UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A FRAMEWORK OF SUBSEASONAL-TO-SEASONAL (S2S) ENSEMBLE HYDROLOGICAL FORECASTING AT A WATERSHED SCALE USING DYNAMICAL PRECIPITATION FORECAST: FORECAST VERIFICATION, ADAPTATION, AND STREAMFLOW PREDICTION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

Doctor of Philosophy

By

LUJUN ZHANG
Norman, Oklahoma
2023

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/340079

A FRAMEWORK OF SUBSEASONAL-TO-SEASONAL (S2S) ENSEMBLE HYDROLOGICAL FORECASTING AT WATERSHED SCALE USING DYNAMICAL PRECIPITATION FORECASTS: FORECAST VERIFICATION, ADAPTATION, AND STREAMFLOW PREDICTION

A DISSERTATION APPROVED FOR THE
SCHOOL OF CIVIL ENGINEERING AND ENVIRONMENTAL SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Tiantian Yang, Chair

Dr. Yang Hong, Co-Chair

Dr. Amy McGovern

Dr. Pierre Kirstetter

Dr. Randall Kolar

© Copyright by LUJUN ZHANG 2023
All Rights Reserved.

Dedication

I dedicate this dissertation to my family, friends, and all those who have been part of my journey thus far. I aspire for this work to bring a small positive change to the world, making it slightly better than it was yesterday.

Acknowledgments

I deeply appreciate Dr. Tiantian Yang, my Ph.D. advisor, for his invaluable guidance throughout this journey. He generously shared his knowledge and expertise in helping me complete my dissertation. But more importantly, I realize that the global pandemic, which has impacted millions of lives, could have deprived me of the opportunity to pursue a Ph.D. if it weren't for the chance offered by him back in 2019. For that, I am grateful to him. Also, my gratitude goes to Dr. Soroosh Sorooshian and Dr. Chiyuan Miao for their encouraging words during my search for a Ph.D. opportunity, which motivated me going forward.

I extend my heartfelt thanks to my Ph.D. committee members: Drs. Yang Hong, Amy McGovern, Pierre Kirstetter, and Randall Kolar. Their valuable suggestions have made this achievement possible. I also want to acknowledge Dr. Taereem Kim and Dr. Shang Gao for their unwavering guidance during this journey. Their wisdom has been my compass, keeping me exploring.

My parents deserve a special mention for their endless support. I feel fortunate to have grown up in a family that not only provided a strong foundation but also taught me essential virtues like honesty and perseverance. Their unwavering belief in me has been my driving force. From my mom Lirong, I have learned the importance of kindness, compassion, and selflessness in everything I do. And from my dad Baorui, I've learned the incredible strength of persistence, dedication, and determination even when faced with challenges. They have not just shaped my education, but also have formed the very core of who I am today.

Last, but certainly not least, I express my sincere gratitude to the National Science Foundation EPSCoR-Track1 award (NSF-OIA-1946093) for providing the essential financial support that has made this dissertation possible.

Table of Contents

Chapter 1. Introduction 1
1.1 Background 1
1.2 Research Questions 6
References 9
Chapter 2. Evaluation of Subseasonal-to-Seasonal (S2S) Precipitation Forecast from the North American Multi-Model Ensemble Phase II (NMME-2) over the Contiguous U.S. 13
2.1 Introduction 14
2.2 Data and Study Regions 21
2.3 Methodology, Procedure, and Evaluation Metrics 24
2.3.1 Percentage Bias (PBIAS) 25
2.3.2 Anomaly Correlation Coefficient (ACC) 25
2.4 Results 26
2.4.1 Forecast Bias 26
2.4.2 Predictive Skill 32
2.5 Discussion 38
2.6 Conclusion 43
References 46
Chapter 3. An Alternative Ensemble Streamflow Prediction (ESP) Approach Using Improved Subseasonal-to-Seasonal (S2S) Precipitation Forecasts from the North America Multi-Model Ensemble Phase II (NMME-2) Dataset 52
3.1 Introduction 53
3.2 Datasets and Study Regions 57
3.2.1 Datasets 57
3.2.2 Target Watersheds 58
3.3 Methodology 60
3.3.1 Experiment Design 60
3.3.2 Evaluation Statistics 63
3.4 Results 64
3.4.1 Calibration of Sac-SMA 64
3.4.2 Forecast Bias and Predictive Skill of Precipitation 65
3.4.3 Predictive Skill of Streamflow 70
3.4.4 Examination of Hydrographs from Selected Extreme Events 74

3.5 Discussion 78
3.6 Conclusion 82
References 85
Chapter 4. Improving Subseasonal-to-seasonal Forecasts in Predicting the Occurrence of Extreme Precipitation Events over the Contiguous U.S. Using Machine Learning Models 88
4.1 Introduction 89
4.2 Data and Study Regions 95
4.3 Experiment Setting and Evaluation Metrics 97
4.3.1 Experiment Setting 97
4.3.1 Pixel-based and Event-based Forecast Evaluation Metrics 99
4.4 Results 105
4.4.1 Model's Sensitivity on Input Variables (Comparison of Scenarios E1-E6) 105
4.4.2 Model's Sensitivity on Hyperparameter (Comparison of Scenarios E1 and E6-E10) 113
4.5 Discussion 121
4.6 Conclusion 126
References 129
Chapter 5. Combining Machine Learning with a Distribution-based Technique to Adapt Subseasonal-to-Seasonal (S2S) Precipitation Forecasts at Watershed Scale for Hydrologic Ensemble Streamflow Prediction (ESP) 136
5.1 Introduction 137
5.2 Datasets and Study Regions 142
5.2.1 Datasets 142
5.2.2 Study Watersheds 143
5.3 Experiment Setting and Evaluation Metrics 144
5.3.1 General Experiment Setting 144
5.3.2 S2S Forecast Adaptation Schemes 147
5.3.3 Evaluation Statistics 150
5.4 Results 152
5.4.1 Calibration of Sac-SMA 152
5.4.2 Predictive Performance of Precipitation 153
5.4.3 Predictive Performance of Streamflow 157
5.4.4 The "Elasticity" of the Added Skill Between Streamflow and Precipitation 160

5.5 Discussion 161
5.6 Conclusion.....164References
..... 167
Chapter 6. Conclusions and Recommendations for Future Work 170
6.1 Conclusions 170
6.2 Recommendations for Future Work 174
References 178
Appendix A 179
References 181
Appendix B 182

List of Figures

Figure 2.1 Climate region partitioning across CONUS per NCEI..... 23
Figure 2.2 The overall pattern of PBIAS for NMME-2 precipitation hindcasts over the CONUS. 26
Figure 2.3 The pattern of PBIAS for NMME-2 precipitation hindcasts over the CONUS at different lead times (i.e., day 1 to 7: week 1, day 8 to 14: week 2, day 15 to 21: week 3, and day 22 to 28: week 4)..... 28
Figure 2.4 The PBIAS of NMME-2 precipitation hindcasts within different seasons (i.e., DJF, MAM, JJA, SON) across CONUS..... 29
Figure 2.5 Regional averaged PBIAS of NMME-2 precipitation hindcasts from the five NMME-2 models and an SMA model for each season (DJF, MAM, JJA, and SON) and lead time (weeks 1 to 4) at the nine NCEI climate regions..... 31
Figure 2.6 The pattern of ACC for NMME-2 precipitation hindcasts over the CONUS at different lead times (weeks 1 to 4)..... 33
Figure 2.7 The pattern of ACC for NMME-2 precipitation hindcasts within different seasons (i.e., DJF, MAM, JJA, SON) at different lead times (weeks 1 to 4) over the CONUS. 34
Figure 2.8 Regional averaged ACC of NMME-2 precipitation hindcasts from the five NMME-2 models and an SMA model for each season (DJF, MAM, JJA, and SON) and lead time (weeks 1 to 4) at the nine NCEI climate regions..... 36
Figure 3.1 Map of the 4 study watersheds (BC is in Texas; SC is in Arkansas; CR is in Louisiana; BP is in Missouri). 59
Figure 3.2 A schematic diagram of the experiment design. 60
Figure 3.3 Simulated and observed daily streamflow hydrographs at the study watersheds. 65
Figure 3.4 Cumulative distribution functions of the S2S precipitation before and after BCSD at four study watersheds. 66
Figure 3.5 QQ-plots of the mean areal S2S precipitation before and after BCSD. 68
Figure 3.6 Predictive skills of S2S precipitation and resampled precipitation over forecast lead times across four study watersheds..... 69
Figure 3.7 Predictive skills of streamflow with S2S precipitation and resampled precipitation over forecast lead times across four study watersheds. 71
Figure 3.8 Predictive performance on daily extreme values within different quantile segments, at different lead times, and across four study watersheds. Red boxes highlight overall better performance of S2S precipitation under different scenarios..... 73
Figure 3.9 Predicted and measured daily hydrograph in November 2002 at the BC watershed. . 75
Figure 3.10 Predicted and measured daily hydrograph in October 2009 at the SC watershed..... 76
Figure 3.11 Predicted and measured daily hydrograph in April 1999 at the CR watershed..... 77
Figure 3.12 Predicted and measured daily hydrograph in January 1990 at the BP watershed. 78
Figure 4.1 Climate region partitioning across CONUS per NCEI..... 97

Figure 4.2 An Illustrating figure of ROC and AUROC.. 104Figure 4.4 The pattern of EFAR from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4). 108

Figure 4.6 The pattern of BSS from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4). 110
Figure 4.7 Spatially averaged AUROC values at 9 NCEI climate regions from experiment scenarios E1-E6, over CONUS and at different lead times (week 1 to 4). 112
Figure 4.8 The pattern of EPOD from experiment scenarios E1, and E6-E10, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4). 114
Figure 4.9 The pattern of EFAR from experiment scenarios E1, and E6-10, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4). 115
Figure 4.10 The pattern of ECSI from experiment scenarios E1, and E6-E10, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4). 117
Figure 4.12 Spatially averaged AUROC values at 9 NCEI climate regions from experiment scenarios E1, and E6-E10, over CONUS and at different lead times (week 1 to 4)..... 120
Figure 5.1 Map of the 4 study watersheds (BC is in Texas; SC is in Arkansas; CR is in Louisiana; BP is in Missouri). 144
Figure 5.2 A schematic diagram of the experiment design. 146
Figure 5.3 Simulated and observed daily streamflow hydrographs at the study watersheds 153
Figure 5.4 QQ-plots of the mean areal S2S precipitation in raw condition, after BCSD post-processing, and after RF-BCSD post-processing..... 154
Figure 5.6 CRPS skill of the mean areal S2S precipitation after BCSD and RF-BCSD adaptation 4 study watersheds and over the 28-day forecast horizon. 157
Figure 5.8. CRPS skill of the streamflow hindcasts resulted from S2S precipitation after BCSD and RF-BCSD adaptation over the forecast horizon at 4 study watersheds. 159
Figure A.1 Conceptual structure of the Sac-SMA model modified from Kim et al. (2021). 180
Figure B.1 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Northwest NCEI climate region, at different lead times (weeks 1 to 4). 182
Figure B.2 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the West NCEI climate region, at different lead times (weeks 1 to 4). 182
Figure B.3 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the West North Central NCEI climate region, at different lead times (weeks 1 to 4). 183
Figure B.4 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Southwest NCEI climate regions, at different lead times (weeks 1 to 4). 183
Figure B.5 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the South NCEI climate region, at different lead times (weeks 1 to 4). 183
Figure B.6 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the East North Central NCEI climate region, at different lead times (weeks 1 to 4). 184

Figure B.7 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Central NCEI climate region, at different lead times (weeks 1 to 4). 184
Figure B.8 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Northeast Central NCEI climate region, at different lead times (weeks 1 to 4). 184
Figure B.9 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Southeast NCEI climate region, at different lead times (weeks 1 to 4). 185

List of Tables

Table 2.1 A list of selected NMME-2 models. 22
Table 3.1 Hydrometeorological datasets used in Chapter 3. 58
Table 3.2 Topography and hydrometeorology information of the study watersheds in Chapter 3. 59
Table 3.3 Contingency table of all possible outcomes for categorical forecasts of binary events... 64
Table 4.1 Designed experiment scenarios... 98
Table 4.2 Contingency table of all possible outcomes for categorical forecasts of binary events... 100
Table 5.1 Hydrometeorological datasets used in Chapter 5. 143
Table 5.2 Topography and hydrometeorology information of the study watersheds in Chapter 5. 144
Table A.1 13 tunable parameters of the Sac-SMA. 180

Abstract

Accurate and reliable streamflow forecasts, especially at a Subseasonal-to-Seasonal (S2S) timescale (spanning 10 to 30 days into the future) could greatly benefit various human socio-economic activities. The Ensemble Streamflow Prediction (ESP) framework is commonly applied and currently in operation for streamflow predictions at S2S timescale. However, ESP's reliance on the randomly resampled historical precipitation has limited its performance and compromised its reliability. As an alternative, the S2S precipitation forecasts derived from coupled general circulation models (GCMs) present an opportunity to overcome the limitations of randomly resampled precipitation in the ESP. Despite this potential, the application of available S2S precipitation forecasts in hydrology has not undergone a comprehensive assessment. Therefore, in this dissertation, multiple S2S precipitation forecast products from the North America Multi-Model Ensemble Phase II (NMME-2) have been collected and analyzed to validate their forecast performance over the contiguous United States (CONUS) as well as to test their hydrologic applicability at a watershed scale. The introductory chapter provides the context and outlines the research questions that motivate this dissertation (**Chapter 1**). A comprehensive evaluation of the performance of the raw S2S precipitation forecasts from NMME-2 over CONUS is conducted (**Chapter 2**). This dissertation further validates the applicability and superiority of S2S precipitation through a standard forecast adaptation technique under the ESP framework at four selected experimental watersheds (**Chapter 3**). In addressing the limitations of extreme precipitation event prediction, a novel Machine Learning (ML)-based post-processing technique is developed to improve the predictive skill of available S2S precipitation forecasts over CONUS (**Chapter 4**). This developed ML-based technique is subsequently applied to the four aforementioned watersheds, effectively adapting raw S2S precipitation for streamflow prediction within the ESP framework (**Chapter 5**). In conclusion, this dissertation underscores

its major findings and offers insights for future research directions (**Chapter 6**). Through its structured exploration of available S2S precipitation forecasts from NMME-2 and their hydrological application, this dissertation contributes to advancing the field of streamflow prediction and offers recommendations for enhancing streamflow forecast accuracy and reliability.

Chapter 1. Introduction

1.1 Background

Skillful streamflow forecasts at a Subseasonal-to-Seasonal (S2S) timescale, defined as 10 to 30 days into the future (Vitart et al., 2017), could provide crucial information, benefit human socio-economic activities, including reservoir operations, irrigation planning, proactive flood mitigations, etc. (Welles et al., 2007; White et al., 2017; Yang et al., 2021).

To perform streamflow predictions at S2S timescale (often termed "extended range" in hydrology), the classical ensemble streamflow prediction (ESP) approach is commonly adopted (Day, 1985; Troin et al., 2021). ESP is also the standard approach for extended-range hydrologic forecasting at the National River Forecast Centers over the United States (Schaake and Larson, 1998).

To apply ESP, a calibrated hydrologic model will be forced with measured precipitation for a hydrologic simulation up to the "current" time (i.e., forecast initializing time) as a first step. By doing so, the watershed's initial hydrologic conditions (IHCs, i.e., mainly soil moisture and/or snow condition) are estimated. Then, a series of precipitation measurements are randomly selected from historical records to form an ensemble of resampled precipitation forecasts, representing possible future precipitation scenarios. Lastly, an ensemble of streamflow predictions is generated through a combination of the calibrated hydrologic model, the estimated IHCs, and the randomly resampled precipitation forecasts.

