperature, specific humidity, surface pressure, downward long and short wave radiation, and u-v components of wind speed</td>
</tr>
<tr>
<td>MPEs</td>
<td>MARFC</td>
<td>~4 x 4</td>
<td>Hourly</td>
<td>Gridded precipitation</td>
</tr>
<tr>
<td>Temperatures</td>
<td>MARFC</td>
<td>~4 x 4</td>
<td>Hourly</td>
<td>Gridded temperature</td>
</tr>
<tr>
<td>Gauge discharge</td>
<td>USGS</td>
<td>-</td>
<td>Hourly</td>
<td>Streamflow</td>
</tr>
</tbody>
</table>
ii) 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 km² 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 1/8th-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.

4.3.3 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 models were used to simulate and forecast flows over the entire period of analysis (years 2004-2009) at the selected gauge locations (Figure 1), but were verified for the warm season only (May-October). We focused on the warm season because flood events are more prevalent in our study area during these months. 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, warm season only (years 2004-2009, May-October). The HL-RDHM simulations were performed for the period 2004-2009, with the year 2003 used as warm-up. To calibrate HL-RDHM, we first manually adjusted the a-priori parameter fields through a multiplying factor; once the manual changes did not yield noticeable improvements in model performance, the multiplying factors were tuned up using the stepwise line search (SLS) algorithm (Kuzmin et al., 2008; Kuzmin, 2009). Out of all the HL-RDHM adjusted parameters, 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. The WRF-Hydro simulations were performed for the period 2004-2009, with the first year used as warm-up. To calibrate WRF-Hydro, 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. Out of all the adjusted parameters, the most sensitive parameters were the soil, groundwater and runoff parameters. After manually calibrating the WRF-Hydro parameters, the most sensitive parameter values were fine-tuned using an optimization algorithm, namely dynamically dimension search (DDS) (Tolson & Shoemaker, 2007). The API-Continuous model was previously calibrated by the MARFC for operational forecasting purposes using a manual approach.

Figure 3 summarizes the models’ performance in simulation mode using the Pearson’s correlation coefficient, $R$, Nash-Sutcliffe efficiency, $NSE$, and percent bias, $PB$, 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.
Figure 4-3. Performance of the hydrological models in simulation mode over the entire period of analysis (2004-2009, May-October): a) Pearson’s correlation coefficient, $R$, b) Nash-Sutcliffe efficiency, $NSE$, and c) percent bias, $PB$, between the daily simulated and observed flows.
4.3.4 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 consisted 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 (training period) to train the postprocessor and the remaining 2 years to verify the forecasts. This was repeated until all the 6 years of forecast data were postprocessed and verified independently of the training period. 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.
4.4 Results and Discussion

4.4.1 CRPSS verification of the single-model forecasts

i) 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 4-4. 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 relatively low performance of the WRF-Hydro could also be due to restrictions in its ability to represent physical processes because of a priori constraints in model parameter values, which neglect the large uncertainty in parameter estimates and large impact that parameters have on model predictions.

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, 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).

**ii) 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). However, since the hydrological models are calibrated with datasets used for cross-validating the postprocessor, the absolute CRPSS for the postprocessed forecasts are not representative of real time conditions.

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
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.

iii) 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 raw members per model, respectively, to obtain a multimodel forecast with QR-BMA; QR-BMA was used to optimize the weights. 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 alone, 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.

The BMA weights associated with the multimodel forecasts tend to reflect the performance of the postprocessed forecasts for the individual models in Figure 4. For example, the API at CNON6 consistently gets a higher weight than the other models, particularly at the longer lead times, while WRF-Hydro at CNON6, CKLN6 and WVYN6 has relatively low BMA weights at the later lead times. Additionally, the weights show that even when the performance of one of the models is dominant, the remaining models may still contribute to improving the multimodel forecasts. This is the case for CNON6 at the later
Figure 4-5. 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.
lead times (e.g., days 6 and 7 in Table S2), where despite the higher weights for API, the HL-RDHM and WRF-Hydro are still assigned some weight.

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. We also computed reliability diagrams, as determined by Brown et al., (2012), for the single model and 9-m multimodel forecasts. The reliability diagrams show that the multimodel forecasts tend, for the most part, to display better reliability than the single model forecasts.

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.
4.4.2 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 alone. 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 (Figures 6a-c), ii) 9-m multimodel (Figures 6d-f), and iii) 33-m multimodel (Figures 6g-i). The experiments are described in Subsection 2.2.1.

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 being more sensitive than the other models to the weather forcing. 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 4-6. 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 raw 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 raw members from each model). The standard deviation of the $CMI$ values varies from 0.02 to 0.06.

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 alone, 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 an important role 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 is a relevant consideration when trying to enhance the skill of streamflow forecasts. Although this is the case here for forecast skill, one would like in the future to examine whether these results apply to other attributes of
forecast quality. In particular, metrics that are more responsive to the ensemble size than the adopted CMI formalism, which was based on the ensemble mean, could be tried.

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.
4.5 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. Although the adopted CMI formalism does not capture all aspects of ensemble forecasts, it allows a robust analysis to decide whether the skill enhancement from multimodel forecasts is dominated by model diversity or is only due to the reduction of noise associated with the ensemble size. 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 only increasing the ensemble size of a single model, may be an 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 only due to the noise reduction associated with the ensemble size but with the ability of each model to contribute additional information to the forecasts.

**4.6 Chapter 4 references**


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


Chapter 5

Subseasonal to seasonal hydroclimatic predictions

5.1 Background and literature review

Hydroclimatic predictions at the subseasonal-to-seasonal (S2S) timescale are critically needed to inform water-related decisions and policies in multiple sectors, including agriculture (Mo et al. 2011), energy (De Felice et al. 2015), and water resources management (Wood and Lettenmaier 2006). The S2S timescale is defined as spanning the period between 2 weeks and 9 months (National Academies of Sciences, Engineering and Medicine 2016). This timescale has remained a major weather-climate prediction gap, since it encompasses the time frame where most of the information from atmospheric initial conditions is lost and the timeframe is too short to be strongly influenced by climate modes of variability (National Academies of Sciences, Engineering and Medicine 2016). However, recent and significant improvements (e.g., improved observational datasets, spatial resolution, model physics, initial conditions, and assimilation techniques) in coupled atmosphere-ocean general circulation models (GCMs) are providing new opportunities to potentially enhance hydrological forecasting at the S2S timescale (Li et al. 2009, Mendoza et al. 2017, Mo et al. 2012, Shukla and Lettenmaier 2011, Yuan and Wood 2012, Yossef et al. 2013). To test this idea, we build and implement a new dynamical-statistical hydroclimatic ensemble prediction system. The system is forced with outputs from the NOAA’s National Centers for Environmental Prediction (NCEP) Climate Forecast System version 2 (CFSv2). CFSv2 is a recently developed, state-of-the-science GCM (Saha et al., 2014) designed to enhance global S2S climate predictions.
Previous S2S hydroclimatic studies have demonstrated the increased value of CFSv2 outputs over previous generation S2S climate forcing datasets (Yuan et al. 2013, Yuan et al. 2014, Yuan and Wood 2012), and shown that new hydrological forecast systems can be built to seamlessly integrate weather and climate forecast timescales (Yuan et al. 2014). They have not, however, proven the ability of S2S climate predictions to support water quality forecasting. We show here for the first time the ability of S2S climate predictions to result in both skillful water quantity and quality forecasts. This has important implications for US national-level, and possibly other regions, operational hydrological forecasting strategies (Demargne et al. 2014, Gochis et al. 2015, Salas et al. 2018). We demonstrate that it is possible to equip current operational forecasting systems, in a cost and time efficient manner, with new prediction capabilities for water quality, which are currently lacking. Water quality predictions are increasingly needed at the S2S timescale, e.g., to anticipate freshwater ecological disasters such as harmful algal blooms (Michalak et al. 2013). For instance, by increasing the lead time at which excess nutrient loadings are predicted, natural resource managers and officials may be able to undertake further in advance mitigating measures to reduce nutrient-related impacts on sensitive water bodies, estuaries and lakes (National Academics of Sciences, Engineering and Medicine 2016).

Several approaches have been adopted to generate S2S hydroclimatic predictions. The approaches can be generally grouped into three main categories: statistical, dynamical, and hybrid. The statistical approach uses an empirical relationship between the streamflow forecasts and potential sources of predictability originating from various processes in the atmosphere, ocean and land (Lee et al. 2018, Mendoza et al. 2017, Pagano et al. 2009, Schepen et al. 2016, Slater and Villarini 2018). Such statistical approaches require long and reliable historical time-series for model fitting, and often lack robustness for nonstationarities, such as those induced by
climate and land-use/land-cover changes (Wood and Schaake, 2008). In the dynamical approach, outputs from GCMs are used as inputs into a hydrological model to produce streamflow forecasts (Yuan and Wood 2012). The hybrid approach combines the strengths of both dynamical and statistical approaches (Mendoza et al 2017), and therefore this approach is implemented here.

Dynamic hydrologic simulation models often include a module to simulate water quality constituents (Abbaspour et al 2015, Bekele and Nicklow 2005, Easton et al 2008). However, such systems require a wide range of spatial and temporal observations to be properly configured for forecasting purposes. Water quality data are scarce, making the effective calibration of associated model parameters a difficult challenge (Jackson-Blake et al 2017). On the other hand, current operational hydrological forecasting systems have been designed primarily for streamflow forecasting purposes (Alfieri et al 2014, Demargne et al 2014, Pagano et al 2016). Upgrading these modeling systems with process-based water quality capabilities would require considerable effort and time. An alternative, the one being proposed here, is to implement a hybrid dynamical-statistical modeling approach that combines dynamic hydroclimatic predictions with statistical models to efficiently generate water quality predictions. Additionally, several recent studies have shown that precipitation is a dominant driver of monthly nutrient loads at the basin scale (Sinha and Michalak 2016, Bastola and Mishra 2015). This supports the possibility of using S2S precipitation predictions to generate skillful nutrient load forecasts months in advance, providing the motivation for the present research.