Such a framework of ESP is fundamentally well-designed. While exploiting the physical characteristics of the hydrologic system through a chosen hydrologic model, ESP also considers the predictability of streamflow from both the IHCs and future precipitation scenarios. More importantly, with the fact that ensemble predictions are generated with resampled precipitation

forecasts, ESP can probabilistically predict future extreme events (i.e., flooding events) over the forecast horizon. This is because, under the classical "stationary" assumption in hydrology, it is believed that historical precipitation records shall repeat themselves and reflect future extreme streamflow scenarios. Furthermore, the employment of ensemble predictions of ESP allows expressions of uncertainty, which makes risk-based decision-making possible (Krzysztofowicz, 2001).

However, the quality of ESP-generated streamflow forecasts at S2S timescale is normally compromised, especially at longer forecast lead times after around 10 days. This is because the predictability of streamflow from the estimated IHCs would normally be lost at S2S timescale, leaving streamflow predictions dominated by the resampled precipitation forecasts (Cao et al., 2021; Shukla et al., 2013; Wood and Lettenmaier, 2008; Yossef et al., 2013). Due to the low deterministic predictive skill of resampled precipitation, instead of accurately predicting the mid-volume streamflow, the ESP framework can only provide probabilistic predictions upon extreme events with its ensemble spreads. Moreover, under the context of global climate change, the classical "stationary" assumption in hydrology might no longer be valid (Milly et al., 2008). This leads to changes in not only the magnitudes but also the frequency of extreme precipitation events over the globe (Miao et al., 2016; Sun et al., 2018). As a result, the resampled precipitation forecasts might not lead to reasonable representations of future extreme precipitation scenarios, which further undermine the fundamentals and strength of ESP in streamflow forecasting.

Dynamical S2S precipitation products are promising alternatives to the resampled precipitation for their overall better streamflow forecasting capabilities under the ESP framework. At short- and/or medium-range (i.e., 2-3 days and 7-10 days, respectively),

Numerical Weather Prediction (NWP) models can provide reliable and skillful forecasts globally (Bauer et al., 2015). In general, NWP has demonstrated high predictability of rainfall from initial atmospheric states, various types of observations, and the predictive skills can be further improved by advanced data assimilation techniques (Alley et al., 2019; Berner et al., 2011; Hamill and Juras, 2006; Lin et al., 2005; Palmer et al., 2004; Ritter and Geleyn, 1992; Shrestha et al., 2013; Sun et al., 2014). At long ranges (i.e., months and seasons), General Circulation Models (GCMs) coupled with dynamical oceanic and land surface components produce more skillful and informative seasonal/climate outlooks than the NWP models, because both the local land surface conditions and sea surface temperature (SST) circulations are incorporated into the GCMs' computation for future weather and climate predictions (Vitart et al., 2017).

However, the predictive skill of precipitation at S2S timescale is often regarded as "poor" or "none" as compared to weather predictions and climate outlooks due to the unique physical feature of the S2S weather. Many efforts have been made to identify additional predictability sources of the weather at S2S timescale, including initial atmospheric conditions (Cohen et al., 2010; Stockdale et al., 2015), initial land surface soil moisture (Asoka and Mishra, 2015; Guo et al., 2011), initial snow conditions (Thomas et al., 2016), initial sea surface temperatures (Chelton and Wentz, 2005), and planetary-scale oceanic patterns such as the El Niño–Southern Oscillation (ENSO), Madden–Julian oscillation (MJO), quasi-biennial oscillation (QBO) (Nardi et al., 2020; Pan et al., 2019). But it is still extremely challenging to provide reliable and accurate precipitation forecasts at S2S timescale. This is because S2S timescale lies in a transitional period between medium-range weather forecasts and longer-range seasonal climate outlooks, where the forecast lead time is sufficiently long that most of the predictability from

the initial atmosphere and land-surface conditions would be lost, but the forecast time horizon is also too short for the variability of the ocean to have a strong influence on local weather (Vitart et al., 2017).

Despite the challenge, many S2S precipitation forecast products have been made available by mission agencies all over the globe in the recent decade. Among existing S2S forecast products, the North America Multi-Model Ensemble (NMME) provides a variety of hydrometeorological forecasts at S2S timescale (Kirtman et al., 2014b). The NMME consists of the outputs from multiple coupled GCMs, each providing independent retrospective forecasts (hereafter referred to as "hindcasts") and real-time forecasts. The NMME phase I project (NMME-1) was initially launched in 2014 and further transitioned into Phase II (NMME-2) in more recent years. Both NMME-1 and NMME-2 datasets provide monthly initialized hindcasts and forecasts with a lead-time of up to 12 months. The major advancement of NMME-2 over the NMME-1 dataset is the provision of dynamically downscaled forecasts, in which the daily S2S precipitation forecasts are provided (Kirtman et al., 2014a; Pegion et al., 2019; Vitart et al., 2017).

There is a good number of existing research investigating the performance as well as hydrologic application of the NMME-1 dataset. For example, Becker et al. (2014) and Krakauer (2019) evaluated the skill of precipitation and temperature forecasts from NMME-1 globally. They reported that forecast bias is commonly presented, while the predictive skill decreases rapidly over forecast lead times. Similar evaluation studies upon NMME-1 monthly hydrometeorological forecasts have been carried out in different regions with similar results, such as in Southern and Southeast Asia, East Africa, and North America (Cash et al., 2019a; Shukla et al., 2019a; Slater et al., 2019a; Xu et al., 2018). Other related studies were conducted

focusing on improving the quality of precipitation forecasts from the NMME-1 dataset. For example, Slater et al. (2017) deployed different multi-model weighting techniques to improve the skill of NMME-1 monthly precipitation forecast across Europe. Xu et al. (2019a) applied several machine learning and wavelet approaches to bias-correct and downscale the monthly precipitation forecast from the NMME-1 dataset over China. Khajehei et al. (2018a) developed a Bayesian ensemble approach based on a Copula function to bias-correct the NMME-1 monthly precipitation forecast over the CONUS.

As compared to the studies on NMME-1, fewer studies investigate the data quality, as well as the hydrologic application of the S2S precipitation forecasts from the newer NMME-2 dataset. Among a limited number of studies, Wanders and Wood (2016) evaluated the precipitation forecast from NMME-2 globally on a bi-weekly basis. More recently, Baker et al. (2019) studied the precipitation forecasts of Climate Forecast System version 2 (CFSv2) from the NMME-2 dataset over the entire CONUS on a bi-weekly basis, and they found the forecast skill of CFSv2 decreases over lead time rapidly, but the forecast biases are insensitive to lead times.

In summary, as a well-recognized framework, the ESP heavily relies on inaccurate resampled precipitation forecasts, which results in inaccurate predictions of streamflow. Further, due to the ongoing climate change, the ability of ESP in predicting flooding events is likely compromised. Available S2S precipitation forecasts generated by GCMs from NMME-2 are promising alternatives to the resampled precipitation under the ESP framework for hydrologic forecasts. However, a limited number of studies exist that focus on NMME-2. Nevertheless, previous studies on precipitation forecasts from NMME-1 at other timescales

might provide some insights. According to previous studies, significant forecast bias and low temporal predictive skills are commonly presented in GCM-generated precipitation forecasts.

1.2 Research Questions

Although available S2S precipitation forecasts from NMME-2 provide a promising alternative to the resampled precipitation under the ESP framework, it is still unclear how the S2S precipitation forecasts from NMME-2 perform over the contiguous United States (CONUS). Therefore, the ***first research question*** this Ph.D. dissertation intends to address is ***(1) What are the levels of accuracy of the S2S ensemble precipitation forecasts from NMME-2 at different forecast lead times, within different seasons, and over different spatial domains across CONUS?***

After obtaining a full understanding of the NMME2 forecast skills, I further selected 4 watersheds located in the South Climate region (per National Center for Environmental Information, i.e., NCEI) with the least anthropogenic activities to test the hydrologic performance of the S2S precipitation forecasts from NMME-2. The raw NMME-2 forecasts are processed with a proposed statistical downscaling and bias correction approach to address the second research question: ***(2) In contrast to the resampled precipitation forecasts, how informative are the bias-corrected and downscaled S2S precipitation forecasts in producing ensemble streamflow forecasts under the ESP framework at typical watershed scales?***

It is anticipated that the S2S precipitation forecasts from NMME-2 shall bring an overall better streamflow prediction under the ESP framework as compared to the resampled precipitation. However, it is also anticipated that the S2S precipitation forecasts from NMME-2 shall only present a marginal level of advantageous predictive skill at longer forecast lead times beyond 2 weeks. Therefore, a novel post-processing approach is developed using a popular

Machine Learning (ML) technique to improve the predictive skill of raw S2S precipitation forecasts and address the research question of ***(3) Can Machine Learning (ML) algorithms provide a more reliable post-processing method with respect to improving the quality of precipitation forecasts over different geographical domains and lead times?***

With an overall better understanding of the S2S precipitation from NMME-2, as well as a novel post-processing approach to further improve the predictive performance of the S2S precipitation forecasts from NMME-2, the ***fourth*** research question this Ph.D. dissertation aims to address is that ***(4) How much of the added predictive skill obtained from the ML-improved S2S precipitation forecasts propagate into the streamflow predictions under the ESP framework at typical watershed scales?***

The remaining content of this dissertation intends to address the aforementioned research questions through the following four chapters. **Chapter 2** (published as Zhang et al., 2021) focuses on evaluating the predictive skill as well as forecast bias of the raw S2S precipitation forecasts from NMME-2 over the CONUS. In addition, the regional analysis is conducted to gain a better understanding of the overall performance of the S2S precipitation over different geo-locations. **Chapter 3** (under review as Zhang et al., 2023c) applies a standard bias correction and statistical downscaling approach to downscale the raw S2S precipitation forecasts from NMME-2 to 4 watersheds lie in the NCEI South Climate Region for streamflow hindcast experiments under the ESP framework. The performance of S2S precipitation from NMME-2 is benchmarked with resampled precipitation. **Chapter 4** (published as Zhang et al., 2023b) applies a Machine learning (ML) technique to incorporate additional forecast variables to improve the predictive skill of the raw S2S precipitation forecasts from NMME-2 for better predictions on extreme events over CONUS. **Chapter 5** (To be submitted to the Journal of

Hydrology as Zhang et al., 2023a) tests the actual hydrologic performance of the S2S precipitation from NMME-2 with the developed ML post-processing approach in streamflow predictions under the ESP framework at the same four watersheds studies in **Chapter 2**. At last, **Chapter 6** summarizes the major findings of this dissertation and gives recommendations for future works.

References

Alley, R.B., Emanuel, K.A. and Zhang, F. 2019. Advances in weather prediction. Science 363(6425), 342-344.

Asoka, A. and Mishra, V. 2015. Prediction of vegetation anomalies to improve food security and water management in India. Geophysical Research Letters 42(13), 5290-5298.

Baker, S.A., Wood, A.W. and Rajagopalan, B. 2019. Developing subseasonal to seasonal climate forecast products for hydrology and water management. JAWRA Journal of the American Water Resources Association 55(4), 1024-1037.

Bauer, P., Thorpe, A. and Brunet, G. 2015. The quiet revolution of numerical weather prediction. Nature 525(7567), 47-55.

Becker, E., den Dool, H.v. and Zhang, Q. 2014. Predictability and forecast skill in NMME. Journal of Climate 27(15), 5891-5906.

Berner, J., Ha, S.-Y., Hacker, J., Fournier, A. and Snyder, C. 2011. Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. Monthly Weather Review 139(6), 1972-1995.

Cao, Q., Shukla, S., DeFlorio, M.J., Ralph, F.M. and Lettenmaier, D.P. 2021. Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in coastal Western US watersheds. Journal of Hydrometeorology 22(6), 1535-1552.

Cash, B.A., Manganello, J.V. and Kinter, J.L. 2019. Evaluation of NMME temperature and precipitation bias and forecast skill for South Asia. Climate dynamics 53, 7363-7380.

Chelton, D.B. and Wentz, F.J. 2005. Global microwave satellite observations of sea surface temperature for numerical weather prediction and climate research. Bulletin of the American Meteorological Society 86(8), 1097-1116.

Cohen, J., Foster, J., Barlow, M., Saito, K. and Jones, J. 2010. Winter 2009–2010: A case study of an extreme Arctic Oscillation event. Geophysical Research Letters 37(17).

Day, G.N. 1985. Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management 111(2), 157-170.

Guo, Z., Dirmeyer, P.A. and DelSole, T. 2011. Land surface impacts on subseasonal and seasonal predictability. Geophysical Research Letters 38(24).

Hamill, T.M. and Juras, J. 2006. Measuring forecast skill: Is it real skill or is it the varying climatology? Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 132(621C), 2905-2923.

Khajehei, S., Ahmadalipour, A. and Moradkhani, H. 2018. An effective post-processing of the North American multi-model ensemble (NMME) precipitation forecasts over the continental US. Climate dynamics 51, 457-472.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P. and Becker, E. 2014a. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bulletin of the American Meteorological Society 95(4), 585-601.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M.P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D.G., Tippett, M.K., Barnston, A.G., Li, S., Rosati, A., Schubert, S.D., Rienecker, M., Suarez, M., Li, Z.E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W.J., Denis,

B. and Wood, E.F. 2014b. The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. Bulletin of the American Meteorological Society 95(4), 585-601.

Krakauer, N.Y. 2019. Temperature trends and prediction skill in NMME seasonal forecasts. Climate Dynamics 53(12), 7201-7213.

Krzysztofowicz, R. 2001. The case for probabilistic forecasting in hydrology. Journal of hydrology 249(1-4), 2-9.

Lin, C., Vasić, S., Kilambi, A., Turner, B. and Zawadzki, I. 2005. Precipitation forecast skill of numerical weather prediction models and radar nowcasts. Geophysical research letters 32(14).

Miao, C., Sun, Q., Duan, Q. and Wang, Y. 2016. Joint analysis of changes in temperature and precipitation on the Loess Plateau during the period 1961–2011. Climate Dynamics 47(9), 3221-3234.

Milly, P.C., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P. and Stouffer, R.J. 2008. Stationarity is dead: Whither water management? Science 319(5863), 573-574.

Nardi, K.M., Baggett, C.F., Barnes, E.A., Maloney, E.D., Harnos, D.S. and Ciasto, L.M. 2020. Skillful all-season S2S prediction of US precipitation using the MJO and QBO. Weather and Forecasting 35(5), 2179-2198.

Palmer, T., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delécluse, P., Déqué, M., Diez, E., Doblas-Reyes, F.J. and Feddersen, H. 2004. Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bulletin of the American Meteorological Society 85(6), 853-872.

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S. and Higgins, W. 2019. Precipitation prediction skill for the West Coast United States: From short to extended range. Journal of Climate 32(1), 161-182.

Pegion, K., Kirtman, B.P., Becker, E., Collins, D.C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D. and Zhu, Y. 2019. The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. Bulletin of the American Meteorological Society 100(10), 2043-2060.

Ritter, B. and Geleyn, J.-F. 1992. A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. Monthly weather review 120(2), 303-325.

Schaake, J. and Larson, L. 1998 Ensemble streamflow prediction (ESP): Progress and research needs, pp. J19-J24.

Shrestha, D., Robertson, D., Wang, Q., Pagano, T. and Hapuarachchi, H. 2013. Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. Hydrology and Earth System Sciences 17(5), 1913-1931.

Shukla, S., Roberts, J., Hoell, A., Funk, C.C., Robertson, F. and Kirtman, B. 2019. Assessing North American multimodel ensemble (NMME) seasonal forecast skill to assist in the early warning of anomalous hydrometeorological events over East Africa. Climate Dynamics 53, 7411-7427.

Shukla, S., Sheffield, J., Wood, E.F. and Lettenmaier, D.P. 2013. On the sources of global land surface hydrologic predictability. Hydrology and earth system sciences 17(7), 2781-2796.

Slater, L.J., Villarini, G. and Bradley, A.A. 2017. Weighting of NMME temperature and precipitation forecasts across Europe. Journal of Hydrology 552, 646-659.

Slater, L.J., Villarini, G. and Bradley, A.A. 2019. Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA. Climate dynamics 53, 7381-7396.

Stockdale, T.N., Molteni, F. and Ferranti, L. 2015. Atmospheric initial conditions and the predictability of the Arctic Oscillation. Geophysical Research Letters 42(4), 1173-1179.

Sun, J., Xue, M., Wilson, J.W., Zawadzki, I., Ballard, S.P., Onvlee-Hooimeyer, J., Joe, P., Barker, D.M., Li, P.-W. and Golding, B. 2014. Use of NWP for nowcasting convective precipitation: Recent progress and challenges. Bulletin of the American Meteorological Society 95(3), 409-426.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S. and Hsu, K.L. 2018. A review of global precipitation data sets: Data sources, estimation, and intercomparisons. Reviews of Geophysics 56(1), 79-107.

Thomas, J.A., Berg, A.A. and Merryfield, W.J. 2016. Influence of snow and soil moisture initialization on sub-seasonal predictability and forecast skill in boreal spring. Climate dynamics 47(1), 49-65.

Troin, M., Arsenault, R., Wood, A.W., Brissette, F. and Martel, J.L. 2021 Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years, Wiley Online Library.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E. and Fuentes, M. 2017. The subseasonal to seasonal (S2S) prediction project database. Bulletin of the American Meteorological Society 98(1), 163-173.

Wanders, N. and Wood, E.F. 2016. Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. Environmental Research Letters 11(9), 094007.

Welles, E., Sorooshian, S., Carter, G. and Olsen, B. 2007. Hydrologic verification: A call for action and collaboration. Bulletin of the American Meteorological Society 88(4), 503-512.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J. and Murray, V. 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorological applications 24(3), 315-325.

Wood, A.W. and Lettenmaier, D.P. 2008. An ensemble approach for attribution of hydrologic prediction uncertainty. Geophysical Research Letters 35(14).

Xu, L., Chen, N., Zhang, X. and Chen, Z. 2018. An evaluation of statistical, NMME and hybrid models for drought prediction in China. Journal of hydrology 566, 235-249.

Xu, L., Chen, N., Zhang, X., Chen, Z., Hu, C. and Wang, C. 2019. Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning. Climate dynamics 53, 601-615.

Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D. and Peng, Q. 2021. A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region. Journal of Hydrology 602, 126723.

Yossef, N.C., Winsemius, H., Weerts, A., van Beek, R. and Bierkens, M.F. 2013. Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. Water Resources Research 49(8), 4687-4699.

Zhang, L., Kim, T., Yang, T., Hong, Y. and Zhu, Q. 2021. Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous US. Journal of Hydrology 603, 127058.

Zhang, L., Yang, T., Gao, S. and Hong, Y. 2023a. Combine Machine Learning with a distribution-based technique to adapt Subseasonal-to-Seasonal (S2S) precipitation forecasts at local watersheds for hydrologic Ensemble Streamflow Prediction (ESP). Journal of Hydrology (to be submitted).

Zhang, L., Yang, T., Gao, S., Hong, Y., Zhang, Q., Wen, X. and Cheng, C. 2023b. Improving Subseasonal-to-Seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous US using machine learning models. Atmospheric Research 281, 106502.

Zhang, L., Yang, T., Gao, S., Yang, H., Fan, M., Lu, D., Xu, H. and Xiao, C. 2023c. An alternative Ensemble Streamflow Prediction (ESP) approach using improved Subseasonal-to-Seasonal (S2S) preciptiation forecasts from the North America Multi-Model Ensemble Phase II (NMME-2) dataset. Journal of Hydrology (under review).

Chapter 2. Evaluation of Subseasonal-to-Seasonal (S2S) Precipitation Forecast from the North American Multi-Model Ensemble Phase II (NMME-2) over the Contiguous U.S.

The major content from this chapter has been published in the *Journal of Hydrology*. © Elsevier. Used with permission.

Zhang, L., Kim, T., Yang, T., Hong, Y., & Zhu, Q. (2021). Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous US. *Journal of Hydrology, 603*, 127058.

Abstract

The second phase of the North America Multi-Model Ensemble (NMME-2) provides globally available Subseasonal-to-Seasonal (S2S) precipitation forecasts with a daily resolution. The S2S precipitation forecasts are getting more and more attention for their potential in providing extended hydrometeorological forcing information for water resources planning. Existing S2S precipitation forecast products suffer from low forecast skills when the lead time increases, which hinders their applicability for watershed-scale hydrologic modeling. It is crucial to evaluate the forecast skill and reveal the strengths and weaknesses of the new NMME-2 dataset before applying it to any hydrologic applications. Specifically, this evaluates the S2S precipitation hindcasts from five NMME-2 models from 1982 to 2011 over the contiguous United States (CONUS). The evaluation focuses on two important aspects of precipitation forecast, including forecast bias and predictive skill with a comprehensive analysis of the seasonal, geographical locations, and lead time dependencies. The AN81d dataset from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) is used as a ground truth reference. Results show that the forecast biases are not sensitive to lead times, but they are seasonally dependent on all NMME-2 models. Overestimations are found in the Western U.S. in cooler seasons, while underestimations are observed in the central regions of

the U.S. in warmer seasons. The predictive skills of NMME-2 models generally decrease as the forecast lead time increases. The simple model averaging (SMA) of five NMME-2 models demonstrates a higher predictive skill than any single NMME-2 model, while the FLORB01 has the lowest skill scores among five NMME-2 models. Spatially, the highest forecast skill scores are observed in coastal areas in the Western U.S. with a one-week lead time.

2.1 Introduction

Precipitation is one of the most important components of the hydrologic cycle (Sorooshian et al., 2011). Accurate and reliable precipitation forecasts with certain lead times could be beneficial in planning and managing social and economic activities, preventing financial and life losses from water-related disasters (Akbari Asanjan et al., 2018; Palmer, 2002). Different forecast products can be categorized as short-, medium- and long-range forecasts based on their lead times. At short- and/or medium-ranges (i.e., 2-3 days and 7-10 days, respectively), Numerical Weather Prediction (NWP) models can provide reliable and skillful forecasts globally (Bauer et al., 2015). The high precipitation forecast accuracy of NWP models at short- and/or medium ranges are due to the high predictability of rainfall from initial atmospheric states, various types of observations, and advanced data assimilation techniques. However, one common critique of the NWP model is that its forecast skill decreases rapidly, and the associated forecast uncertainty increases dramatically when the forecast lead time extends two weeks and beyond (Alley et al., 2019; Berner et al., 2011; Hamill and Juras, 2006; Lin et al., 2005; Palmer et al., 2004; Ritter and Geleyn, 1992; Shrestha et al., 2013; Sun et al., 2014). This is because NWP models heavily rely on the initial states of the atmosphere, and the predictability coming from the initial states dissipates rapidly over lead time. At longer ranges (i.e., months, seasons, years, and even decades), Earth System Models (ESMs) and General

Circulation Models (GCMs) coupled with dynamic oceanic and land surface components produce more skillful and informative climate outlooks than the NWP models, because ESMs and couped GCMs are able to incorporate both the local land surface conditions and sea surface temperature (SST) circulations into the computation for future weather and climate predictions (Vitart et al., 2017).

A forecast gap lies in the transitional period between medium-range weather forecasts and longer-range seasonal climate outlooks. This transitional period is also referred to as the Subseasonal-to-Seasonal (S2S) timescale, which defines a specific time range beyond 10 days and up to 30 days into the future (White et al., 2017). At the S2S time range, the forecast lead time is sufficiently long that most of the predictability from the initial conditions would be lost, but it is also too short for the variability of the ocean to have a strong influence on local weather (Vitart et al., 2017). This unique physical feature of weather predictability made precipitation forecast at the S2S range notoriously challenging and also makes the S2S forecast often considered as a "desert of predictability" (White et al., 2017).

The S2S hydrometeorological forecasts are important information and have great potential in providing seamless streamflow and flood forecasts at the S2S range. Traditionally, river forecast centers and weather service centers over the globe issue probabilistic seasonal streamflow forecasts forced by seasonal and/or monthly climate outlooks (Wood and Lettenmaier, 2006). However, this kind of seasonal streamflow forecast can only reflect an increased or decreased risk of flooding, but it does not have the ability to predict floods at the S2S range. Meanwhile, the S2S hydrometeorological forecasts provide an opportunity to assist streamflow forecasts, which not only reflect the flooding risks but also provides additional information regarding the timing, frequency, or severity of potential floods within seasons

(White et al., 2015). In addition, accurate S2S hydrometeorological forecasts could help the operation of reservoirs in scheduling optimal water supplies and hydropower generation, given foreseeable dry and wet water conditions (Sankarasubramanian et al., 2009). Despite the potential usage and benefits, the S2S hydrometeorological forecast appears to be a newer research area in recent years as compared with the weather and climate forecasts at other ranges (Vitart and Robertson, 2018).

Previous studies also concluded that the predictability of S2S forecast came from several sources, including initial atmospheric conditions (Cohen et al., 2010; Stockdale et al., 2015), initial land surface soil moisture (Asoka and Mishra, 2015; Guo et al., 2011), initial snow conditions (Thomas et al., 2016), and initial sea surface temperatures (Chelton and Wentz, 2005). In some recent studies, the planetary-scale oceanic patterns are found to be the main predictability source of S2S forecasts, and these interconnection climate indices include the El Niño–Southern Oscillation (ENSO), Madden–Julian oscillation (MJO), quasi-biennial oscillation (QBO) (Nardi et al., 2020; Pan et al., 2019). There are many existing efforts and programs that focus on the hydrometeorological forecasts at S2S timescale, such as the European Center for Medium Range Weather Forecasts (ECMWF), the Environment Canada (EC), the Japan Meteorological Agency (JMA), the China Meteorological Administration (CMA), etc. Each of those agencies uses different coupled GCMs and ESMs to obtain the precipitation forecasts at the S2S range with different temporal and spatial resolutions.

The North America Multi Model Ensemble (NMME) is a multi-agency initiated and collaborative program that provides a variety of hydrometeorological forecasts at S2S timescale (Kirtman et al., 2014a). The NMME consists of the outputs from multiple coupled GCMs and ESMs, each providing independent retrospective forecasts (hereafter referred to as “hindcasts”)

and real-time forecasts. The NMME phase 1 project (NMME-1) was initially launched in 2014 and further transitioned into Phase 2 (NMME-2) in more recent years. Both NMME-1 and NMME-2 datasets provide monthly initialized hindcasts and forecasts with a lead time of up to 12 months. The major advancement of NMME-2 over the NMME-1 dataset is the provision of dynamically downscaled forecasts, in which the new dataset provides precipitation forecasts with a daily temporal resolution at the S2S scale, while the outputs from the NMME-1 dataset only provide forecasts with monthly resolution (Kirtman et al., 2014b).

There is a good number of existing research that investigated the forecast quality and accuracy of the hydrometeorological forecasts from NMME-1 dataset. For example, Becker et al. (2014) and Krakauer (2019) evaluated the skill of precipitation and temperature forecasts from NMME-1 globally, and they found that the simple model averaging (SMA) of NMME-1 models showed better forecast skills than any individual NMME-1 models and the forecast skills varied depending on geographical regions and seasons. Similar evaluation studies upon NMME-1 monthly hydrometeorological forecasts have been carried out in different regions. For example, Cash et al. (2019b) observed significant systematic error in both precipitation and temperature forecasts from NMME-1 in two Southern Asia regions and found the highest forecast skills were observed at the shortest lead times. Shukla et al. (2019b) found the skill scores of precipitation forecasts from the NMME-1 were higher during ENSO years over East Africa. Slater et al. (2019b) evaluated the skill of precipitation and temperature forecasts from NMME-1 over seven geological regions of the continental United States. Slater et al. (2019b) found the highest forecast skill was generally observed at the shortest lead time, and the performance of NMME-1 forecast were spatially and seasonally dependent. The study from Slater et al. (2019b) also consistently showed that higher forecast skills can be gained by

averaging multiple NMME-1 models and the skill of hydrometeorological forecast from NMME-1 quickly declined to marginal levels as lead time increases. To address the low accuracy issues associated with the raw NMME-1 forecasts, many follow-on studies focused on improving precipitation forecasts from the NMME-1 dataset. For example, Slater et al. (2017) deployed different multi-model weighting techniques to improve the skill of NMME-1 monthly precipitation forecast across Europe. Xu et al. (2019b) applied several machine learning and wavelet approaches to bias-correct and downscale the monthly precipitation forecast from the NMME-1 dataset over China. Khajehei et al. (2018b) developed a Bayesian ensemble approach based on a Copula function to bias-correct the NMME-1 monthly precipitation forecast over the CONUS.

As compared to the studies on the NMME-1 dataset, there are fewer studies that investigate the quality of S2S precipitation forecasts from the newer NMME-2 dataset. Among a limited number of studies, Wanders and Wood (2016) evaluated the precipitation forecast from NMME-2 globally on a bi-weekly basis. By aggregating the evaluation result into three global regions, including the tropics, extratropic and northern latitudes, they found the forecasts skill decreased over lead time as well as the performance discrepancies between different NMME-2 models. Zhou and Kim (2018) evaluated the ability of NMME-2 in predicting the wintertime atmospheric rivers (AR) and moisture flux over the Northeast Pacific in response to ENSO. They found NMME-2 dataset has significant regional biases in anomalous landfalling AR frequency, which highlighted the challenges in forecasting regional precipitation events. More recently, Baker et al. (2019) studied the precipitation forecasts of Climate Forecast System version 2 (CFSv2) from the NMME-2 dataset over the entire CONUS on a bi-weekly basis, and they found the forecast skill of CFSv2 decreases over lead time rapidly, but the

forecast biases were insensitive to lead times. Guo and Nie (2020) evaluated the daily precipitation forecasts of CFSv2 over east China. Their result indicated that the raw precipitation forecasts of CFSv2 were substantially biased, and the extreme precipitation events over east China have been underestimated by CFSv2. Becker et al. (2020) studied the performance of precipitation forecasts from NMME-2 globally. But Becker et al. (2020) aggregated the daily forecast values from NMME-2 into monthly values since their study focused on seasonal scale on proving the improvement of forecast quality corresponding to the iteration of NMME models. To summarize the existing studies focusing on precipitation forecasts from NMME-2, existing research either (1) targeted the evaluation of NMME-2 at a spatial scale that is too large to provide valuable reference for regional hydrologic studies, or (2) were only focusing on monthly forecast values and overlooked the S2S forecasts from NMME-2, or (3) have only included a certain member of NMME-2 into the study and did not comprehensively evaluate all NMME-2 members as a whole with the consideration of lead times, seasonality and precipitation's geographical characteristics.