The objective of this letter is to demonstrate the ability of S2S climate forecasts to result in both skillful water quantity (streamflow) and quality (nutrients and sediments) predictions using a new dynamical-statistical ensemble prediction system. This system is comprised of: i) hydrometeorological and water quality observations, ii) S2S climate reforecasts from the CFSv2,
iii) distributed hydrological model, and iv) statistical water quantity/quality model. The system is used to generate S2S ensemble predictions of streamflow, total nitrogen (TN), total phosphorus (TP), and total suspended sediments (TSS) at biweekly and monthly timescales out several months. The generated water quantity and quality forecasts are used to address the following questions: How skillful are the CFSv2 flow forecasts? Can the CFSv2 flow forecasts be used to anticipate nutrient and sediment loads? What is the relative importance of different sources of uncertainty (climatic and hydrological) in the flow, and nutrient and sediment load forecasts? How does the skill of the water quantity and quality forecasts varies with different forecasting conditions (e.g., lead time, season, and flow threshold) and watershed characteristics (e.g., basin size and land cover)?

5.2 Materials and methods

5.2.1 Forecasting and observational datasets

We use S2S precipitation and near-surface air temperature predictions from the CFSv2 (http://cfs.ncep.noaa.gov/) as forcing to generate S2S water quantity and quality predictions. The CFSv2 is a fully coupled dynamical prediction system representing the interactions between the ocean, land and atmosphere. The system runs at T126L64 spatial resolution (~0.94° Gaussian grid spacing or ~ 100 km) and 6-hourly temporal resolution (Saha et al 2014). We utilize a total of 15 years (2002-2016) of CFSv2 reforecast data. CFSv2 consists of 4 control runs per day at 00, 06, 12, and 18 Coordinated Universal Time (UTC) cycles, out 9 months and initiated every 5th day beginning in January 1st of each year. We use 90 days of CFSv2 forecasts because previous findings have shown that CFSv2 forecasts for precipitation and temperature tend to be
the most skillful at lead times <90 days (Tian et al 2017). For both precipitation and temperature forecasts, time-lagged ensembles consisting of 8 members are used. The ensembles are obtained by combining 4 runs per day from 2 consecutive initialization dates. The CFSv2 data are bilinearly interpolated onto the 4 x 4 km² grid cell resolution of the hydrological model used. Note that the CFSv2 is a global S2S climate forecast system, so that our proposed methods have potential to be implemented in basins across the US.

For the hydrometeorological observational datasets, we use multi-sensor precipitation estimates (MPEs), gridded near-surface air temperature, and streamflow observations at selected US Geological Survey (USGS) gages. These observational datasets are used to calibrate and verify the hydrological model, perform the hydrological model simulations, and obtain initial conditions for the forecasting runs. The forecasting runs are performed for the period 2002-2016. Both hourly gridded MPEs and near-surface air temperature data at 4 x 4 km² are obtained from the NOAA’s Middle Atlantic River Forecast Center. Daily streamflow observations for the selected gaged locations are obtained from the USGS (https://waterdata.usgs.gov/nwis/rt). Additionally, we use monthly estimates of TN, TP, and TSS loads at the selected gage locations to build and test a new statistical model for water quality forecasting. The water quality data are obtained from the Chesapeake Bay Nontidal Network Stations (https://www.sciencebase.gov/catalog/item/59c65eb1e4b017cf313f0aea) (Moyer 2016).

5.2.2 Dynamical-statistical approach to S2S water quantity and quality predictions
We propose and describe a general (i.e., that can be applied to any US basin) dynamical-statistical approach for generating consistent S2S water quantity and quality ensemble forecasts. The approach combines dynamical CFSv2-based flow predictions with a statistical model to generate nutrient and sediment loads at S2S timescales. To build and implement this dynamical-statistical system, the following three main modeling components are integrated: i) statistical climate preprocessor, ii) distributed hydrological model, and iii) statistical water quantity/quality model. The statistical climate preprocessor is used to correct systematic biases in the CFSv2 predictions. GCM outputs from the CFSv2 are often characterized by the presence of systematic biases (Saha et al 2014). These biases in the climate forcing propagate through the hydrological model to result in biased streamflow predictions (Cloke and Pappenberger 2009). Statistical preprocessing techniques are used to correct such biases (Crochemore et al 2016, Lucatero et al 2018, Shukla and Lettenmaier 2011, Yuan and Wood 2012). We use a logistic regression model to statistically preprocess the CFSv2 forecasts (Messner et al 2014a-b, Yang et al 2017). This model has been successfully used before with weather forecasts (Sharma et al 2018) but is tested here for the first time with climate forecasts. Different logistic regression models are used for near-surface temperature and precipitation forecasts to take into consideration that precipitation is a censored variable, i.e., it can only take positive values. The logistic regression models used have the important advantage of accounting for forecast heteroscedasticity by using the ensemble spread as a predictor. Additionally, they allow obtaining the full continuous predictive probability density function of the forecast variables.

The statistically preprocessed CFSv2 predictions are then used as forcing to a hydrological model to generate the S2S streamflow forecasts. NOAA’s Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM) is used as the spatially distributed hydrological
model (Koren et al. 2004). HL-RDHM is run at a 4-km spatial resolution and daily time step for
the selected gage locations.

The streamflow forecasts from HL-RDHM are statistically postprocessed to remove any
systematic biases using a quantile regression (QR) model (Koenker 2005). Different QR models
are used to correct systematic biases in the streamflow forecasts and to generate bias-corrected
TN, TP and TSS load forecasts. The QR model has the relevant advantages of accounting for
non-Gaussian error distributions and being less sensitive to the tail behavior of the forecasts. The
latter is because QR results in conditional quantiles rather than conditional means and,
consequently, the regression coefficients are robust with respect to outliers. The QR model is
implemented here separately for each lead time and basin considered. Note that some of the
statistical models commonly used to estimate water quality constituents (Hirsch et al. 2010,
Preston and Brakebill 1999, Runkel et al. 2004) rely on multiple linear regression techniques,
making them susceptible to outliers and requiring that model errors do not deviate significantly
from normality, which is often not the case. The QR model employed here overcomes these
drawbacks.

Next, we discuss the statistical preprocessor and statistical postprocessor in detail.

i) Statistical preprocessor

We use logistic regression to preprocess the CFSv2 forecasts (Messner et al. 2014a-b, Yang
et al. 2017). Different regression models are used for temperature and precipitation forecasts to
take into consideration that precipitation is a censored variable, i.e., it can only take on values
≥0. The logistic regression models used were recently developed for weather forecasts. They
have the advantage of accounting for forecast heteroscedasticity by using the ensemble spread as
a predictor. In addition to deriving coherent probabilistic forecasts, these logistic regression
models allow obtaining the full continuous predictive probability density function of the forecast variables. The logistic regression models are described next.

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

$$P(y \leq q \mid x) = \Lambda[\delta(x)],$$

where $\Lambda(.)$ denotes the cumulative distribution function of the standard logistic distribution, $y$ is the forecast variable, $q$ is a specified threshold, $x$ is a predictor variable that depends on the forecast members, and $\delta(x)$ is a linear function of the predictor variable $x$.

Equation (1) requires separate logistic regressions 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 avoid these inconsistencies, Wilks (2009) included a transformation of the predictand thresholds as an additional predictor variable in the logistic regression

$$P(y \leq q \mid x) = \Lambda[\omega(q) - \delta(x)],$$

where the transformation $\omega(.)$ is a monotone nondecreasing function. In addition to avoiding negative probabilities, equation (2) has the advantage that fewer parameters need to be estimated; instead of having a linear function $\delta(x)$ for each threshold, $\delta(x)$ is now the same for all the thresholds. This is especially advantageous for small training datasets.

Furthermore, to appropriately utilize the uncertainty information in the ensemble spread, Messner et al (2014a) proposed the Heteroscedastic extended logistic regression (HXLR) postprocessor. HXLR uses an additional predictor vector $\phi$ to control the dispersion of the logistic predictive distribution,

$$P(y \leq q \mid x) = \Lambda \left\{ \frac{\omega(q) - \delta(x)}{\exp[\eta(\phi)]} \right\},$$

where $\eta(.)$ is another linear function of the predictor variable $\phi$. 
The functions \( \delta(.) \) and \( \eta(.) \) are defined as
\[
\delta(x) = a_0 + a_1 x, \quad (38)
\]
and
\[
\eta(\varphi) = b_0 + b_1 \varphi, \quad (39)
\]
where \( a_0, a_1, \) and \( b_0, b_1 \) are parameters that need to be estimated.