Thus, more detailed evaluations of the S2S precipitation forecasts from the NMME-2 data across CONUS are still critically needed. From existing studies, although the evaluation results of the seasonal precipitation forecast from NMME over CONUS may have some consistency and similarities, it is still unknown how exactly do precipitation forecasts from different NMME-2 models perform at the S2S range over the entire CONUS. In addition, the merit of the S2S precipitation forecast from NMME-2 in forecasting extreme events has not been verified at large scales. And these missing pieces of research, in return, limit further hydrologic applications of the NMME-2 dataset, since potential maximum streamflow prediction is one of

the most essential and desired outcomes of hydrologic forecasts at the S2S range (Day, 1985; Gobena and Gan, 2010)

To fill these gaps, as well as to provide valuable reference information and knowledge for future hydrologic research on NMME-2, this chapter aims to answer the following research questions: ***1) How does the S2S precipitation forecast from NMME-2 perform over the entire CONUS? 2) What are the differences between S2S precipitation forecasts from different NMME-2 models in terms of their performance over CONUS? 3) What is the forecast skill of the individual NMME-2 models at different regions over the CONUS? 4) Do certain NMME-2 models outperform others with the consideration of certain regions, lead times, and seasons? and 5) what are the NMME-2 model's performance in predicting extreme precipitation events over the CONUS?***

To answer these research questions, in this chapter, I evaluated the S2S precipitation forecasts from five NMME-2 models, and their grand ensemble (i.e., all five NMME-2 models as a whole set) is collected and analyzed over the CONUS. The study period is from 1982 to 2011. All five NMME-2 models selected in this chapter provide daily S2S precipitation forecast, except for the CFSv2, which was already studied by Baker et al. (2019). The AN81d dataset generated from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) is used as the reference dataset. The forecast bias and forecast skill are examined since forecast bias and poor forecast skill are two major obstacles in applying precipitation forecasts to hydrologic simulations (Zalachori et al., 2012). In addition, the ability of NMME-2 S2S precipitation forecasts in predicting extreme precipitation events are further evaluated and compared to the benchmark performance of the historical resampled precipitation forecasts. In this evaluation study, I separate the forecast lead time by weeks (i.e., from week one to week

four to cover the whole S2S forecast range), since the forecast skills scores are dependent on the lead time. Comparison and analysis on the forecast data quality are further conducted over nine National Centers for Environmental Information (NCEI) climate regions and four seasons, which are more inclusive and detailed as compared to the existing NMME-2 evaluation studies mentioned above. Specifically, this chapter extends the existing study from Baker et al. (2019) and Wanders and Wood (2016), whereas my study presents a more representative evaluation of five NMME-2 models and the evaluation spans for the full S2S range and covers the entire CONUS, on a finer weekly temporal resolution.

The rest of this **Chapter 2** is organized as follows: In **Section 2.2**, I present data and study regions. **Section 2.3** describes the evaluation metrics and methodologies. **Sections 2.4** and **2.5** present the results and discussions, respectively. The main conclusions and findings are summarized in **Section 2.6**.

2.2 Data and Study Regions

There are a total of seven different models available in the NMME-2 dataset. In this chapter, I used five NMME-2 models that provide daily precipitation forecasts covering the S2S range. Table 2.1 presents the basic information of the five selected NMME-2 models, which includes the Canadian Coupled Climate Model version 3 and 4 (CanCM3, CanCM4) from the Canada's Climate Model Center (CMC), the Community Climate System Model 4.0 (CCSM4) from the National Center for Atmospheric Research (NCAR), the Forecast-oriented Low Ocean Resolution model using parameter set B (FLORB01) from the Geophysical Fluid Dynamics Laboratory (GFDL), and the Goddard Earth Observing System version 5 model (GEOS5) from the National Astronautics and Space Administration (NASA). Each NMME-2 model generates ensemble forecasts through perturbed physics strategy and/or under different initial conditions:

The CanCM3, CanCM4, CCSM4, and GEOS5 models consist of 10 ensemble members and the FLORB01 model consists of 12 ensemble members. The study period was set from 01/01/1982 to 12/31/2011 (30 years), which overlaps with the hindcast/forecast periods for all five NMME-2 models. All NMME-2 hindcast/forecast datasets are available at: https://www.earthsystemgrid.org/search.html?Project=NMME.

Table 2.1 A list of selected NMME-2 models.

Model	Temporal Resolution	Spatial Resolution	Lead Time (Days)	Ensemble members	Reference
CanCM3	Daily	1°×1°	up to 365	10	Merryfield et al. (2013)
CanCM4	Daily	1°×1°	up to 365	10	Merryfield et al. (2013)
CCSM4	3-hourly/Daily	1°×1°	up to 365	10	Vertenstein et al. (2010)
FLORB01	Daily	0.5°×0.625°	up to 365	12	Delworth et al. (2012)
GEOS5	Daily	1°×1°	up to 274	10	Vernieres et al. (2012)

The daily precipitation dataset AN81d generated from the Parameter-elevation Regressions on Independent Slopes Model (hereafter referred to as PRISM) is used as a reference precipitation dataset in this chapter. The PRISM data is available from 1981 to near-present in gridded format with a spatial solution of 4 km (~0.04 degree) across the CONUS. The PRISM data combines surface observations with a digital elevation model to account for the orographic enhancement of precipitation. In addition to rain-gauge records, the PRISM data also incorporates the Radar measurement into account when producing corrected data over the central and eastern U.S. regions (Daly and Bryant, 2013). Since the PRISM dataset does not incorporate assimilated information from numerical weather forecasting models or meteorological reanalysis, it represents an independent dataset that is suitable for hydrologic studies (Radcliffe and Mukundan, 2017). Numerous hydrologic studies have used PRISM

precipitation data as a reliable reference for model evaluation, bias-correction for remotely sensed precipitation estimation products, and forecast verification studies (Ashfaq et al., 2016; Oubeidillah et al., 2014; Prat and Nelson, 2015; Widmann and Bretherton, 2000).

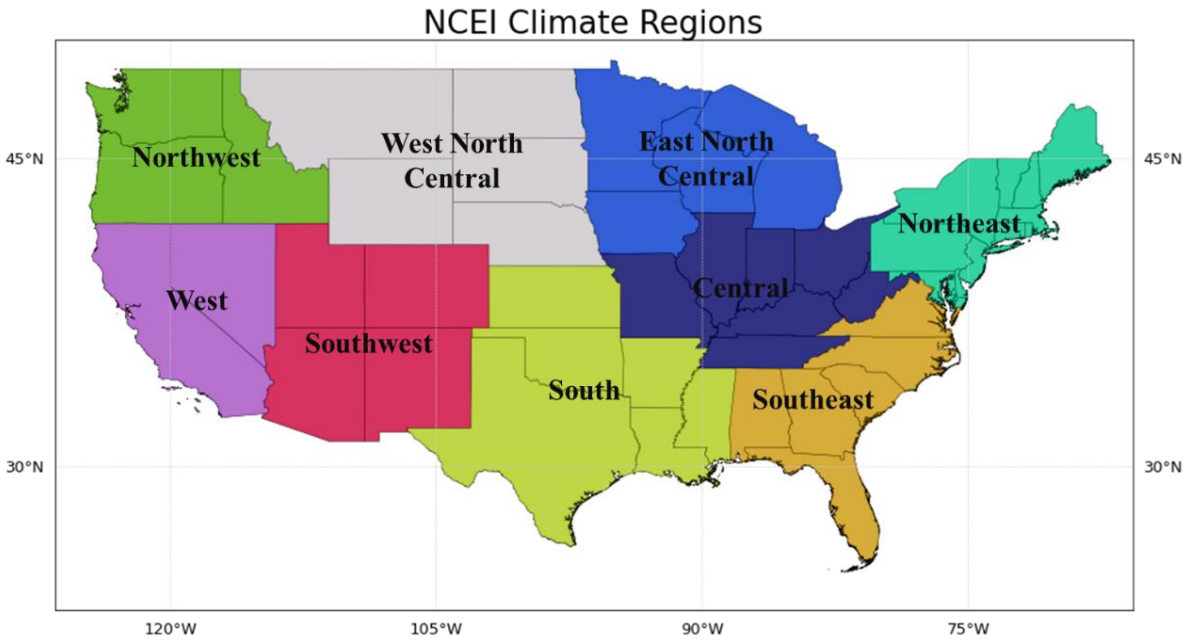

Figure 2.1 Climate region partitioning across CONUS per NCEI.

Figure 2.1 shows nine different climate regions across the CONUS, which are defined by the National Centers for Environmental Information (NCEI) (Karl and Koss, 1984). These nine climatic regions separate the CONUS into Northwest, West North Central, East North Central, Northeast, Central, West, Southwest, South, and Southeast regions. Within the NECI climate regions, the Sierra Nevada Mountain and the Rocky Mountain are across the Northwest, West, Southeast, and West North Central regions, and the Appalachian Mountain covers parts of the Northeast and Southeast regions. In this chapter, I evaluate the precipitation forecast of the NMME-2 at each pixel across the CONUS with spatial analysis according to the spatially averaged results over these nine climate regions. This is because rainfall presents different

physical and dynamical features and patterns over different climates, i.e., orographic elevation induced rainfall, frontal precipitation, and/or, convective systems.

2.3 Methodology, Procedure, and Evaluation Metrics

In this chapter, I first collected the NMME-2 precipitation forecasts and PRISM dataset and conducted data pre-processing. The NMME-2 precipitation forecasts are initialized at the beginning of months. NMME-2 monthly initialized forecasts provide forecast values with the lead time up to 1 year (365 or 366 days). All collected NMME-2 precipitation forecasts are truncated to 28 days (4 weeks) and then aggregated into weekly forecast values (e.g., day 1 to 7 as of week 1; day 8 to 14 as of week 2; day 15 to 21 as of week 3; and day 22 to 28 as of week 4). Spatially, both NMME-2 and PRISM datasets were re-gridded into 0.25° resolution using the nearest neighbor method for consistency. The grand ensemble of all selected NMME-2 models (i.e., a total of 52 different realizations in Table 2.1) is also constructed after the data pre-processing. The evaluation of forecast accuracy on individual NMME-2 model and the 52 ensemble is conducted by the seasonality of December-January-February (DJF), March-April-May (MAM), June-July-August (JJA), and September-October-November (SON). In other words, the evaluation experiment in this chapter comprehensively considers the forecast skill and model bias by different climate regions, forecast lead times, as well as the seasonality over the CONUS.

In this chapter, I use the commonly accepted approaches of pixel-based and spatial forecast evaluation metrics. Two evaluation metrics are included, i.e., the percentage bias (PBIAS) and the anomaly correlation coefficient (ACC). The PBIAS and ACC of (1) the ensemble means of individual NMME-2 models and (2) the SMA of the grand ensemble, are computed to provide overall bias and forecast skill in a quantitative manner.

2.3.1 Percentage Bias (PBIAS)

The PBIAS measurement reflects the degree of the under- and/or over-estimations of precipitation forecast that are vital for potential future hydrologic applications. To quantify the bias pattern of NMME-2 precipitation forecasts over the CONUS, the PBIAS of the ensemble means of single NMME-2 models and the SMA of the grand ensemble are computed with Equation (2.1).

$$PBIAS = \frac{\bar{x}-\bar{y}}{\bar{y}} \times 100\% \quad (2.1)$$

Where $\bar{x}$ is long-term mean value of precipitation forecasts at a certain location over CONUS, $\bar{y}$ is the long-term mean reference precipitation at a certain location over CONUS. And PBIAS is the percentage differences between mean forecast values and mean reference values over the study period. Positive PBIAS values indicate overestimations by forecasts, while negative PBIAS values indicate underestimations by forecasts.

2.3.2 Anomaly Correlation Coefficient (ACC)

The ACC is a widely used metric in the climate prediction community. It measures the degree of association between forecast and observed deviation from climatology. The advantage of ACC over some other metrics is that ACC can separate effects due to the existence of forecast bias in evaluating forecast skill. The ACC score of 1 indicates that the forecast provides perfect information and a score of zero means the forecast contains no information at all. The ACC skill scores of the ensemble mean of each individual NMME-2 model and the SMA of the grand ensemble are calculated following Equation (2.2) (Murphy and Epstein, 1989).

$$ACC = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \quad (2.2)$$

where x is the forecast/hindcast precipitation anomalies at a certain lead time and y is reference precipitation anomalies at the same lead time, n is the total number of hindcasts/forecast values made for that lead time, and ACC is the anomaly correlation coefficient skill score for the forecasts/hindcasts.

2.4 Results

2.4.1 Forecast Bias

Figure 2.2 presents the overall PBIAS of NMME-2 precipitation hindcasts over CONUS. It consists of the results from five individual NMME-2 models (CanCM3, CanCM4, FLORB01, GEOS5, and CCSM4) and the SMA of the grand ensemble of five NMME-2 models. In Figure 2.2, the positive bias in cooler colors (blue) is associated with overestimation and the negative bias in warmer colors (red) is associated with underestimation.

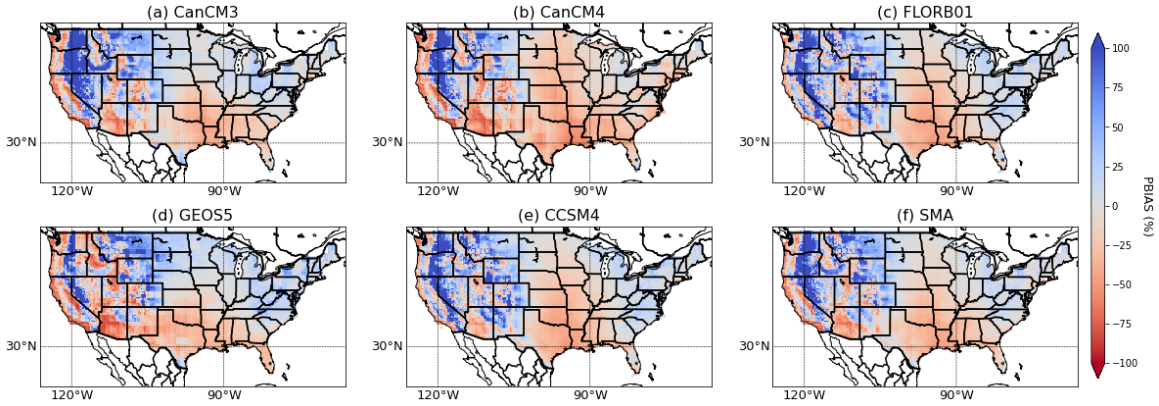

Figure 2.2 The overall pattern of PBIAS for NMME-2 precipitation hindcasts over the CONUS.

According to Figure 2.2, the highest level of model PBIAS of the NMME-2 dataset appears in the central and western U.S., and the forecast biases are relatively lower in eastern regions than that in the western regions. This similar spatial variability of overall PBIAS can be observed across individual NMME-2 models and the SMA. In addition, I notice that all five NMME-2 models demonstrate both positive and negative PBIAS with a “mosaic-like” pattern

over the Northwest, West, West North Central, and Southwest regions (Figure 2.2a, 2.2b, 2.2c, 2.2d, and 2.2e), where the Sierra Nevada Mountains and the Rocky Mountains ranges. On the contrary, across the Northwest and Southeast regions where Appalachian Mountain is located, the GEOS5 model (Figure 2.2d) demonstrates a significant level of overestimation, while other NMME-2 models show a relatively smaller level of biases. In the South region, all five NMME-2 models are consistently underestimating precipitation. The SMA (Figure 2.2f) has a very similar pattern of spatial variability in the overall pattern of PBIAS compared to the five individual single NMME-2 models.

The following Figure 2.3 presents the PBIAS of NMME-2 precipitation hindcasts across CONUS at different lead times (i.e., week one to week four). According to Figure 2.3, only some minor differences are observed in different climate regions, and no obvious lead-time dependences of PBIAS are observed for all NMME-2 members. At all lead times, the spatial pattern of weekly PBIAS is very similar to the overall pattern of PBIAS observed in Figure 2.2. The SMA also shows similar behavior to the five individual NMME-2 models at all lead times. Combining the results from both Figures 2.2 and 2.3, I observe that the overall model biases are not sensitive to lead times, though individual models could be associated with different levels of PBIAS and the model errors are also spatially varying across the CONUS.