Maximum likelihood estimation with the log-likelihood function is used to estimate the parameters associated with equation (3) (Messner et al. 2014a-b). One variation of the HELR that can accommodate nonnegative variables that are continuous for positive values and have a natural threshold at zero, such as precipitation amounts, is heteroscedastic censored logistic regression (HCLR) (Messner et al. 2014b). To implement the HCLR model, the predicted probability or likelihood \( \pi_i \) of the \( i^{th} \) observed outcome, accounting for the fact that \( y \geq 0 \), is determined as follows
\[
\pi_i = \begin{cases} 
\lambda \left[ \o(0) - \delta(x) \right] / \left[ \exp[\eta(\varphi)] \right], & y_i = 0 \\
\lambda \left[ \o(y_i) - \delta(x) \right] / \left[ \exp[\eta(\varphi)] \right], & y_i > 0,
\end{cases} \quad (40)
\]
where \( \lambda[.] \) denotes the likelihood function of the standard logistic function. In essence, HCLR fits a logistic error distribution with point mass zero to the transformed predictand. To transform the precipitation ensembles, we use a square root transformation, which has been shown to work well under various conditions (Yang et al., 2017; Sharma et al., 2018). Consequently, the predictor variables \( x \) and \( \varphi \) are the mean and standard deviation of the square root transformed precipitation ensemble forecasts.

HXLR is implemented using the daily-average temperature forecasts, whereas HCLR is employed using daily precipitation accumulations. However, to verify the preprocessed CFSv2
forecasts, biweekly and monthly values are used since this is timescale at which the streamflow and nutrient loads forecasts need to be generated. Both preprocessor models are applied to each CFSv2 grid cell and lead times ranging from 1 to 90 days in the JRB. At each grid cell, the preprocessor models are implemented for the period 2002-2016 using a leave-one-out cross-validation approach. For this, we select a stationary period of 14-years for training and the remaining 1-year is used for verification purposes. This is repeated for each season and year, until all the 15 years of data are preprocessed and verified independently of the training period. Thus, to forecast a given season and specific lead time, we use \( \sim 10,080 \) forecasts (i.e., 8 members \( \times \) 90 days per season \( \times \) 14 years). Finally, the Schaake shuffle method (Clark et al., 2004) is applied to each individual forecast lead time to maintain the observed space-time variability in the preprocessed CFSv2 temperature and precipitation forecasts.

**ii) Statistical postprocessor**

Quantile Regression (QR) (Koenker 2005) is employed to both i) statistically postprocess the ensemble streamflow forecasts, and ii) generate the ensemble nutrients and sediment load forecasts. The QR model has the advantages of accounting for non-Gaussian error distributions and being less sensitive to the tail behavior of the forecasts (e.g., streamflow). The latter is because QR results in conditional quantiles rather than conditional means and, consequently, the regression coefficients are robust with respect to outliers. The QR model is implemented here separately for each lead time and basin. The QR model estimates lead time \( t \)-specific quantiles of a conditional distribution of streamflow,

\[
\Phi_t = \{V_{\tau_1}, V_{\tau_2}, \ldots, V_{\tau_T}\},
\]

where \( \tau \) is the number of quantiles \( \tau \) considered. In the case that \( \tau \) is sufficiently large and the quantiles \( \tau \) covers the entire domain of \([0, 1]\), we consider \( \Phi_t \) to be a continuous distribution.
For every lead time $t$ considered, and for every quantile $\tau$, QR model between the verification $V$ and predictor variable, i.e., forecast $F$, is developed as:

$$V_t^\tau = c_t^\tau + d_t^\tau F^\tau,$$

In equation (9), $c_t^\tau$ and $d_t^\tau$ are the regression parameters for quantile interval $\tau$ at time $t$. These parameters are determined separately by minimizing the sum of the residuals from a training dataset as follows:

$$\arg\min_{c, d} \sum_{t=1}^T \Omega_t \left[ (v_{t,j}^\tau - (c_t^\tau + d_t^\tau f_{t,j}^\tau)) \right].$$

$v_{t,j}^\tau$ and $f_{t,j}^\tau$ are the $j^{th}$ paired forecast-verification samples from a total of $J$ samples; $\Omega_t^\tau$ is the QR function for the $\tau^{th}$ quantile defined as:

$$\Omega^\tau = \begin{cases} \{ r_{t,j}^\tau & \text{if } \psi_{t,j}^\tau < 0 \\ \psi_{t,j}^\tau & \text{otherwise} \end{cases}$$

and $\psi_{t,j}^\tau$ is the residual term computed as the difference between the verification $v_{t,j}^\tau$ and the linear QR estimate $(c_t^\tau + d_t^\tau f_{t,j}^\tau)$ for any quantile. The resulting minimization problem in equation (10) is solved using linear programming via the interior point method.

In order to postprocess the raw ensemble streamflow forecasts generated from HL-RDHM, we implemented QR as discussed in equations (9)-(11). For this, we used equation (9) with streamflow observations as the verification $V$ and ensemble mean forecast as the predictor variable $F$. Optimal parameter values in equation (9) are then computed using equations (10) and (11) from the training dataset. Finally, during the verification period, these optimal regression parameters are used to fit equation (9) and generate calibrated ensemble streamflow forecasts for any quantile. The QR postprocessor is employed here with biweekly accumulated streamflow forecasts. We therefore aggregated daily streamflow forecasts and generate biweekly values. The
postprocessor is implemented for the years 2003-2016, using the same leave-one-out approach used for the preprocessor.

Here we use QR postprocessor to relate streamflow to TN, TP and TSS loads at each of the selected gage locations. That is, monthly nutrients and sediments loads are generated by using monthly accumulated streamflow as a predictor variable in QR model (equation 9). Basically, QR model is trained using monthly accumulated streamflow, and the available observations of nutrients and sediment loads. Note that, the use of available observations of nutrients and sediment loads in the training dataset facilitate to obtain the bias-corrected load estimates. In order to obtain the simulated nutrients and sediment loads, log-linear relationship is employed in equation (9) using the respective load observations as verification $V$ and simulated streamflow as the predictor variable $F$. However, forecasts of nutrients and sediment loads are generated by using the raw ensemble streamflow forecast mean as the predictor variable $F$ in equation (9). Finally, ensemble nutrients and sediment load forecast are determined using equations (9)-(11) by varying the values of $\tau$ over the entire domain of $[0, 1]$.

### 5.2.3 Implementation and verification strategy

We implement the dynamical-statistical ensemble prediction system in the James River Basin (JRB). The JRB is the third largest tributary to the Chesapeake Bay, after the Susquehanna and Potomac River basins. The Chesapeake Bay is the largest estuary in the US. Nutrient pollution is a major environmental and economic concern in the Chesapeake Bay (Pyke et al 2008). We select the JRB because S2S water quantity and quality predictions are particularly challenging in this basin. This is due, among other factors, to the basin having little influence from known climate teleconnections and having only a relatively small fraction of agricultural land cover,
which is typically a major source of excess nutrients in rivers. Therefore, we expect that any
gains in S2S water quantity and quality predictability in this basin would be indicative of
potential gains in many other US basins with more favorable forecasting conditions, e.g., where
teleconnections are stronger and agricultural activities more dominant. The JRB has an overall
drainage area of 17,504 km², and a mean annual precipitation and temperature of ~1,118 mm and
~13.15 °C, respectively (Kang and Sridhar 2017). The climate in the JRB is considered humid
subtropical. The land cover distribution is 75% forest, 15% cropland and pasture, and ~ 7% urban. In the JRB, we select six different USGS daily gage stations as the streamflow and water
quality forecast locations (Table 1). The selected locations are the Bullpasture River at
Williamsville (BPRV2; USGS gauge 02015700), Calfpasture River above Mill Creek at Goshen
(GOHV2; USGS gauge 02020500), James River at Blue Ridge (JBIv2; USGS gauge 02024752),
Rivanna River at Palmyra (PYAV2; USGS gauge 02034000), James River at Cartersville
(CARV2; USGS gauge 02035000), and James River near Richmond (RMDV2; USGS gauge
0203700). This last gage location is at the overall outlet of the basin, just upstream of the tidal
section of the river.
### Table 5-1: Main characteristics of the study basins.