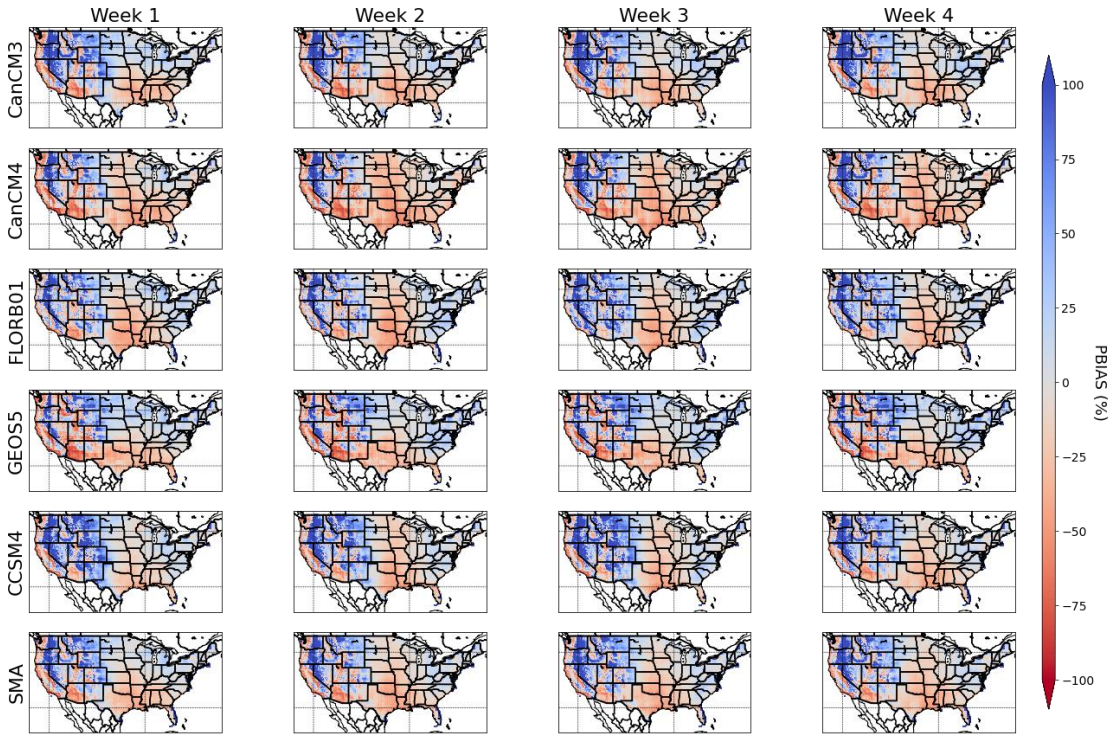

Figure 2.3 The pattern of PBIAS for NMME-2 precipitation hindcasts over the CONUS at different lead times (i.e., day 1 to 7: week 1, day 8 to 14: week 2, day 15 to 21: week 3, and day 22 to 28: week 4).

In the following Figure 2.4, I present the seasonal PBIAS of NMME-2 precipitation hindcasts over CONUS. According to Figure 2.4, both the individual NMME-2 model and the SMA exhibit a strong seasonal variation of PBIAS over CONUS. Individual model biases may demonstrate significant changes from positive to negative or vice-versa from season to season. Specifically, in DJF, all NMME-2 models show significant overestimation in most parts of the Northwest, West, West North Central, and Southwest regions. In MAM, the general overestimation of precipitation is reduced as compared to that in DJF. In MAM, I also notice that the spatial patterns of the obtained PBIAS are very similar to the overall patterns of PBIAS as shown in prior Figure 2.2, suggesting the MAM is the most representative season to show model's error variability in space. In JJA, all NMME-2 models show a significant level of underestimation in the Northwest, West, Southwest, South, and West North Central regions. In

SON, the level of underestimation by all models is smaller than that in JJA, while the level of overestimation increased in the Northwest region.

The results in Figure 2.3 indicate strong seasonal dependences of model PBIAS at different regions across CONUS, especially in DJF and JJA. Some noticeable model discrepancies are observed in different seasons. For example, in SON, the CanCM4 exhibits a more significant level of underestimation than other NMME-2 models in the Southeast, South, Southwest, and West North Central regions. In MAM, the GEOS5 demonstrates a higher level of overestimation than other NMME-2 models in the West North Central region, while a higher level of underestimation is observed with the same GEOS5 as compared to other models in the Southwest, South, and Southeast regions.

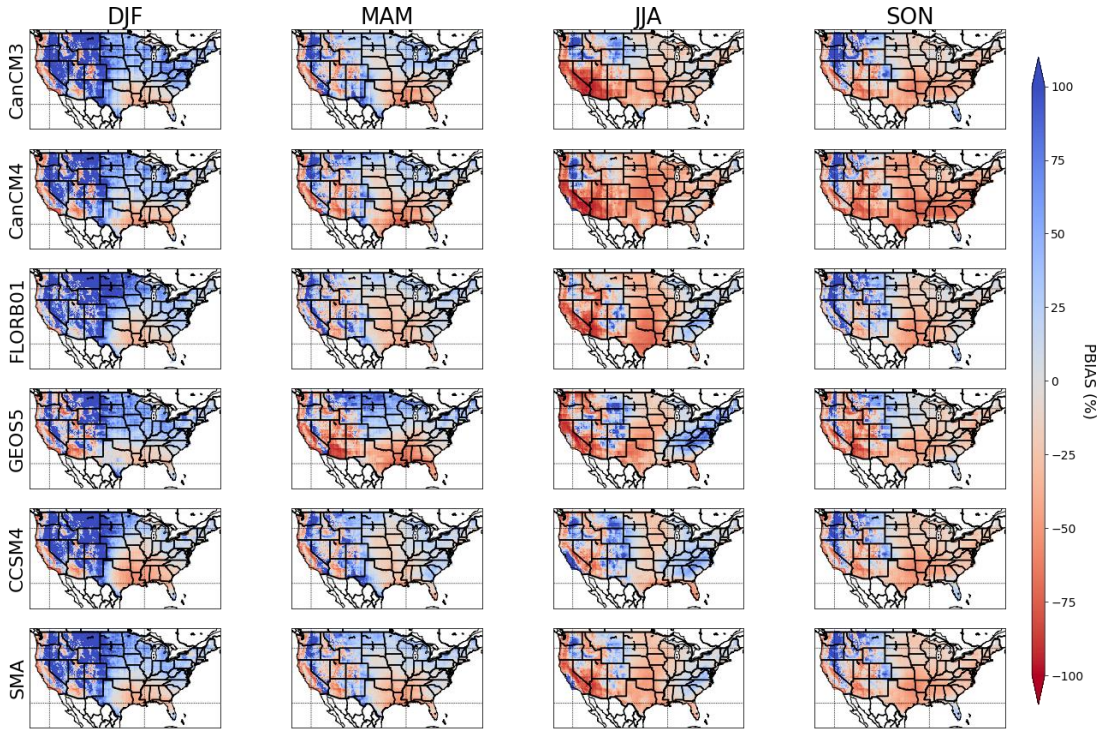

Figure 2.4 The PBIAS of NMME-2 precipitation hindcasts within different seasons (i.e., DJF, MAM, JJA, SON) across CONUS.

To better illustrate the large-scale spatial patterns of model biases with a joint consideration of seasonality and lead time dependences, I further computed and presented the regional

averaged PBIAS values of NMME-2 over nine NCEI climate regions for different seasons (DJF, MAM, JJA, and SON) and lead times (weeks 1-4) in the following Figure 2.5. In Figure 2.5, the cooler colors indicate overestimation, while the warmer colors indicate underestimation. The numbers in each cell of Figure 2.5 indicate the actual PBIAS values obtained for each model, lead time, and season. According to the heat map of Figure 2.5, the largest overestimations made by all NMME-2 models occur in the West North Central region in DJF. Significant underestimations are also observed in the warmer seasons of SON and JJA at most of the regions over CONUS except for the Northwest. The layout pattern of nine NCEI climate regions in Figure 2.5 (from left to right) is consistent with the geographic patterns of those regions in the real-world layout (from west to east). According to the quantitative values presented in Figure 2.5 (from left to right), there is an apparent decreasing trend of PBIAS from west to east. This numerical trend and decreasing pattern are more evident than that of different seasonality and lead times. Specifically, the overall overestimation in the Northwest, West, and West North Central regions observed in Figure 2.2 appears to be mainly contributed by the overestimations that occurred in DJF. The underestimation observed in Figure 2.2 seems to be mainly contributed by the underestimations that occurred in warmer seasons of JJA and SON. In terms of the different lead times, I found that the PBIAS values are not sensitive to the lead times since the colors of grid boxes in Figure 2.5 remain relatively consistent along the x-axis within most of the subplots. Only CanCM4, FLORB01, CCSM4, and SMA models show some noticeable variations of PBIAS over lead times in the West, the West North Central, and the Southwest regions, in JJA and SON.

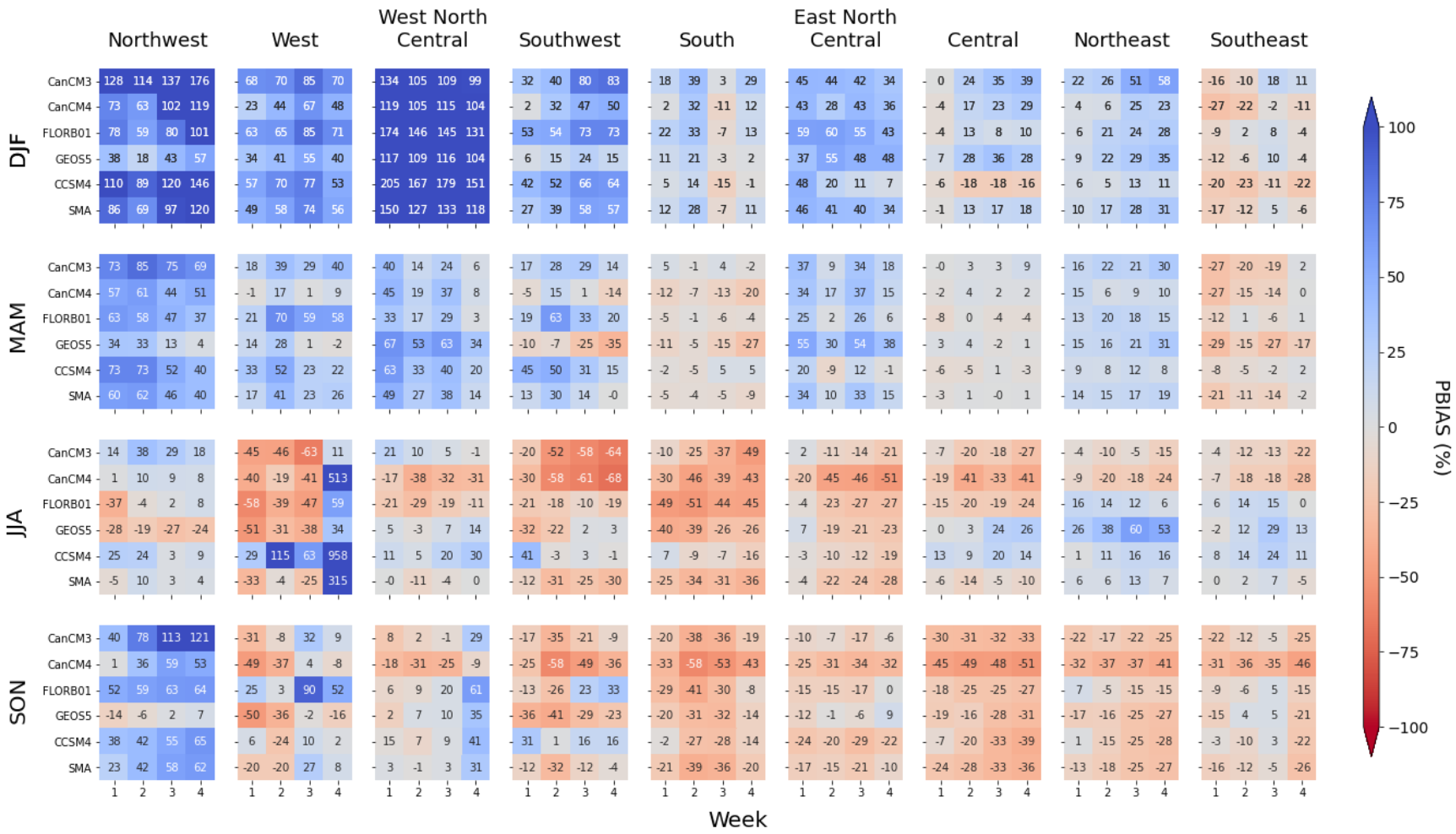

Figure 2.5 Regional averaged PBIAS of NMME-2 precipitation hindcasts from the five NMME-2 models and an SMA model for each season (DJF, MAM, JJA, and SON) and lead time (weeks 1 to 4) at the nine NCEI climate regions.

2.4.2 Predictive Skill

To evaluate the predictive skills of NMME-2 precipitation hindcasts at different lead times (weeks 1 to 4) across CONUS, I further calculate the ACC values of the ensemble means of five NMME-2 models and the SMA of the grand ensemble of NMME-2. The results are presented in Figure 2.6. Warmer colors in Figure 2.6 indicate higher ACC scores and cooler colors in Figure 2.6 indicate lower ACC scores. The red-colored numbers on the lower-right corners of each subplot in Figure 2.6 are the spatially averaged ACC value over the entire CONUS.

According to the results in Figure 2.6, except for the FLORB01, the forecast skill of all other NMME-2 models shows a decreasing trend over lead times as their CONUS-averaged ACC values decrease rapidly from week one (0.34 to 0.43) to week two (0.19 to 0.26) and remain at a marginal level at week three (0.10 to 0.17) and week four (0.09 to 0.10). The CONUS-averaged ACC values of FLORB01 are consistently lower than other NMME-2 models, especially at week one and week two. The SMA shows a higher forecast skill than any other individual NMME-2 models at all lead times over the entire CONUS.

In addition, the results in Figure 2.6 also indicate that spatially, all NMME-2 models have the highest forecast skill over the coastal areas in the Northwest and West regions as compared to other regions in CONUS. Higher ACC values can be observed in the Northwest, West, West North Central, and some parts of the Southwest Central regions, as compared with the CONUS-averaged ACC values. However, the ACC values in the South, Central, some parts of the Southwest, and Northeast regions are consistently lower than in other regions, especially at the longer lead times (week two, week three, and week four).

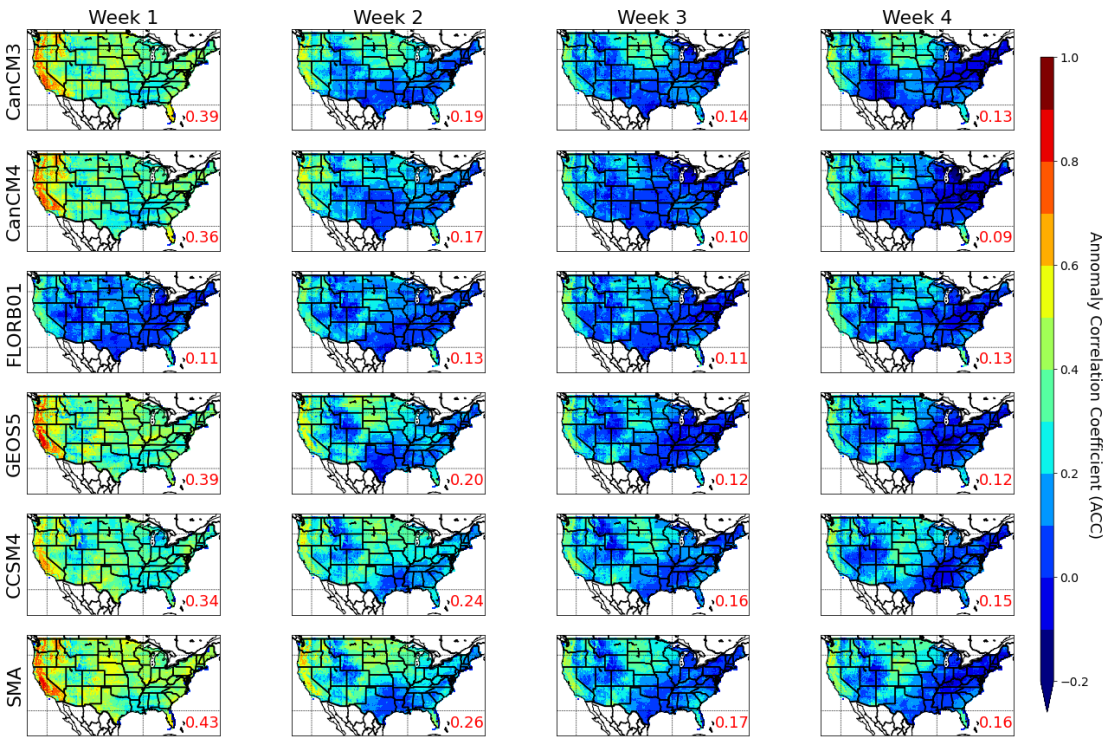

Figure 2.6 The pattern of ACC for NMME-2 precipitation hindcasts over the CONUS at different lead times (weeks 1 to 4).

In the following Figure 2.7, I present the ACC values in different seasons of (a)DJF, (b)MAM, (c)JJA, and (d)SON at different lead times (week one to week four). The difference between Figure 2.7 and Figure 2.6 is that the prior Figure 2.6 distinguishes the evaluation by forecast lead times but does not separate the forecasts made in different seasons. On the contrary, in Figure 2.7, I further group the forecasts into different seasons and lead times for evaluation. The warmer colors in Figure 2.7 indicate higher ACC scores, and the cooler colors in Figure 2.7 indicate lower ACC scores. The red-colored numbers on the lower-right corners of each subplot in Figure 2.7 are the spatially averaged ACC value over the entire CONUS.

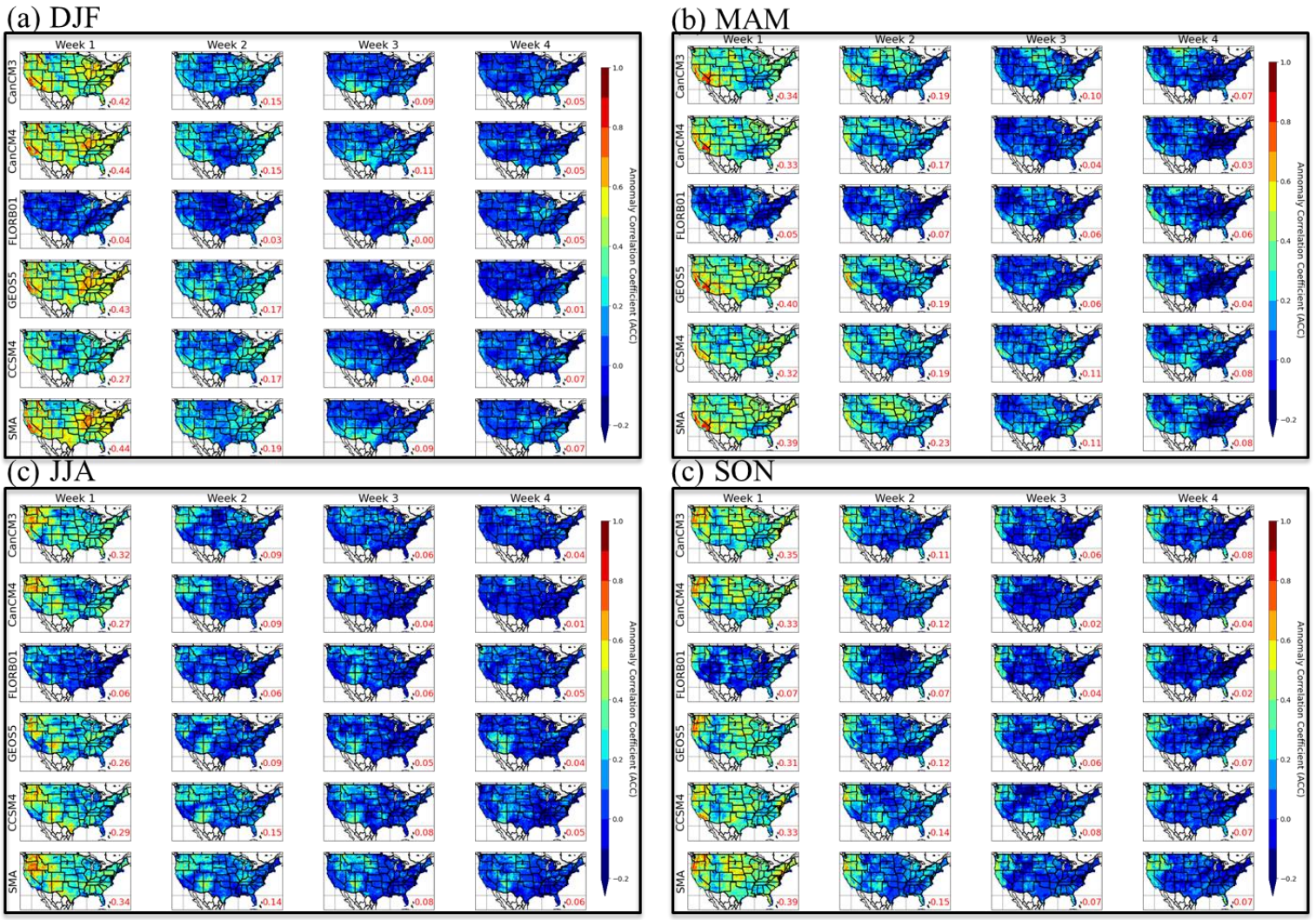

Figure 2.7 The pattern of ACC for NMME-2 precipitation hindcasts within different seasons (i.e., DJF, MAM, JJA, SON) at different lead times (weeks 1 to 4) over the CONUS.

In Figure 2.7, the weekly pattern of the ACC values in all seasons generally follows similar behaviors, as shown in Figure 2.6. The ACC values of the NMME-2 models decrease rapidly from week 1 to week 2 and show marginal forecast skill in week 3 and week 4 in all seasons. The FLORB01 model is still an outlier compared with the other NMME-2 models as it continuously presents lower CONUS-averaged ACC values. The SMA of all NMME-2 models shows the highest CONUS-averaged ACC values at almost all lead times in all seasons. It is also found that there are seasonal dependencies in the forecast skills according to the lead times. Taking SMA results as an example, the highest ACC values at week one and week three are observed in DJF and MAM, respectively. Although overall marginal forecast skill is observed at week 3 and week 4 in all seasons, a moderate forecast skill (0.3 to 0.4) can be observed in some areas within different seasons. In addition, at week three and week four, acceptable forecast skills are observed (i) over the South, Southwest, some parts of the West, the West North Central, and the Southeast in DJF; (ii) over the North Central, the South, and the coastal area of the West regions in MAM; (iii) over the Southwest and the West North Central regions in JJA; and (iv) over the Northwest, and the coastal areas of the West regions in SON, respectively.

Figure 2.8 presents the regionally averaged ACC scores of NMME-2 precipitation hindcasts in different seasons (DJF, MAM, JJA, and SON) and at different lead times (weeks 1-4) over nine NCEI climate regions. In Figure 2.8, the lighter colors indicate lower ACC values, and the darker colors indicate higher ACC values. The numbers in the figure are the actual ACC scores. Similar to the presentation of Figure 2.5, Figure 2.8 follows the pattern that the layout of nine NCEI climate regions from left to right in Figure 2.8 is corresponding to the real-world geography from western to eastern U.S. per the real-world layout.

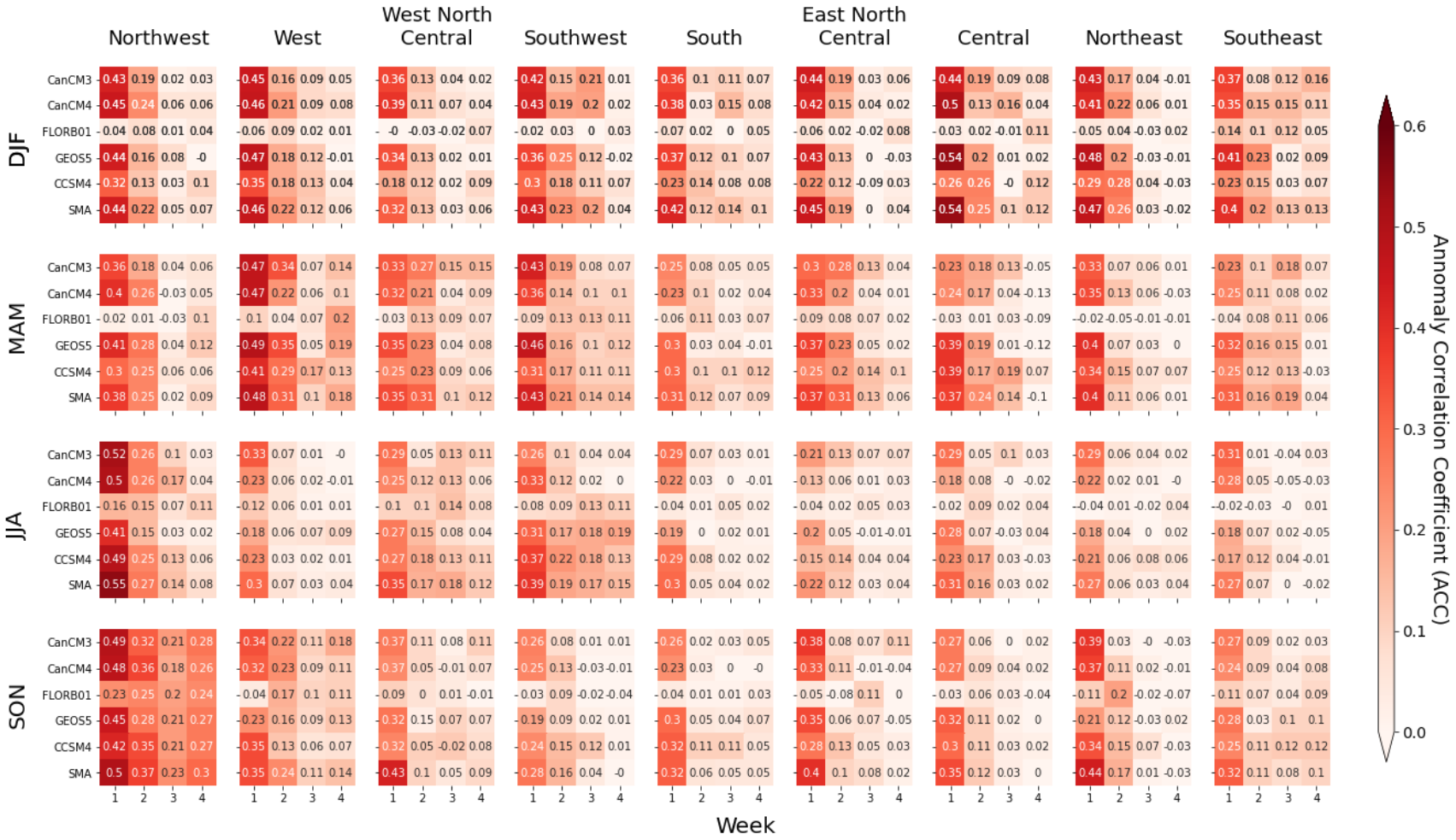

Figure 2.8 Regional averaged ACC of NMME-2 precipitation hindcasts from the five NMME-2 models and an SMA model for each season (DJF, MAM, JJA, and SON) and lead time (weeks 1 to 4) at the nine NCEI climate regions.

The results in Figure 2.8 indicate a spatially varying pattern of the precipitation forecast skill of NMME-2 models, when the seasonality and lead time dependencies are jointly considered. Specifically, in Figure 2.8, there is a decreasing trend of ACC values from the west to the east, especially in MAM and JJA. In addition, I observe that the ACC values are relatively higher at the shortest lead time (i.e., week 1), as compared to that at longer lead times (i.e., weeks 2 to 4) across all regions and seasons. According to Figure 2.8, in general, the highest overall ACC value is likely to appear in the winter season (i.e., DJF). On the contrary, the NMME-2 models tend to produce the lowest forecast skill in the summer season (i.e., JJA) except for the Northwest and Southwest regions. Except for the FLORB01 model, all other NMME-2 models show similar behaviors in most of the NECI climate regions and the different models show similar performance over most of the lead times. The SMA shows the highest ACC values than any single NMME-2 model under most of the scenarios, which is expected as the SMA approach can eliminate the outliners and produce a more conservative forecast as compared to each single NMME-2 model.

As a summary of this **Section 3.2**, the following three important findings are evidenced by the presented results: (1) the NMME-2 precipitation forecast skill consistently decreases over the lead times across the CONUS and in all seasons; (2) The raw forecast skill of NMME-2 dataset is generally higher in cooler seasons than that in warmer seasons; And (3) spatially, the raw forecast skill of all five employed NMME-2 models tends to be higher in western regions than in eastern regions across the CONUS, and the performance of different models is similar to each other, except for the FLORB01 model.

2.5 Discussion

The PBIAS result shown in prior **Section 2.4.1** can be explained along with the geographical characteristics of the CONUS and rainfall mechanisms. The overall pattern of PBIAS in the winter DJF season showed significant overestimations in the Northwest, the West, and West North Central regions. For these regions, precipitation events are dominated by synoptic-scale extratropical cyclones (ECs) and atmospheric rivers (ARs) related weather in cooler seasons (Zhang et al., 2019b). A number of previous studies have reported that GCMs tended to produce too much low-volume precipitation in comparison with reference (also known as the "drizzle effect") when simulating synoptic precipitation events (Hill, 1993; Maraun, 2013). Thus, I suspect the observed overestimations at the above-mentioned regions in cooler seasons are primarily due to the "drizzle effect". However, other studies also suggested that GCMs tended to underestimate high-volume precipitations related to convective weather systems in the Western US (Norris et al., 2021). My evaluation is also somehow limited in this regard as it only reflects aggregated bias conditions of NMME-2, which entangles all potential factors. Thus, more detailed examinations of NMME-2 in these regions should be carried out to attribute the biases more rigorously. Although significant overestimation is generally observed in the DJF, there is also an underestimation in some areas within the same regions in DJF. These mixed behaviors of over- and under-estimations could be attributed to the complex terrains brought by the Sierra Nevada Mountain and the Rocky Mountains, which are likely to trigger orographic precipitations with the nearby climatic regions. This also suggests that the raw coarse spatial resolution of NMME-2 should be considered through proper bias corrections and downscaling before any further hydrologic applications. In contrast to what has been observed in DJF, an overall underestimation of precipitation from all NMME-2 models is

observed in the South, the Southwest and part of the West North Central regions in warmer seasons, especially in JJA. These regions lie in the middle of the continent, with nearly half of the precipitation contributed by mesoscale convective systems (MCSs) in warmer seasons (Easterling et al., 2017; Fritsch et al., 1986; Nesbitt et al., 2006). But for GCMs, convection systems are neither perfectly parameterized nor resolved (Moncrieff, 2019). Moreover, MCSs with even smaller spatial scales normally operate within typical GCM grids, which surpass the capability of GCMs and ESMs (Eden et al. 2012). And for that reason, it is reasonable to suspect the observed underestimations over these regions in warmer seasons are associated with sub-grid convective precipitation events.

For forecast skills, one major discrepancy was identified between FLORB01 and other NMME-2 models, as shown in Figure 2.6. The FLORB01 model showed consistently lower forecast skill throughout all lead times, while other NMME-2 models show a higher forecast skill at week 1 and rapidly decreases to marginal levels at week 3 and week 4. This result agrees with the previous evaluation study on monthly NMME-1 precipitation forecasts by Slater et al. (2019), in which the authors found the FLORB01 model sometimes does not display higher skill at the shortest lead time (one month). The obtained result in this chapter further shows that within the one-month lead time (i.e., week one to week four), the FLORB01 model still presents poorer performance compared with other NMME models. My results (Figure 2.9) also exhibit that the application of the grand ensemble with larger ensemble sizes generally shows better statistics than individual NMME-2 models with smaller ensemble sizes. It strongly supports that creating a grand model ensemble with bigger ensemble sizes through multiple techniques (e.g., multi-model, perturbed physics, and perturbed initial conditions, etc.) is an effective way to increase forecast skill and forecast reliability.