<table>
<thead>
<tr>
<th>Location of outlet</th>
<th>NWS id</th>
<th>USGS id</th>
<th>Area [km²]</th>
<th>Outlet latitude</th>
<th>Outlet longitude</th>
<th>Minimum daily flow&lt;br&gt;(m³ s⁻¹)</th>
<th>Maximum daily flow&lt;br&gt;(m³ s⁻¹)</th>
<th>Mean daily flow&lt;br&gt;(m³ s⁻¹)</th>
<th>Climatological flow&lt;br&gt;(Pr&lt;0.05)&lt;br&gt;(m³ s⁻¹)</th>
<th>Climatological flow&lt;br&gt;(Pr&gt;0.95)&lt;br&gt;(m³ s⁻¹)</th>
<th>Forest (%)</th>
<th>Pasture (%)</th>
<th>Developed Cropland (%)</th>
<th>Streamflow Simulation&lt;br&gt;(NSE)</th>
<th>TN (NSE)</th>
<th>TP (NSE)</th>
<th>TSS (NSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Williamsville, Virginia</td>
<td>BPRV2</td>
<td>02015700</td>
<td>285</td>
<td>38°11'43&quot;</td>
<td>79°34'14&quot;</td>
<td>0.71 (0.62)</td>
<td>123.46 (246.36)</td>
<td>4.62 (4.33)</td>
<td>0.98 (0.91)</td>
<td>14.01 (13.28)</td>
<td>81.15</td>
<td>14.60</td>
<td>0.28</td>
<td>0.70</td>
<td>0.73</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Goshen, Virginia</td>
<td>GOHV2</td>
<td>02020500</td>
<td>365</td>
<td>37°59'16&quot;</td>
<td>79°29'38&quot;</td>
<td>0.09 (0.02)</td>
<td>232.76 (620.14)</td>
<td>5.07 (4.78)</td>
<td>0.19 (0.16)</td>
<td>19.48 (18.07)</td>
<td>89.45</td>
<td>6.69</td>
<td>0.24</td>
<td>0.71</td>
<td>0.69</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Palmyra, Virginia</td>
<td>PYAV2</td>
<td>02034000</td>
<td>1717</td>
<td>37°51'28&quot;</td>
<td>78°15'58&quot;</td>
<td>0.69 (0.15)</td>
<td>656.95 (1925.55)</td>
<td>20.23 (20.22)</td>
<td>2.02 (1.96)</td>
<td>58.05 (60.03)</td>
<td>64.91</td>
<td>19.88</td>
<td>0.33</td>
<td>0.77</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Big Island, Virginia</td>
<td>JBV2</td>
<td>02024752</td>
<td>7967</td>
<td>37°33'19&quot;</td>
<td>79°22'03&quot;</td>
<td>13.67 (13.68)</td>
<td>1458.32 (1458.32)</td>
<td>99.10 (99.78)</td>
<td>19.62 (19.57)</td>
<td>325.64 (330.59)</td>
<td>81.08</td>
<td>11.68</td>
<td>0.45</td>
<td>0.73</td>
<td>0.85</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Cartersville, Virginia</td>
<td>CARV2</td>
<td>02035000</td>
<td>16192</td>
<td>37°40'15&quot;</td>
<td>78°05'10&quot;</td>
<td>13.42 (12.46)</td>
<td>2860.00 (7928.7)</td>
<td>99.78 (99.78)</td>
<td>31.71 (30.58)</td>
<td>625.80 (605.98)</td>
<td>69.67</td>
<td>14.82</td>
<td>0.39</td>
<td>0.73</td>
<td>0.85</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>Richmond, Virginia</td>
<td>RMDV2</td>
<td>02037500</td>
<td>17490</td>
<td>37°33'47&quot;</td>
<td>77°32'50&quot;</td>
<td>12.46 (12.46)</td>
<td>3001.59 (8381.79)</td>
<td>198.76 (197.00)</td>
<td>33.13 (19.33)</td>
<td>656.95 (622.97)</td>
<td>66.77</td>
<td>4.23</td>
<td>0.86</td>
<td>0.70</td>
<td>0.75</td>
<td>0.70</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*aThe numbers in parenthesis are the historical values based on entire available period of record.

bPr>0.95 indicates flows with exceedance probability of 0.05.

cPr<0.05 indicates flows with nonexceedance probability of 0.05.

dDifferent land cover percentages are calculated using the National Land Cover Database 2011.
The annual average daily discharge at the outlet (RMDV2) is 78,000 m$^3$/s for the chosen study period 2003-2016. The year 2003 shows the highest annual average streamflow (151,000 m$^3$/s) and annual TN load (905,600 lbs), while the lowest annual average streamflow (42,500 m$^3$/s), and the lowest annual TN (145,820 lbs), TP (20,969 lbs), and TSS load (9,902,700 lbs) occur in the year 2008. The selected forecast locations show different behaviors in terms of the nutrient loads. RMDV2, CARV2 and GOHV2 exhibit low variations in TN concentrations despite large variations in discharge, implying that these basins behave chemostatically. In contrast, TN concentrations decrease with increasing discharge at JBIV2, PYAV2 and BPRV2, implying that dilution is the dominant process controlling TN concentrations in these basins. Further information about the selected basins is included in Table 1.

To verify the skill of the water quantity and quality predictions, we use the mean Continuous Ranked Probability Skill Score (CRPSS) and the mean Brier Skill Score (BSS). The CRPSS is used because it accounts for ensemble or probabilistic information when computing the prediction skill. The BSS is used because it remains proper for conditioned predicted variables, i.e., it allows computing the skill of a predicted variable conditioned on another. 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’ cumulative distribution function and the corresponding observations. Thus, the CRPS accounts for both the forecast ensemble spread, or uncertainty, and forecast accuracy. Let the probability distribution function of the ensemble forecasts be $p(z)$ and the actual value be $z_a$, then the CRPS is defined as

$$CRPS = \int_{-\infty}^{\infty} \left[ P_f(z) - P_a(z) \right]^2 dz,$$

where $P_f(z)$ and $P_a(z)$ are cumulative functions given by
\[ P_f(z) = \int_{-\infty}^{z} p(z)dz \] and
\[ P_s(z) = H(z - z_s), \]
respectively. \( H \) is the Heaviside step function which is 1 if the argument is positive and zero otherwise. To measure the skill of the forecasting system relative to a reference system, the associated skill score or CRPSS is computed as
\[ CRPSS = 1 - \frac{\overline{CRPS}_m}{\overline{CRPS}_r}, \]
where \( \overline{CRPS}_m \) and \( \overline{CRPS}_r \) are the average CRPS values for the main forecasting system (i.e., the system to be evaluated) and reference forecast system, respectively. 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 use sampled climatology as the reference forecasting system.

The BS is analogous to the mean square error over the \( n \) forecast-observation pairs, but where the forecast is a probability and the observation is either 0 or 1. The BS is given by
\[ BS = \frac{1}{n} \sum_{i=1}^{n} \left[ P_f(q) - P_s(q) \right]^2, \]
where the probability of \( z_i \) to exceed a fixed threshold \( (q) \) is
\[ P_s(q) = P[z_i > q], \quad \text{and} \]
\[ P_f(q) = \begin{cases} 1, & f_i > q; \\ 0, & \text{otherwise.} \end{cases} \]
In order to measure the skill score of the main forecast system with respect to the reference forecast, the associated skill score or BSS is computed as
\[ BSS = 1 - \frac{BS_m}{BS_r}. \]
where $BS_{m}$ and $BS_{r}$ are the $BS$ values for the main and reference forecasting system, respectively. Similar to the CRPSS, any positive values of the $BSS$, from 0 to 1, indicate that the main forecasting system performed better than the reference one. Thus, a $BSS$ of 0 indicates no skill and a $BSS$ of 1 indicates perfect skill.

Besides the CRPSS and $BSS$, we also employ a commonly used deterministic metric, namely the correlation coefficient, to measure the degree of linear association between the mean ensemble forecasts and corresponding observations. Confidence intervals for the verification metrics are determined using the stationary block bootstrap technique (Politis and Romano 1994). Both the raw and postprocessed streamflow forecasts are verified for biweekly and monthly accumulations. Nutrients and sediments loads are verified for monthly load forecasts. The verification metrics (CRPSS, BSS, and correlation coefficient) are all computed following a leave-one-out cross-validation approach so that all the data are used for verification. For this, we select one year for verification purposes, and the remaining years for training. This is repeated for every year until all the available data have been postprocessed and verified independently of the training period.

5.3 Results

5.3.1 Performance of the water quantity and quality simulations

To assess the performance of HL-RDHM, the model is used to generate daily streamflow simulations for the entire period of analysis (years 2003–2016) and all the selected basins, except
JBIV2. For JBIV2, streamflow simulations are generated for the period 2007-2016 since streamflow data are not available prior to 2007. We aggregate the daily simulated streamflow values to the monthly timescale, and verify those values against monthly accumulated observed flows. The model performance is assessed using the Nash-Sutcliffe efficiency (NSE) index (Nash and Sutcliffe 1970). The NSE values for the streamflow, TN, TP, and TSS simulations are included in Table 1. Overall, the performance of the streamflow simulations is satisfactory with NSE>0.70 for all the basins. The simulated streamflow tends to display similar NSE values across the different basins.

The QR statistical model is used to simulate TN, TP and TSS loads in the selected basins. Note that the QR model is used to both simulate and forecast water quality by using the simulated and forecasted streamflow, respectively, as predictors. We find, overall, that the performance of the simulated TN, TP and TSS loads with the QR model is comparable to that of streamflow, with NSE values greater than 0.6 (Table 1). For most of the basins, the NSE values for the simulated TN loads are slightly better than those for the TP and TSS loads, which tend to be very similar across basins. In general, the performance of the QR model in simulating TN, TP, and TSS loads is satisfactory.

5.3.2 Skill of the raw and preprocessed CFSv2 predictions

Prior to evaluating the water quantity and quality forecasts, we assess the performance of the CFSv2 forecasts using the CRPSS as the measure of forecast skill. The CRPSS (relative to sampled climatology) is determined for both the raw and preprocessed CFSv2 near-surface temperature (Figures 1a-b) and precipitation (Figures 1c-d) ensemble forecasts. The forecast verification is done for lead times of up to 3 months using both biweekly and monthly
aggregations. To verify the CFSv2 ensemble predictions, each CFSv2 grid cell is treated as a separate verification unit. Thus, for the JRB, the CRPSS values are obtained by averaging the verification results from the different verification units that fall within the basin boundary.

The raw biweekly ensemble, near-surface temperature forecasts demonstrate positive skill up to ~5 weeks, beyond which the skill continues to degrade becoming worse than the sampled climatology (Figure 1a). In the case of raw ensemble precipitation forecasts, positive skill is observed only for lead times up to ~3 weeks for biweekly accumulations (Figure 1c); hence, the skill of monthly accumulations is low for month 1 and negligible or negative beyond that (Figure 1d). The CFSv2 raw precipitation forecasts demonstrate lower skill than the near-surface temperature forecasts across all the lead times for both biweekly and monthly accumulations (Figure 1).