The spatial dependencies of the forecast skill of all employed NMME-2 models are evident. Higher forecast skills are observed over the coastal areas of the Northwest, the West, and West North Central regions at all lead times. Within these regions, the winter precipitation is mainly associated with cyclonic (synoptic) scale weather (Cayan and Roads, 1984). Because of the life cycle of cyclones and their large spatial scales, the corresponding precipitation events are generally easier for GCMs to predict compared to other types of precipitation events (i.e., convective and orographic) (Kumar et al., 2011; Zhu et al., 2014). Even at longer lead times (weeks two to week four), the forecast skills in these regions are still higher than that in other regions. It is also noticeable that the areas and seasons showing significant overestimations seem more likely to achieve higher forecast skills if cross examining the results presented in Figure 2.5 and Figure 2.8.

Such results indicate many challenges remain when applying the NMME-2 S2S precipitation forecast to hydrologic modeling and real-world water resources planning. The first issue that needs to be addressed is the substantial forecast bias. Although my result has demonstrated the good potential of NMME-2 in predicting extreme events, the raw forecast values need to be bias-corrected before assisting real-world flood predictions (Brown and Seo, 2010; Tiwari et al., 2021). Currently, most of the popular bias correction approaches heavily rely on the "stationary assumption", where statistical moments are generalized from the historical records and will be used to correct the raw forecast values (Teutschbein and Seibert, 2012). However, some recent studies have identified trends in not only frequencies but also magnitudes of extreme precipitation events in various regions over the globe (Madsen et al., 2014; Miao et al., 2015; Sun et al., 2014), which likely undermines the efficacy of existing approaches in bias correcting extreme precipitation forecast values. And the authors believe that

properly considering the “non-stationary” in bias correction, especially for extreme values, might be a major challenge for future studies.

Efforts have also been made to improve the marginal forecast skill of the S2S precipitation forecasts in order to utilize hydrometeorological forecasts from multiple sources and models to assist ensemble hydrologic forecasting. I believe that one promising direction is to construct a super NMME-2 ensemble, and further apply more sophisticated multi-model averaging techniques to improve the forecast data quality (Ji et al., 2020; Sloughter et al., 2007; Yang et al., 2018). Multi-model ensemble with proper data quality control could further increase the precipitation forecast skill across different temporal and spatial resolutions. I also observe that in most cases, the SMA method can produce slightly higher forecast skills than any individual NMME-2 models at all lead times. Nevertheless, the improvements of SMA are still limited to individual models. I also believe that advanced model ensemble technique needs to be used together with the fundamental enhancements of the GCMs and ESMs, providing better precipitation predictability and less system and random errors in the S2S forecasts. For example, this could be done by advancing the current land surface components of GCMs (Dirmeyer et al., 2018; Zhou et al., 2020) or through better sub-grid convective parameterizations (Eden et al., 2012).

From the perspective of the hydrology community, exploiting the benefits of deep learning techniques to post-process S2S precipitation forecasts might be another feasible alternative (Akbari Asanjan et al., 2018; Weyn et al., 2021), given the slow advances in fundamentally enhancing the physical dynamics of coupled GCMs or ESMs. Most of the previous studies have tried to correct the bias of GCM-generated precipitation forecast separately without considering the forecast skill. However, the result of my study has shown forecast bias and predictive skill

are somehow connected with each other. The performance of precipitation forecasts can essentially be attributed to different climate and weather patterns and mechanisms (Eden et al., 2012; Kirtman et al., 2014). Only correcting bias may lead to the degradation of forecast skills (Mendoza et al., 2017). Recently, it has been reported that taking advantage of deep learning techniques to handle additional resolved atmospheric forecast variables to improve the S2S precipitation forecast skills first and then correcting the bias may provide some major improvements (Miao et al., 2019; Pan et al., 2019).

Another limitation of this chapter is the use of statistical measures. In this chapter, I mainly use three measurements to compare the NMME-2 forecasts, namely, the PBIAS, ACC, QPOD, and QFAR. These three selected statistical indices measure different aspects of the forecast data quality, and they reflect different characteristics of precipitation events. The number of statistical measurements used in this chapter is still limited, and there are many other popular metrics were not examined in this chapter, such as the Mean Squared Error, the Ranked Probabilistic Skill Score (RPSS), the Continuous RPSS (CRPSS), and other Categorical skill metrics, including the False Alarm Ratio (FAR), the Probability of Detection (POD), the Critical Success Index (CSI), and Equitable Threat Score (ETC), etc. I suggest other practitioners evaluate as many skill scores and metrics as possible to comprehensively reveal the pros and cons of the S2S forecast data.

For hydrologists, it is also important to apply the improved and corrected S2S precipitation forecasts for hydrologic forecasts at watersheds with different hydrometeorological conditions (Li et al., 2019). By doing so, more detailed examinations of S2S precipitation forecasts on a daily basis can be conducted. This would overcome the limitation of my present study, in which only the NMME-2 S2S forecasts on a weekly basis are studied. One of the major motivations

for my study is to identify trends and quantify the data quality of the NMME-2 forecasts over the entire CONUS, and therefore, the streamflow simulation capabilities and forecast data improvements will be investigated in future studies. Based on the obtained results in this chapter, I expect that it is promising to use machine learning methods to improve the NMME-2 S2S daily precipitation forecasts. These new machine learning approaches allow auxiliary information to be considered during the bias-correction process, such as the forecast lead times, seasonality, regional factors, and relevant atmospheric forecast variables, which are all found to be important information in improving the accuracy of S2S precipitation forecasts.

To summarize, based on the presented studies, I strongly believe future studies on NMME-2 S2S precipitation forecasts may include (1) proper bias corrections and downscaling of the NMME-2 S2S precipitation forecasts at different regions across the CONUS; (2) application of deep learning approaches to provide more accurate and reliable NMME-2 precipitation forecasts, especially at longer lead times; (3) hydrologic applications of the NMME-2 S2S precipitation forecasts with ESP framework to further investigate the efficacy of NMME-2 S2S precipitation forecasts over traditional historical resample forecasts in streamflow forecasting.

2.6 Conclusion

In this chapter, the S2S precipitation forecasts of NMME-2 are comprehensively evaluated across CONUS and during a hindcast period of 1982-2011 by computing forecast bias (PBIAS) and forecast skill (ACC). The spatial, seasonal, and lead time dependence of the performance of NMME-2 daily precipitation forecasts has been analyzed over nine NCEI climate regions. My results in this Chapter highlight the strengths and weaknesses of the NMME-2 S2S precipitation forecast and its potential for hydrologic applications. The major findings and conclusions of this chapter are summarized below.

1. The NMME-2 S2S precipitation forecasts show substantial biases across CONUS. The forecast biases also demonstrate strong spatial and seasonal dependencies, but I found the model biases are not sensitive to forecast lead times. Five selected NMME-2 models and their SMA show similar spatial bias patterns across the CONUS. Based on my spatial analysis, significant overestimations of the NMME-2 forecasts are observed in the western (Northwest, West, and West North Central) regions in DJF season. And significant model underestimations are identified in the South region of CONUS in warmer seasons of JJA and SON.
2. In terms of forecast skill, a major discrepancy exists between the FLORB01 and the rest of NMME-2 models. Specifically, the FLORB01 model consistently shows a lower precipitation forecast skill, especially at week-1 and week-2 lead times, as compared to other NMME-2 models. The forecast skill of the rest of NMME-2 models is the highest at week-1, tends to decrease rapidly from week-1 to week-2, and remains at a marginal level at week-3 and week-4 across all regions and in all seasons. Spatially, all NMME-2 models show higher forecast skills in the western areas (Northwest and West regions).
3. The SMA of five NMME-2 members shows better deterministic forecast skill than any single NMME-2 model under most comparison scenarios. Thus, I believe more sophisticated model averaging techniques could further increase the deterministic forecast skills of NMME-2.

Acknowledgments

The material is based upon work supported by the National Science Foundation under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3, and the National Science Foundation under Grant No. NSF1802872. This work is partially supported by the U.S. Department of

Energy (DOE Prime Award # DE-IA0000018) and the Natural Science Foundation of Jiangsu Province, China (BK20180403).

References

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J. and Peng, Q. 2018. Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. Journal of Geophysical Research: Atmospheres 123(22), 12,543-512,563.

Alley, R.B., Emanuel, K.A. and Zhang, F. 2019. Advances in weather prediction. Science 363(6425), 342-344.

Ashfaq, M., Rastogi, D., Mei, R., Kao, S.C., Gangrade, S., Naz, B.S. and Touma, D. 2016. High-resolution ensemble projections of near-term regional climate over the continental United States. Journal of Geophysical Research: Atmospheres 121(17), 9943-9963.

Asoka, A. and Mishra, V. 2015. Prediction of vegetation anomalies to improve food security and water management in India. Geophysical Research Letters 42(13), 5290-5298.

Baker, S.A., Wood, A.W. and Rajagopalan, B. 2019. Developing subseasonal to seasonal climate forecast products for hydrology and water management. JAWRA Journal of the American Water Resources Association 55(4), 1024-1037.

Bauer, P., Thorpe, A. and Brunet, G. 2015. The quiet revolution of numerical weather prediction. Nature 525(7567), 47-55.

Becker, E., den Dool, H.v. and Zhang, Q. 2014. Predictability and forecast skill in NMME. Journal of Climate 27(15), 5891-5906.

Becker, E., Kirtman, B.P. and Pegion, K. 2020. Evolution of the North American multi-model ensemble. Geophysical Research Letters 47(9), e2020GL087408.

Berner, J., Ha, S.-Y., Hacker, J., Fournier, A. and Snyder, C. 2011. Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. Monthly Weather Review 139(6), 1972-1995.

Brown, J.D. and Seo, D.-J. 2010. A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. Journal of Hydrometeorology 11(3), 642-665.

Cash, B.A., Manganello, J.V. and Kinter, J.L. 2019. Evaluation of NMME temperature and precipitation bias and forecast skill for South Asia. Climate dynamics 53(12), 7363-7380.

Cayan, D.R. and Roads, J.O. 1984. Local relationships between United States West Coast precipitation and monthly mean circulation parameters. Monthly Weather Review 112(6), 1276-1282.

Chelton, D.B. and Wentz, F.J. 2005. Global microwave satellite observations of sea surface temperature for numerical weather prediction and climate research. Bulletin of the American Meteorological Society 86(8), 1097-1116.

Cohen, J., Foster, J., Barlow, M., Saito, K. and Jones, J. 2010. Winter 2009–2010: A case study of an extreme Arctic Oscillation event. Geophysical Research Letters 37(17).

Daly, C. and Bryant, K. 2013. The PRISM climate and weather system—an introduction. Corvallis, OR: PRISM climate group.

Day, G.N. 1985. Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management 111(2), 157-170.

Delworth, T.L., Rosati, A., Anderson, W., Adcroft, A.J., Balaji, V., Benson, R., Dixon, K., Griffies, S.M., Lee, H.-C. and Pacanowski, R.C. 2012. Simulated climate and climate change in the GFDL CM2. 5 high-resolution coupled climate model. Journal of Climate 25(8), 2755-2781.

Dirmeyer, P.A., Chen, L., Wu, J., Shin, C.-S., Huang, B., Cash, B.A., Bosilovich, M.G., Mahanama, S., Koster, R.D. and Santanello, J.A. 2018. Verification of land–atmosphere coupling in forecast models, reanalyses, and land surface models using flux site observations. Journal of hydrometeorology 19(2), 375-392.

Easterling, D.R., Arnold, J., Knutson, T., Kunkel, K., LeGrande, A., Leung, L.R., Vose, R., Waliser, D. and Wehner, M. 2017. Precipitation change in the United States.

Eden, J.M., Widmann, M., Grawe, D. and Rast, S. 2012. Skill, correction, and downscaling of GCM-simulated precipitation. Journal of Climate 25(11), 3970-3984.

Fritsch, J., Kane, R. and Chelius, C. 1986. The contribution of mesoscale convective weather systems to the warm-season precipitation in the United States. Journal of Applied Meteorology and Climatology 25(10), 1333-1345.

Gobena, A. and Gan, T. 2010. Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system. Journal of hydrology 385(1-4), 336-352.

Guo, Y. and Nie, H. 2020. Summertime daily precipitation statistics over East China in CFSv2. Physics and Chemistry of the Earth, Parts A/B/C 115, 102841.

Guo, Z., Dirmeyer, P.A. and DelSole, T. 2011. Land surface impacts on subseasonal and seasonal predictability. Geophysical Research Letters 38(24).

Hamill, T.M. and Juras, J. 2006. Measuring forecast skill: Is it real skill or is it the varying climatology? Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 132(621C), 2905-2923.

Hill, C.D. 1993. Forecast problems in the western region of the National Weather Service: An overview. Weather and forecasting 8(2), 158-165.

Ji, L., Zhi, X., Simmer, C., Zhu, S. and Ji, Y. 2020. Multimodel ensemble forecasts of precipitation based on an object-based diagnostic evaluation. Monthly Weather Review 148(6), 2591-2606.

Karl, T. and Koss, W.J. 1984. Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983.

Khajehei, S., Ahmadalipour, A. and Moradkhani, H. 2018. An effective post-processing of the North American multi-model ensemble (NMME) precipitation forecasts over the continental US. Climate dynamics 51(1), 457-472.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P. and Becker, E. 2014a. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bulletin of the American Meteorological Society 95(4), 585-601.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M.P., Becker, E., Peng, P., Tripp, P., Huang, J., DeWitt, D.G., Tippett, M.K., Barnston, A.G., Li, S., Rosati, A., Schubert, S.D., Rienecker, M., Suarez, M., Li, Z.E., Marshak, J., Lim, Y.-K., Tribbia, J., Pegion, K., Merryfield, W.J., Denis, B. and Wood, E.F. 2014b. The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. Bulletin of the American Meteorological Society 95(4), 585-601.

Krakauer, N.Y. 2019. Temperature trends and prediction skill in NMME seasonal forecasts. Climate Dynamics 53(12), 7201-7213.

Kumar, A., Chen, M. and Wang, W. 2011. An analysis of prediction skill of monthly mean climate variability. Climate dynamics 37(5), 1119-1131.

Li, W., Chen, J., Li, L., Chen, H., Liu, B., Xu, C.-Y. and Li, X. 2019. Evaluation and bias correction of S2S precipitation for hydrological extremes. Journal of Hydrometeorology 20(9), 1887-1906.

Lin, C., Vasić, S., Kilambi, A., Turner, B. and Zawadzki, I. 2005. Precipitation forecast skill of numerical weather prediction models and radar nowcasts. Geophysical research letters 32(14).

Madsen, H., Lawrence, D., Lang, M., Martinkova, M. and Kjeldsen, T. 2014. Review of trend analysis and climate change projections of extreme precipitation and floods in Europe. Journal of Hydrology 519, 3634-3650.

Maraun, D. 2013. Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. Journal of Climate 26(6), 2137-2143.

Mendoza, P.A., Wood, A.W., Clark, E., Rothwell, E., Clark, M.P., Nijssen, B., Brekke, L.D. and Arnold, J.R. 2017. An intercomparison of approaches for improving operational seasonal streamflow forecasts. Hydrology and Earth System Sciences 21(7), 3915-3935.