After statistically preprocessing the raw CFSv2 forecasts, both the near-surface temperature and precipitation forecasts are more skillful than the raw forecasts (Figure 1). The relative improvements in skill are generally greater at the later forecast lead times. As was the case with the raw forecasts, the preprocessed near-surface temperature forecasts remain more skillful than the precipitation forecasts. The biweekly preprocessed, near-surface temperature forecasts remain skillful up to ~7 weeks (Figure 1a), with ~2 weeks of skill gain compared to the raw forecasts. This skill gain in the biweekly forecasts results in enhanced forecast skill at the monthly timescale (Figure 1b), especially for months 1 and 2. A similar trend in skill improvement is observed in the preprocessed precipitation forecasts. The preprocessed biweekly accumulated precipitation forecasts show ~2 weeks of skill gain (Figure 1c), resulting the skillful biweekly accumulations up to ~5 weeks.
Figure 5-1. CRPSS (relative to sampled climatology) of the CFSv2 raw and preprocessed ensemble (a)-(b) near-surface temperature and (c)-(d) precipitation forecasts versus the biweekly and monthly forecast lead times. The 90% confidence intervals, determined using the block bootstrap technique, are shown with shaded colors.
5.3.3 Skill of the S2S streamflow predictions

By using the statistically preprocessed CSFv2 precipitation and near-surface temperature forecasts as forcing to the distributed hydrological model HL-RDHM, we generate S2S raw ensemble streamflow forecasts. The raw forecasts are then statistically postprocessed to remove any systematic biases using the QR model. The CRPSS relative to sampled climatology is used to assess the skill of the raw and postprocessed ensemble streamflow forecasts (Figure 2). The CRPSS is computed for all the basins using the biweekly accumulated streamflow forecasts. Overall, the results show that the CFSv2 outputs are able to produce skillful ensemble streamflow forecasts at S2S timescales (Figure 2). The skill performance, however, of the forecasts varies with the forecast lead time and basin size. The basins with area less than 2000 km² are considered small (BPRV2, GOHV2 and PYAV2), whereas large basins are those with area greater than 7500 km² (JBIV2, PYAV2, and RMDV2). For small basins (Figures 2a-c), the raw biweekly ensemble streamflow forecasts show positive skill up to ~3 (BPRV2 and GOHV2) and 4 weeks (PYAV2), whereas large basins (Figures 2d-f) remain skillful up to ~4 (JBIV2) and 6 weeks (CARV2). The lower skill in the small basins could be due to the reduced ability of the CFSv2 outputs to capture the spatial variability of the climate variables in these basins, in particular precipitation, given the CFSv2 coarse spatial resolution (~100 x 100 km²).

After statistically postprocessing the S2S ensemble streamflow forecasts, the forecast skill improves relative to the raw forecasts across lead times (Figure 2). In general, the relative improvements in skill are greater at the later forecast lead times. The relative skill improvements can be as low as ~5% (CARV2 and RMDV2 at week 2, Figures 2e-f) to as high as ~30% (RMDV2 at week 6, Figure 2f). The skill gain achieved by postprocessing with respect to the raw forecasts is ~2 weeks. The CRPSS is also calculated (not shown) for monthly accumulations.
Figure 5-2. CRPSS (relative to sampled climatology) of the raw and postprocessed ensemble streamflow forecasts versus the biweekly forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. The 90% confidence intervals, determined using the block bootstrap technique, are shown with shaded colors.
Both the raw and postprocessed monthly accumulations generally demonstrate higher skill than the respective biweekly accumulations. The higher skill in the monthly accumulated flows is likely due, in part, to temporal averaging and the accumulation of skill from earlier periods.

### 5.3.4 Skill of the S2S nutrient and suspended sediment load predictions

Using the raw, monthly ensemble streamflow forecasts as the predictor in the QR model, we generate forecasts of monthly nutrients and sediment loads for the selected basins. The quality of the TN, TP, and TSS forecasts is assessed using both the correlation coefficient (Figures 3a-c) and CRPSS (relative to sampled climatology) (Figures 3d-f). Both metrics are computed for monthly loads and for lead times of 1 to 3 months. Overall, the CFSv2-based forecasts result in skillful nutrient and sediments load predictions at lead times of 1 to 3 months (Figure 3). In terms of both metrics (correlation coefficient and CRPSS), it is apparent that large basins (RMDV2, CARV2 and JBIV2) remain skillful at all the lead times for the TN, TP and TSS forecasts, while small basins (PYAV2, GOHV2 and BPRV2) tend only to have positive skill up to month 2. In general, TN demonstrates better forecasting ability than both TP and TSS, with TP outperforming TSS. Although these trends are apparent in all the basins and lead times, they tend to be emphasized at the later lead time of 3 months.

Compared to the skill of the streamflow forecasts, the TN, TP and TSS load forecasts show similar skill for the month 1 lead time. Beyond month 1, however, some of the load forecasts have greater CRPSS values than the streamflow forecasts. This can be explained by the nature of the empirical power law relationship between the nutrient concentration $C$ and the streamflow $Q$ (Zhang et al 2016), i.e., $C = \alpha Q^\beta$ where $\alpha$ and $\beta$ are coefficients. For example, higher
Figure 5-3. Correlation coefficient and CRPSS (relative to sampled climatology) of the (a)-(d) total nitrogen, (b)-(e) total phosphorous, and (c)-(f) total suspended sediment load forecasts versus the monthly forecast lead time for all the selected locations.
predictability of TN at RMDV2 and CARV2 is associated with their chemostatic behavior. These basins show the highest CRPSS values at month 3 among all the basins and water quality variables. Under chemostatic conditions, the power law relationship is linear with $C$ collapsing to the value of the coefficient $\alpha$ as the exponent $\beta$ goes to zero. This linear relationship between flow and load allows the TN predictions at CARV2 and RMDV2 to stay skillful at month 3. In the case of GOHV2, despite TN being chemostatic, the short observational record, which results in limited data for training the QR model, as well as the low skill of the streamflow forecasts make the TN forecast skill in this basin worse than sampled climatology at the month 3 lead time.

5.3.5 Effect of spatial scale, seasonality, and forecast threshold on the skill of the water quantity and quality predictions

We examine the effect of spatial scale (basin size), seasonality (warm and cool seasons), and forecast threshold (low versus high forecast values) on the skill of the streamflow, TN, TP, and TSS predictions. To assess the effect of basin size and seasonality on the forecast skill, the CRPSS values for the monthly streamflow, TN, TP and TSS forecasts at the month 1 lead time are plotted against the drainage area for the entire year (Figure 4a), cool season (Figure 4b), and warm season (Figure 4c). The cool season is defined as covering the months of October-March and the warm season as April-September. We find that monthly streamflow forecast skill increases with the basin size (Figure 4), and the linear trend is significant for both seasons ($p$-value<0.05) (Figure 4). The skill of the streamflow forecasts is appreciably higher during the
Figure 5-4. CRPSS (relative to sampled climatology) of the streamflow, total nitrogen load, total phosphorous load, and total suspended sediment load forecasts at the month 1 lead time versus the basin area for the (a) entire year, (b) cool season, and (c) warm season.
cool season (Figure 4b) than the warm one (Figure 4c). This is mostly due to the CFSv2-precipitation forecast skill being higher in the cool season (Peng et al., 2013; Tian et al., 2017). The lower CFSv2 skill in the warm season is partly due to challenges with the GCM in predicting convective precipitation and tropical cyclones (Moore et al. 2015). Both the nutrient and sediment load forecasts also show increasing skill with the basin size. However, this trend is only significant for the warm season TN and cool season TP. The skill of the nutrient and sediment load forecasts is also better in the cool season than in the warm season. This is mainly due to the higher skill of the cool season streamflow forecasts.

To understand the quality of the streamflow forecasts during low and high flow conditions, we employ the BSS (Figure 5). The low forecast category represents flows with nonexceedance probability, Pr, less than 0.05 (Pr<0.05) while the high flow category is for Pr>0.95. We find that the BSS of the S2S streamflow forecasts is greater for high flows than low flows (Figure 5), with high flows remaining skillful up to lead times of ~4 to 5 weeks while the skill of low flows is only ~2 to 3 weeks (Figure 5). There are two reasons for this. First, since the high flows result from the direct response of the basin to the precipitation events and the low flows are mostly dominated by subsurface processes, any skill in the precipitation forecasts propagates through the hydrological model directly to the high flows. Second, the performance of the hydrological model in simulation mode across the different basins is better for high flows than low flows, with an average NSE of 0.76 and 0.55 for high and low flows, respectively.
Figure 5-5. BSS (relative to sampled climatology) of the raw ensemble streamflow forecasts versus the biweekly forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. Results are shown for both low flows (flows with nonexceedance probability, Pr, less than 0.05) and high flows (flows with exceedance probability less than 0.05 or Pr>0.95).
The behavior of the water quality forecasts during low and high load conditions (Figure 6) mimics that of the streamflow forecasts (Figure 5). For example, the $BSS$ values for TN, TP, and TSS (Figure 6) are comparable to those for streamflow at the month 1 lead time (Figure 5), and the $BSS$ values are higher during high load conditions than low conditions (Figure 6). The higher skill for the high load forecasts is due, in part, to the higher skill of high flows. We also assess the effect of land cover on prediction skill. However, since the percent land cover distribution in the basins is relatively similar (Table 1), we do not find any marked trends between the skill of the water quantity/quality predictions and land cover, perhaps with the exception of PYAV2. This is the most anthropogenically disturbed basin out of the 6 selected basins in that it has the lowest percent of forest land cover, ~65%, and the largest percent of developed and pasture land cover, ~32%. PYAV2 is the only basin that has negative $CRPSS$ skill at the month 3 lead time for TP and TSS (Figure 3e-f). This could in part be due to the increased anthropogenic influence in this basin. For example, it seems possible that human decisions (e.g., stormwater management), not considered in the hydrological model, could be influencing TP and TSS budgets.
Figure 5-6. BSS (relative to sampled climatology) of the total nitrogen (TN), total phosphorous (TP) and total suspended sediment (TSS) load forecasts versus the monthly forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. Results are shown for both low loads (loads with nonexceedance probability, Pr, less than 0.05) and high loads (loads with exceedance probability less than 0.05 or Pr>0.95).
5.3.6 Uncertainty of the streamflow, nutrient load, and suspended sediment load predictions

To quantify the relative importance of different sources of uncertainties (climate and hydrological) in the S2S streamflow forecasts, we employ the CRPSS relative to both the observed and simulated values (Figure 7). When the S2S streamflow forecasts are verified relative to the observed values, the CRPSS accounts for the effect on forecast skill of both climate and hydrological uncertainties, whereas the forecasts verified relative to simulated values mainly account for climate uncertainties. The difference between the two, i.e., the CRPSS relative to observed values minus the CRPSS relative to simulated ones, provides an estimate of the effect of hydrological uncertainties on the streamflow forecast skill. We find that, regardless of the basin size, hydrological uncertainties have the strongest influence on streamflow forecast skill at the initial lead time (Shukla and Lettenmaier 2011). For instance, for BPRV2 at the month 1 lead time (Figure 7a), the CRPSS values relative to the simulated and observed flows are ~0.30 and ~0.10, respectively, suggesting a reduction of ~65% skill due to hydrological uncertainties. As the lead time grows, hydrological uncertainty become less pronounced and climate uncertainty starts to dominate the streamflow forecast skill. This is demonstrated in Figure 7 by the close proximity between the two CRPSS values (observation and simulation) beyond the 6 weeks lead time.

The CRPSS is also used to quantify the impact of different sources of uncertainties (hydrological, climate and statistical) on the nutrient and sediment load forecasts (Figure 8). In this case, the CRPSS verification relative to observed values (solid line) accounts for the statistical model uncertainties along with the climate and hydrologic uncertainties, while the
Figure 5-7. CRPSS (relative to both sampled climatological flow observations and simulations) of the raw ensemble streamflow forecasts versus the biweekly forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2. The 90% confidence intervals, determined using the block bootstrap technique, are shown with shaded colors.
Figure 5-8. CRPSS (relative to both sampled climatological load observations and simulations) of total nitrogen (TN), total phosphorous (TP) and total suspended sediment (TSS) load forecasts versus forecast lead time for the six selected locations in the JRB: (a) BPRV2, (b) GOHV2, (c) PYAV2, (d) JBIV2, (e) CARV2, and (f) RMDV2.
verification relative to simulated values mainly accounts for climate uncertainties. The difference between the two provides an estimate of both the hydrological and statistical model uncertainties. We find that at the initial lead time (month 1) the major source of uncertainty is hydrological and statistical (Figure 8). As the lead time grows, the hydrological and statistical uncertainty become less pronounced compared to the climate uncertainty. In contrast to Figure 7, the solid and dashed lines do not tend to converge towards each other with increasing lead time. This suggests that there is potential for improvements in water quality modeling to enhance the skill of nutrient and sediment load forecasts even at the later lead times.

5.4 Discussion and conclusions

In this study, we build and implement a new dynamical-statistical ensemble prediction system to generate S2S water quantity and quality forecasts. The system consists of a logistic regression statistical preprocessor, distributed hydrologic model and a quantile regression statistical postprocessor. The quantile regression model is used both to remove any systematic biases in the raw ensemble streamflow forecasts and to generate the nutrient and sediment load forecasts. The system is implemented for the years 2002-2016 along a major tributary of the Chesapeake Bay. Overall, we find that the dynamical CFSv2-based forecasts, when combined with the quantile regression model, can generate skillful flow, nutrient load, and suspended sediment load forecasts at lead times of 1 to 3 months.

Although this finding is demonstrated for the JRB, the results can be generalized to other places where the quality of the CFSv2 forecasts is similar to that found for the JRB. Indeed, the S2S near-surface temperature and precipitation forecast skill values obtained for the JRB are comparable to values reported for many other locations across the US (Chen et al 2013, Peng et
The CFSv2 near-surface temperature and precipitation forecast skills do not tend to vary markedly across the US (Chen et al. 2013, Peng et al. 2013, Saha et al. 2014, Tian et al. 2017), with some exceptions. For example, near-surface temperature skill is lower in parts of the US Northwest and South regions, whereas precipitation demonstrates minimum skill in parts of the US Rockies and Midwest region (Peng et al. 2013, Tian et al. 2017). Aside from these regions of minimal CFSv2 skill, the skill values of the CFSv2-based water quantity and quality forecasts for the JRB would be representative of potential skill values in many other locations in the US. This, together with the fact that CFSv2 is a global forecasting system, makes our results relevant to other basins.

We find that the flow forecasts can be skillful up to 7 weeks for large basins, which goes beyond the skill limit observed for the CFSv2 precipitation outputs (~5 weeks). This may be due to the propagation of skill from previous weeks through the serial dependency of flows. Nonetheless, the quality of the CFSv2 forecasts severely limits hydroclimatic predictions beyond 2 months. To further extend the lead time of streamflow predictions, it seems necessary to improve the spatial resolution of the CFSv2 forecasts. This would allow to resolve finer scale atmospheric phenomena and, as consequence, potentially improve precipitation and streamflow predictions. This is particularly relevant for small basins. Our results show that the skill-drainage area relationship is strong for streamflow, indicating a reduction in skill for small basins. This is also the case for TN and TP. Having more spatially resolved climate forcing could contribute to enhancing hydroclimatic skill across spatial scales. This could also improve the warm season forecasts. There is a noticeable difference in the performance of the water quantity and quality forecasts during the cool and warm season. The forecasts are more skillful in the cool season. This is due to the higher skill of the CFSv2 forcing in this season. Thus, improving the warm
season CFSv2 forecasts, in the future, will enhance the performance of the dynamical-statistical approach.

There is also a marked difference in the skill of high and low flows/loads. The higher skill for the high loads is due to the higher skill of the high flow forecasts. These results are relevant to water quality management as monthly high flows have been associated with increasing eutrophication and hypoxia in coastal waters (Kaushal et al 2008). It also identifies an area where improvements to the hydrological model could be made. The model in simulation mode performs better for high flows than low flows. Nonetheless, the BSS for the high flows declines with increasing lead time and becomes comparable to the BSS of low flows beyond month 1. This relatively rapid decline in skill suggests that initial land surface states (e.g., soil moisture) can have a strong effect on the forecast skill. Improving the representation of these initial states, in the future, could likely serve to prolong the forecasts’ lead times.

We believe the findings of this study are relevant to water quality management applications. One area of application is in the prevention and mitigation of harmful algal blooms. For example, in the Great Lakes, high precipitation events, coupled with agricultural practices and land use change, have increased nutrient loading into Lake Erie, which has led to the episodic occurrence of extensive harmful algal blooms over recent years (Michalak et al 2013). Our ability to predict nutrient loads 3 months ahead could support building early warning systems for harmful algal blooms and help make better informed decisions for nutrient-related management plans and actions. Since water quantity and nutrient loads are effective predictors of bloom size across different geographic regions (Stumpf et al 2012), our dynamic-statistical approach should be readily transferable to other freshwater systems and coastal locations, such as the Gulf of
Mexico and Great Lakes, where the prediction of bloom occurrence, extent, and timing is crucial to implementing realistic bloom mitigation strategies.

Our findings have implications for operationalizing S2S water quality predictions in the US and potentially other places. By building and testing a modular dynamical-statistical ensemble prediction system, we demonstrate a feasible approach to the generation of water quality forecasts. The approach is more cost effective and computationally efficient than fully implementing a process-based water quality model. This makes it readily implementable in an operational forecasting setting. For instance, an existing operational seasonal streamflow forecast product could be used with the quantile regression model to generate nutrient and suspended sediment load forecasts. This study also identifies relevant timescales and sources of uncertainties that could assist in the future when trying to implement process-based water quality models operationally. In addition, the outputs from the dynamical-statistical ensemble prediction system could be used as boundary conditions (e.g., riverine inflows and associated nutrients and sediments concentrations) to operational marine hydrodynamic and biogeochemical forecasting models. Ultimately, this study shows that it is possible to use current operational streamflow forecast systems, e.g., the National Water Model (Salas et al 2018), to generate skillful water quality predictions several months in advance. This would have the benefit of making predictions available to the public, providing support to a wide range of water quality management strategies and creating new opportunities for yet unknown applications.

5.5 Chapter 5 references


Bekele E G and Nicklow J W 2005 Multiobjective management of ecosystem services by integrative watershed modeling and evolutionary algorithms. *Water Resources Research* 41 W10406


Cloke H and Pappenberger F 2016 Ensemble flood forecasting: A review *Journal of Hydrology* **375** 613-626

Crochemore L, Ramos M H and Pappenberger F 2016 Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts *Hydrology and Earth System Sciences* **20** 3601-3618


Demargne J *et al* 2014 The science of NOAA's operational hydrologic ensemble forecast service *Bulletin of the American Meteorological Society* **95** 79-98

Dirmeyer P A 2013 Characteristics of the water cycle and land–atmosphere interactions from a comprehensive reforecast and reanalysis data set: CFSv2 *Climate dynamics* **41** 1083-1097


Easton Z M *et al* 2017 Short-term forecasting tools for agricultural nutrient management. *J. Environ. Qual.* **46** 1257–1269

Han H, Allan J D and Scavia D 2009 Influence of climate and human activities on the relationship between watershed nitrogen input and river export *Environmental Science & Technology* **43** 1916-1922

Hirsch R M, Moyer D L and Archfield S A 2010 Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs *JAWRA Journal of the American Water Resources Association* **46** 857-880


Howarth R *et al* 2012 Nitrogen fluxes from the landscape are controlled by net anthropogenic nitrogen inputs and by climate *Frontiers in Ecology and the Environment* **10** 37-43


Kaushal S S *et al* 2008 Interaction between urbanization and climate variability amplifies watershed nitrate export in Maryland *Environmental Science & Technology* **42** 5872-5878


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

Lee D, Ward P J and Block P 2018 Identification of symmetric and asymmetric responses in seasonal streamflow globally to ENSO phase *Environmental Research Letters* **13** 044031

Li H, Luo L, Wood E F and Schaake J 2009 The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting *Journal of Geophysical Research: Atmospheres* **114** D04114

Lucatero D, Madsen H, Refsgaard J C, Kidmose J and Jensen K H 2018 Seasonal streamflow forecasts in the Ahlergaard catchment, Denmark: the effect of preprocessing and post-processing on skill and statistical consistency *Hydrology and Earth System Sciences* **22** 3601-3617


Moyer D 2016 Nitrogen, phosphorus, and suspended-sediment loads and trends measured in nine Chesapeake Bay tributaries: Water years 1985–2015 *US Geological Survey data release* http://dx doi org/105066/F7Q81B5N

National Academies of Sciences, Engineering, and Medicine 2018 Review of the New York City Department of Environmental Protection Operations Support Tool for Water Supply


Peng P, Barnston A G and Kumar A 2013 A comparison of skill between two versions of the NCEP climate forecast system (CFS) and CPC’s operational short-lead seasonal outlooks *Weather Forecast* 28 445–462


Pyke C R et al 2008 Climate Change and the Chesapeake Bay: State-of-the-Science Review and Recommendations. *A Report from the Chesapeake Bay Program Science and Technical Advisory Committee (STAC), Annapolis, MD* 59 pp

Preston S D and Brakebill J W 1999 Application of spatially referenced regression modeling for the evaluation of total nitrogen loading in the Chesapeake Bay watershed, USGS.

Raymond P A, David M B and Saiers J E 2012 The impact of fertilization and hydrology on nitrate fluxes from Mississippi watersheds *Current Opinion in Environmental Sustainability* 4 212-218
Runkel R L, Crawford C G, and Cohn T A 2004 Load Estimator (LOADEST): A FORTRAN program for estimating constituent loads in streams and rivers 2328-7055


Salas F R et al 2018 Towards Real-Time Continental Scale Streamflow Simulation in Continuous and Discrete Space JAWRA Journal of the American Water Resources Association 54 7-27


Sharma S et al 2017 Eastern U.S. verification of ensemble precipitation forecasts Wea. Forecasting 32 117–139

Sharma S, Siddique R, Reed S, Ahnert P, Mendoza P and Mejia A 2018 Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system Hydrology and Earth System Sciences 22 1831-1849


Shukla S and Lettenmaier D 2011 Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill Hydrology and Earth System Sciences 15 3529-3538
Sinha E and Michalak A M 2016 Precipitation dominates interannual variability of riverine nitrogen loading across the continental United States *Environmental science & technology* 50 12874-12884

Sinha E, Michalak A M and Balaji V 2017 Eutrophication will increase during the 21st century as a result of precipitation changes *Science* 357 405–408

Slater L J and Villarini G 2018 Enhancing the predictability of seasonal streamflow with a statistical-dynamical approach *Geophysical Research Letters* 45 6504–6513


Taka M, Aalto J, Virkanen J and Luoto M 2016 The direct and indirect effects of watershed land use and soil type on stream water metal concentrations *Water Resources Research* 52 7711-7725

Tian D, Wood E F and Yuan X 2017 CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States *Hydrology and Earth System Sciences* 21 1477-1490


Yang X, Liu Q, Luo X and Zheng Z 2017 Spatial regression and prediction of water quality in a watershed with complex pollution sources *Scientific Reports* 7 8318


Yuan X, Wood E F, Luo L and Pan M 2011 A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction *Geophysical research letters* **38** L13402


Yuan X, Wood E F, Roundy J K and Pan M 2013 CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States *Journal of Climate* **26** 4828-4847


Chapter 6
Conclusions and recommendations

6.1 Summary and conclusions

This dissertation aims at developing and advancing a robust, regional hydrologic ensemble prediction system (RHEPS) in order to comprehensively study, understand and quantify the implications of ensemble weather/climate variables and other sources of uncertainty on hydrological predictions. The goal with the RHEPS is to enhance the skill and reliability of hydrological forecasts by facilitating the integration and rigorous verification of new system components, enhanced physical parameterizations, and distributed observations. The RHEPS is comprised by the following system components: ensemble weather/climate forecasts, hydrological/land-surface models, statistical weather preprocessor, statistical streamflow postprocessor, and verification strategy. The RHEPS is implemented from small (~10^2 km^2) to regional basin scales (~10^4 km^2), and from short-range (~1-3 days) to seasonal (~3 months) temporal scales within the US Middle Atlantic region. Overall, results from RHEPS show that implementing both preprocessing and postprocessing ensures the most skill improvements in streamflow forecasting, but postprocessing alone can often be a competitive alternative. In addition, RHEPS demonstrate improved skill of multimodel forecasts relative to single model ones, and the observed skill improvement is found to dominated by model diversity rather than by increased ensemble size alone. RHEPS also demonstrate the ability of climate forecasts to support water quality predictions at S2S timescales. Specific conclusions associated with each objective of this dissertation are provided next.
In objective 1 (O1), ensemble precipitation forecasts across the eastern United States were verified, specifically, version 2 of the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System Reforecast (GEFSRv2) and Short Range Ensemble Forecast (SREF) system, as well as NCEP’s Weather Prediction Center probabilistic quantitative precipitation forecast (WPC-PQPF) guidance. The forecasts were verified using multisensor precipitation estimates and various metrics conditioned upon seasonality, precipitation threshold, lead time, and spatial aggregation scale. The verification results indicate that, across the eastern United States, precipitation forecast bias decreases and the skill and reliability improve as the spatial aggregation scale increases; however, all the forecasts exhibit some underforecasting bias. The skill of the forecasts is appreciably better in the cool season than in the warm one. The WPC-PQPFs tend to be superior, in terms of the correlation coefficient, relative mean error, reliability, and forecast skill scores, than both GEFSRv2 and SREF, but the performance varies with the River forecast center and lead time. Based on GEFSRv2, medium-range precipitation forecasts tend to have skill up to approximately day 7 relative to sampled climatology.

In objective 2 (O2), RHEPS was implemented to investigate 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). To implement the RHEPS, 1 to 7 days weather forecasts from the GEFSRv2 was used to force NOAA's Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM) and generate raw ensemble streamflow forecasts. Streamflow ensembles were generated with the RHEPS for four different forecasting scenarios involving no statistical processing, preprocessing alone, postprocessing alone, and both preprocessing and postprocessing. The results show that the scenarios that implement preprocessing and postprocessing separately tend to perform similarly in terms of forecast skill,
although the postprocessing-alone scenario is often more effective. The scenario involving both preprocessing and postprocessing consistently outperforms the other scenarios. In some cases, however, the differences between this scenario and the scenario with postprocessing alone are not as significant. It is concluded that implementing both preprocessing and postprocessing ensures the most skill improvements, but postprocessing alone can often be a competitive alternative.

In objective 3 (O3), RHEPS was used to investigate if the skill of hydrologic multimodel forecast is significantly larger than that of a single model forecast, and examine if the observed improvements are dominated by model diversity or increasing ensemble size. To generate the multimodel ensembles, a new statistical postprocessor, namely quantile regression-Bayesian model averaging (QR-BMA) was developed and implemented. 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, an information-theoretic measure, namely conditional mutual information, was used to quantify the skill enhancements from the multimodel forecasts. Ensemble streamflow forecasts at lead times from 1 to 7 days are generated using three hydrological models: Antecedent Precipitation Index (API)-Continuous, Hydrology Laboratory-Research Distributed Hydrologic Model, and Weather Research and Forecasting Hydrological modeling system. Weather ensemble forecasts from the National Centers for Environmental Prediction 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2) are used as forcing to the hydrological models. The result shows that after bias-correcting the streamflow forecasts from each model their skill performance becomes comparable. 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. The results highlight benefits of hydrological multimodel forecasting for improving streamflow predictions.

In objective 4 (O4), RHEPS was used to generate subseasonal to seasonal (S2S) water quantity (streamflow) and quality (nutrient and sediment loads) forecasts. For this purpose, a new dynamical-statistical approach was developed and implemented. The approach combines dynamical CFSv2-based flow predictions with a statistical model to generate nutrient and sediment loads at S2S timescales. By implementing the system along a major tributary of the Chesapeake Bay, it is concluded that the CFSv2-based streamflow predictions when combined with the quantile regression model can generate skillful flow, nutrient load, and sediment load forecasts at lead times of 1 to 3 months. The forecast skill is appreciably higher in the large basins over the small ones, high flows/loads over the low flows/loads, and cool season over the warm season. Overall, the forecast skill is dominated by hydrological uncertainty at the initial lead times, whereas climate uncertainties become more dominant as lead time grows. Through the dynamical-statistical approach, the system comprises a cost and time effective solution to operationalize S2S water quality prediction.

6.2 Recommendations for future work

There are scopes of future work related to every objective of this dissertation. Objective 1 (O1) verifies the quality of the ensemble precipitation forecasts from GEFSRv2, SREF, and the WPC-PQPFs. Although verification results provided useful diagnostic information regarding the quality of the ensemble precipitation forecasts, they did not provide direct information on how to
improve the underlying numerical weather prediction models. To better understand the physical and environmental conditions associated with forecast errors and skill, more weather variables than just precipitation could be considered. Additionally, the verification could be focused on a few high-impact events or unusual weather scenarios.

Objective 2 (O2) quantify the contributions of statistical processing to the RHEPS. Future research, however, could be focused on specific flood events to understand how distinct hydrological processes contribute or constrain forecast quality. 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. 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.

From objective 3 (O3), it is demonstrated that the skill enhancement obtained by hydrological multimodel forecast is dominated by model diversity, rather than by increased ensemble size alone. Although this is the case here for forecast skill, future research could examine whether these results apply to other attributes of forecast quality. In particular, metrics that are more responsive to the ensemble size than the adopted CMI formalism, which was based on the ensemble mean, could be tried.

Objective 4 (O4) shows the ability of S2S hydroclimatic predictions to generate skillful water quantity and quality predictions using a dynamical-statistical approach. To further improve the forecast skill at S2S timescales, future research could be focused on improving spatial
resolution of the CFSv2 forecasts to resolve finer scale atmospheric phenomena, and also in the representation of initial land surface states (e.g., soil moisture) in hydrologic modeling.
Appendix

Verification metrics

**Modified correlation coefficient (R_m)**

The modified version of the correlation coefficient, called as modified correlation coefficient $R_m$, 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:

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

where $\sigma_s$ and $\sigma_q$ denote the standard deviation of the simulated and observed flows, respectively.

**Percent bias (PB)**

The percent bias (PB) measures the average tendency of the predicted values to be larger or smaller than the reference values. The PB is given by

$$\text{PB} = \frac{\sum_{i=1}^{N} S_i - O_i}{\sum_{i=1}^{N} O_i} \times 100,$$

where $S_i$ and $O_i$ 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 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:
\[
\text{NSE} = 1 - \frac{\sum_{i=1}^{N} (s_i - o_i)^2}{\sum_{i=1}^{N} (s_i - \bar{o}_i)^2},
\]
where \( s_i \), \( o_i \), and \( \bar{o}_i \) are the simulated, observed, and mean observed flow, respectively, at time \( i \).

**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
\[
\text{RME} = \frac{\sum_{i=1}^{n} (\bar{X}_i - Y_i)}{\sum_{i=1}^{n} Y_i},
\]
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 \( k \) 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.

**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

$$\text{BS} = \frac{1}{n} \sum_{i=1}^{n} \left[ F_{x_i}(q) - F_{y_i}(q) \right]^2,$$

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

$$F_{x_i}(q) = P[X_i > q],$$

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

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

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):

$$\text{BSS} = 1 - \frac{\text{BS}_{\text{main}}}{\text{BS}_{\text{reference}}},$$

where $\text{BS}_{\text{main}}$ and $\text{BS}_{\text{reference}}$ 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.

**Reliability diagram**

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

$$
\bar{F}_{X_i}(q) = \frac{1}{|I_k|} \sum_{i \in I_k} F_{X_i}(q), \text{ where } I_k = \{ i: X_i \in B_k \}, \quad (A9)
$$

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

$$
\bar{F}_{Y_i}(q) = \frac{1}{|I_k|} \sum_{i \in I_k} F_{Y_i}(q), \text{ where } F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise}. \end{cases} \quad (A10)
$$

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

**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

$$
\text{CRPS} = \int_{-\infty}^{\infty} \left[ F_x(q) - F_y(q) \right]^2 dq. \quad (A11)
$$

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:

$$
\text{CRPSS} = 1 - \frac{\text{CRPS}_{\text{main}}}{\text{CRPS}_{\text{reference}}}, \quad (A12)
$$
where CRPS is averaged across \( n \) pairs of forecasts and observations to calculate the mean CRPS of the main forecast system (\( \text{CRPS}_{\text{main}} \)) and reference forecast system (\( \text{CRPS}_{\text{reference}} \)). The CRPSS ranges from \(-\infty\) 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 \( \text{CRPS}_{\text{main}} \) into different components. Specifically, we consider the CRPS reliability (\( \text{CRPS}_{\text{rel}} \)) and potential (\( \text{CRPS}_{\text{pot}} \)) such that

\[
\text{CRPS}_{\text{main}} = \text{CRPS}_{\text{rel}} + \text{CRPS}_{\text{pot}}. \tag{A13}
\]

**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

\[
\text{PoD} = \frac{\sum_{i=1}^{n} I_{X_i}(F_{X_i}(q) > d | Y_i > q)}{\sum_{i=1}^{n} I_{Y_i}(Y_i > q)}, \tag{A14}
\]

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

\[
\text{PoFD} = \frac{\sum_{i=1}^{n} I_{X_i}(F_{X_i}(q) > d | Y_i \leq q)}{\sum_{i=1}^{n} I_{Y_i}(Y_i \leq q)}. \tag{A15}
\]

The relationship between PoD and PoFD is assumed bivariate normal such that
PoD = \phi[a + b\phi^{-1}(PoFD)], \quad (A16)

where \( a = \frac{\mu_{PoD} - \mu_{PoFD}}{\sigma_{PoD}} \), \( b = \frac{\sigma_{PoFD}}{\sigma_{PoD}} \), and \( \phi \) is the cdf of the standard normal distribution. \( \mu_{PoD} \) and \( \mu_{PoFD} \) are the means while \( \sigma_{PoD} \) and \( \sigma_{PoFD} \) 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.

**Implementation of the Expectation Maximation Algorithm**

We describe here the steps followed to implement the EM algorithm. The description uses the variables and notation previously defined in Subsection 4.2.1. To implement the EM algorithm, the latent variable \( z_{ki}^i \) 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

\[
\begin{align*}
  w_{k,iter-1}^i &= \frac{1}{K}, \quad \text{and} \\
  \sigma_{k,iter-1}^2 &= \frac{1}{K} \sum_{i=1}^{T} \frac{\sum_{k=1}^{K} \left( \Delta_{NQT}^{t,i} - f_{k,NQT}^{t,i} \right)^2}{T},
\end{align*}
\]

allowing the calculation of an initial log-likelihood

\[
l(\theta_{iter-1}) = \sum_{i=1}^{T} \log \left( \sum_{k=1}^{K} w_{k,iter-1}^i g(\Delta_{NQT}^{t,i} | f_{k,NQT}^{t,i}, \sigma_{k,iter-1}^{2,i}) \right),
\]

\[
(A19)
\]

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 \( z_{k}^{i,j} \) for each time step is estimated given the initial values of the weight and variance as

\[
\hat{z}_{k,\text{iter}}^{i,j} = \frac{w_{k,\text{iter}-1}^{i} g \left( \Delta_{\text{NQT}}^{i,j} \mid f_{k,\text{NQT}}^{i,j}, \sigma_{k,\text{iter}-1}^{z} \right)}{\sum_{k=1}^{K} w_{k,\text{iter}-1}^{i} g \left( \Delta_{\text{NQT}}^{i,j} \mid f_{k,\text{NQT}}^{i,j}, \sigma_{k,\text{iter}-1}^{z} \right)}.
\]  

(A20)

In the subsequent maximization step, the values of the weight and variance are updated using the current estimate of \( z_{k,\text{iter}}^{i,j} \) as follows

\[
\begin{align*}
\hat{w}_{k,\text{iter}}^{i} &= \frac{1}{T} \sum_{t=1}^{T} \hat{z}_{k,\text{iter}}^{i,j}, \text{ and} \\
\hat{\sigma}_{k,\text{iter}}^{z} &= \frac{\sum_{t=1}^{T} \hat{z}_{k,\text{iter}}^{i,j} \left( \Delta_{\text{NQT}}^{i,j} - f_{k,\text{NQT}}^{i,j} \right)^{2}}{\sum_{t=1}^{T} \hat{z}_{k,\text{iter}}^{i,j}}. 
\end{align*}
\]  

(A21)

The log-likelihood function is then recomputed using the updated weight and variance as

\[
I(\theta_{\text{iter}}) = \sum_{t=1}^{T} \log \left( \sum_{k=1}^{K} w_{k,\text{iter}}^{i} g \left( \Delta_{\text{NQT}}^{i,j} \mid f_{k,\text{NQT}}^{i,j}, \sigma_{k,\text{iter}}^{z} \right) \right).
\]  

(A22)

The expectation and maximization steps are iterated until the improvement in the log-likelihood is no less than some pre-defined tolerance, i.e. \( \left| \left( I(\theta_{\text{iter}}) - I(\theta_{\text{iter}-1}) \right) \right| < tol \), in this case \( tol=10^{-6} \).
VITA
Sanjib Sharma

EDUCATION

**PhD in Hydrology and Hydrometeorology**
Department of Civil and Environmental Engineering
The Pennsylvania State University, USA
Summer 2019

**MS in Water Resources Engineering**
Department of Civil and Environmental Engineering
Southern Illinois University, USA
2015

**BE in Water Resources Engineering**
Institute of Engineering
Tribhuvan University, Nepal
2011

PUBLICATIONS


PROFESSIONAL MEMBERSHIPS

- American Geophysical Union (AGU)
- ASCE EWRI Risk, Uncertainty and Probabilistic Analysis (RUPA) committee
- Quarterly Franklin Membership
- Engineer in Training (EIT), Michigan Board, USA
- Nepal Engineering Association (NÉA)