Merryfield, W.J., Lee, W.-S., Boer, G.J., Kharin, V.V., Scinocca, J.F., Flato, G.M., Ajayamohan, R., Fyfe, J.C., Tang, Y. and Polavarapu, S. 2013. The Canadian seasonal to interannual prediction system. Part I: Models and initialization. Monthly weather review 141(8), 2910-2945.

Miao, C., Ashouri, H., Hsu, K.-L., Sorooshian, S. and Duan, Q. 2015. Evaluation of the PERSIANN-CDR daily rainfall estimates in capturing the behavior of extreme precipitation events over China. Journal of Hydrometeorology 16(3), 1387-1396.

Miao, Q., Pan, B., Wang, H., Hsu, K. and Sorooshian, S. 2019. Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network. Water 11(5), 977.

Moncrieff, M.W. 2019. Toward a dynamical foundation for organized convection parameterization in GCMs. Geophysical Research Letters 46(23), 14103-14108.

Murphy, A.H. and Epstein, E.S. 1989. Skill scores and correlation coefficients in model verification. Monthly weather review 117(3), 572-582.

Nardi, K.M., Baggett, C.F., Barnes, E.A., Maloney, E.D., Harnos, D.S. and Ciasto, L.M. 2020. Skillful all-season S2S prediction of US precipitation using the MJO and QBO. Weather and Forecasting 35(5), 2179-2198.

Nesbitt, S.W., Cifelli, R. and Rutledge, S.A. 2006. Storm morphology and rainfall characteristics of TRMM precipitation features. Monthly Weather Review 134(10), 2702-2721.

Norris, J., Hall, A., Chen, D., Thackeray, C.W. and Madakumbura, G.D. 2021. Assessing the Representation of Synoptic Variability Associated With California Extreme Precipitation in CMIP6 Models. Journal of Geophysical Research: Atmospheres 126(6), e2020JD033938.

Oubeidillah, A.A., Kao, S.-C., Ashfaq, M., Naz, B.S. and Tootle, G. 2014. A large-scale, high-resolution hydrological model parameter data set for climate change impact assessment for the conterminous US. Hydrology and Earth System Sciences 18(1), 67.

Palmer, T., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delécluse, P., Déqué, M., Diez, E., Doblas-Reyes, F.J. and Feddersen, H. 2004. Development of a European

multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bulletin of the American Meteorological Society 85(6), 853-872.

Palmer, T.N. 2002. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 128(581), 747-774.

Pan, B., Hsu, K., AghaKouchak, A. and Sorooshian, S. 2019a. Improving precipitation estimation using convolutional neural network. Water Resources Research 55(3), 2301-2321.

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S. and Higgins, W. 2019b. Precipitation prediction skill for the West Coast United States: From short to extended range. Journal of Climate 32(1), 161-182.

Prat, O. and Nelson, B. 2015. Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). Hydrology and Earth System Sciences 19(4), 2037.

Radcliffe, D. and Mukundan, R. 2017. PRISM vs. CFSR precipitation data effects on calibration and validation of SWAT models. JAWRA Journal of the American Water Resources Association 53(1), 89-100.

Ritter, B. and Geleyn, J.-F. 1992. A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. Monthly weather review 120(2), 303-325.

Sankarasubramanian, A., Lall, U., Souza Filho, F.A. and Sharma, A. 2009. Improved water allocation utilizing probabilistic climate forecasts: Short-term water contracts in a risk management framework. Water Resources Research 45(11).

Shrestha, D., Robertson, D., Wang, Q., Pagano, T. and Hapuarachchi, H. 2013. Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. Hydrology and Earth System Sciences 17(5), 1913-1931.

Shukla, S., Roberts, J., Hoell, A., Funk, C.C., Robertson, F. and Kirtman, B. 2019. Assessing North American multimodel ensemble (NMME) seasonal forecast skill to assist in the early warning of anomalous hydrometeorological events over East Africa. Climate Dynamics 53(12), 7411-7427.

Slater, L.J., Villarini, G. and Bradley, A.A. 2017. Weighting of NMME temperature and precipitation forecasts across Europe. Journal of Hydrology 552, 646-659.

Slater, L.J., Villarini, G. and Bradley, A.A. 2019. Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA. Climate Dynamics 53(12), 7381-7396.

Sloughter, J.M.L., Raftery, A.E., Gneiting, T. and Fraley, C. 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Monthly Weather Review 135(9), 3209-3220.

Sorooshian, S., AghaKouchak, A., Arkin, P., Eylander, J., Foufoula-Georgiou, E., Harmon, R., Hendrickx, J.M., Imam, B., Kuligowski, R. and Skahill, B. 2011. Advanced concepts on remote sensing of precipitation at multiple scales. Bulletin of the American Meteorological Society 92(10), 1353-1357.

Stockdale, T.N., Molteni, F. and Ferranti, L. 2015. Atmospheric initial conditions and the predictability of the Arctic Oscillation. Geophysical Research Letters 42(4), 1173-1179.

Sun, J., Xue, M., Wilson, J.W., Zawadzki, I., Ballard, S.P., Onvlee-Hooimeyer, J., Joe, P., Barker, D.M., Li, P.-W. and Golding, B. 2014a. Use of NWP for nowcasting convective precipitation: Recent progress and challenges. Bulletin of the American Meteorological Society 95(3), 409-426.

Sun, Q., Kong, D., Miao, C., Duan, Q., Yang, T., Ye, A., Di, Z. and Gong, W. 2014b. Variations in global temperature and precipitation for the period of 1948 to 2010. Environmental monitoring and assessment 186(9), 5663-5679.

Teutschbein, C. and Seibert, J. 2012. Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. Journal of hydrology 456, 12-29.

Thomas, J.A., Berg, A.A. and Merryfield, W.J. 2016. Influence of snow and soil moisture initialization on sub-seasonal predictability and forecast skill in boreal spring. Climate dynamics 47(1), 49-65.

Tiwari, A.D., Mukhopadhyay, P. and Mishra, V. 2021. Influence of bias correction of meteorological and streamflow forecast on hydrological prediction in India. Journal of Hydrometeorology.

Vernieres, G., Rienecker, M.M., Kovach, R. and Keppenne, C.L. 2012. The GEOS-iODAS: Description and evaluation.

Vertenstein, M., Craig, T., Middleton, A., Feddema, D. and Fischer, C. 2010. CCSM4. 0 User's Guide. National Center for Atmospheric Research, Boulder, 70pp.[Available online at.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E. and Fuentes, M. 2017. The subseasonal to seasonal (S2S) prediction project database. Bulletin of the American Meteorological Society 98(1), 163-173.

Vitart, F. and Robertson, A.W. 2018. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. npj Climate and Atmospheric Science 1(1), 1-7.

Wanders, N. and Wood, E.F. 2016. Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. Environmental Research Letters 11(9), 094007.

Weyn, J.A., Durran, D.R., Caruana, R. and Cresswell-Clay, N. 2021. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. arXiv preprint arXiv:2102.05107.

White, C., Franks, S. and McEvoy, D. 2015. Using subseasonal-to-seasonal (S2S) extreme rainfall forecasts for extended-range flood prediction in Australia. Proceedings of the International Association of Hydrological Sciences 370, 229-234.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J. and Murray, V. 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorological applications 24(3), 315-325.

Widmann, M. and Bretherton, C.S. 2000. Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States. Journal of Climate 13(11), 1936-1950.

Wood, A.W. and Lettenmaier, D.P. 2006. A test bed for new seasonal hydrologic forecasting approaches in the western United States. Bulletin of the American Meteorological Society 87(12), 1699-1712.

Xu, L., Chen, N., Zhang, X., Chen, Z., Hu, C. and Wang, C. 2019. Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning. Climate dynamics 53(1), 601-615.

Yang, T., Tao, Y., Li, J., Zhu, Q., Su, L., He, X. and Zhang, X. 2018. Multi-criterion model ensemble of CMIP5 surface air temperature over China. Theoretical and Applied Climatology 132(3), 1057-1072.

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T. and Gailhard, J. 2012. Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. Advances in Science and Research 8(1), 135-141.

Zhang, Z., Ralph, F.M. and Zheng, M. 2019. The relationship between extratropical cyclone strength and atmospheric river intensity and position. Geophysical Research Letters 46(3), 1814-1823.

Zhou, T., Chen, Z., Zou, L., Chen, X., Yu, Y., Wang, B., Bao, Q., Bao, Y., Cao, J. and He, B. 2020. Development of climate and earth system models in China: Past achievements and new CMIP6 results. Journal of Meteorological Research 34(1), 1-19.

Zhou, Y. and Kim, H.-M. 2018. Prediction of atmospheric rivers over the North Pacific and its connection to ENSO in the North American multi-model ensemble (NMME). Climate Dynamics 51(5), 1623-1637.

Zhu, H., Wheeler, M.C., Sobel, A.H. and Hudson, D. 2014. Seamless precipitation prediction skill in the tropics and extratropics from a global model. Monthly Weather Review 142(4), 1556-1569.

Chapter 3. An Alternative Ensemble Streamflow Prediction (ESP) Approach Using Improved Subseasonal-to-Seasonal (S2S) Precipitation Forecasts from the North America Multi-Model Ensemble Phase II (NMME-2) Dataset

This chapter has been submitted in its current form to the *Journal of Hydrology* and is now under review.

Zhang, L., Yang, T., Gao, S., Hong, Y., Fan, M., Dan L., Xu, H & Xiao C. (2021). An alternative Ensemble Streamflow Prediction (ESP) approach using improved Subseasonal-to-Seasonal (S2S) precipitation forecasts from the North America Multi-Model Ensemble Phase II (NMME-2) dataset. *Journal of Hydrology*, (under review).

Abstract

Subseasonal-to-Seasonal (S2S) precipitation forecast products are useful tools to assist water and energy resource planning. Most of the S2S forecasts are made available by collaborations among many mission agencies all over the globe. However, a substantial amount of forecast bias and a marginal level of predictive skills are commonly presented in many existing S2S precipitation forecast datasets. Previous studies found that in the South Climate Region of the U.S., the available S2S precipitation performs consistently worse than in other climate regions, which raises the question of whether S2S precipitation could lead to better streamflow and flood forecasting within this challenging region from the practical use perspective. In this chapter, I carried out a series of streamflow hindcast experiments using four S2S precipitation forecast products from North America Multi-Model Ensemble Phase II (NMME-2) at four selected watersheds in the South climate region of the U.S. The performance of S2S precipitation is benchmarked against the historically resampled precipitation under the ensemble streamflow prediction (ESP) framework. Evaluation statistics of the Kling-Gupta Efficiency (KGE), probability of detections (POD), and false alarm ratios (FAR) are employed for a comprehensive evaluation of the predictive performance of the resampled and S2S

precipitation in streamflow hindcasts. The results show that, compared to the resampled precipitation, S2S precipitation leads to overall better streamflow predictions. Evaluation of daily extreme streamflow values within different quantile segments (i.e., 75%-90%, 90-95%, 95%-99%, and above 99%) also suggests the overall advantages of S2S precipitation in improving streamflow predictions. In addition, a few high-volume extreme streamflow events were selected and plotted as time series for visual inspections, which confirms the effectiveness of S2S precipitation in streamflow predictions. Overall, the S2S precipitation has shown promising performance, highlighting its potential broader applications for river and flood forecasting under the classical ESP framework.

3.1 Introduction

Skillful streamflow predictions at Subseasonal-to-Seasonal (S2S) timescales, defined as 10 to 30 days into the future (Vitart et al., 2017), could provide crucial information and benefit human socio-economic activities, including reservoir operations, irrigation planning, proactive flood mitigations, etc. (Welles et al., 2007; White et al., 2017; Yang et al., 2021).

To perform streamflow predictions at S2S timescale, the classical ensemble streamflow prediction (ESP) approach is the most commonly adopted (Day, 1985) approach, and it is also the standard guidance for streamflow forecasting used by the National Weather Center and its 13 River Forecast Centers over the United States (Schaake and Larson, 1998). The classical ESP approach consists of the following three major steps. First, a calibrated hydrologic model will be forced with measured precipitation for a hydrologic simulation up to the “current” time (i.e., forecast initializing time) as a first step. By doing so, the initial hydrologic conditions (IHCs, i.e., mainly soil moisture and/or snow condition) of the watershed are estimated. Then, a series of precipitation measurements are randomly selected from historical records to represent

possible future precipitation scenarios. Last, an ensemble of streamflow predictions is generated through a combination of the calibrated hydrologic model, the estimated IHCs, and the randomly resampled precipitation forecasts from historical measurements.

The key advantage of the ESP framework is to consider the predictability of streamflow from both the IHCs and future precipitation scenarios. But more importantly, by employing ensemble predictions generated with resampled precipitation forecasts, ESP can predict future extreme events (i.e., flooding events) probabilistically. This is because, under the classical "stationary" assumption in hydrology, it is believed that historical precipitation records should repeat themselves and reflect future extreme streamflow scenarios.

However, the quality of ESP-generated streamflow forecasts at S2S timescale is usually compromised. This is because the predictability of streamflow from the estimated IHCs cannot sustain over S2S timescale, leaving streamflow predictions dominated by the inaccurate resampled precipitation at an extended forecast lead times (Cao et al., 2021; Shukla et al., 2013; Wood and Lettenmaier, 2008; Yossef et al., 2013). As a result, ESP cannot guarantee accurate predictions of the mid-volume streamflow, but it can only provide probabilistic predictions upon extreme events with its ensemble spreads. Moreover, under the context of global climate change, the resampled precipitation forecasts might not lead to reasonable representations of future extreme precipitation scenarios (Miao et al., 2016; Sun et al., 2018), which undermines the fundamentals of ESP in predicting extreme events (i.e., the "stationary" assumption in hydrology). Therefore, more advanced forecast products that provide reliable and accurate S2S precipitation forecasts are critically needed under the ESP framework.

In the recent decade, dynamical precipitation forecast products from a series of coupled General Circulation Models (GCMs) have been made available by mission agencies across

different countries (Kirtman et al., 2014; Pegion et al., 2019; Vitart et al., 2017). These available S2S precipitation forecast products are generated by coupled GCMs that simulate the evolution of the "current" state of the Earth System through a consideration of dynamic land-ocean-atmosphere interactions. Therefore, in theory, available S2S precipitation forecast products should provide more accurate and reliable precipitation predictions at S2S timescale, which could enable overall superior streamflow forecasts at S2S timescale under the ESP framework.

However, a series of evaluation studies have reported the inconsistent performance of available S2S precipitation forecast products. It is commonly recognized that available S2S precipitation forecasts only show a marginal level of forecast skill after week 2 and generally have a substantial amount of forecast bias at different locations over the globe. More recently, two studies by Zhang et al. (2021) and showed similar results with the S2S precipitation forecasts from the North America Multi-Model Ensemble Phase II (NMME-2) over the contiguous United States (CONUS). More specifically, according to the spatial analysis conducted by Zhang et al. (2021) and Zhang et al. (2023b) based on the climatic region partition by National Center for Environmental Information (NCEI), the S2S precipitation forecasts from NMME-2 perform consistently worse in the NCEI South climate region (including Oklahoma, Texas, Arkansas, and Louisiana) than in other regions of the United States.

Given such compromised performance of available S2S precipitation forecast products, especially in the NCEI South climate region, it appears questionable whether available S2S precipitation forecasts lead to better streamflow predictions under the ESP framework than benchmark forecasts with resampled precipitation. Therefore, the following research questions

are raised: ***(1) How do available S2S precipitation forecasts perform at watersheds in the Southern region of the United States, in comparison to the resampled precipitation forecasts? (2) Can available S2S precipitation forecasts lead to overall better streamflow predictions in comparison to the original ESP with resampled precipitation forecasts? And more specifically, (3) Can the S2S precipitation forecast outperform the resampled precipitation under the ESP framework in terms of predicting extreme streamflow values at a daily resolution?***

To answer these research questions, in this chapter, I carried out a series of ESP hindcast experiments at S2S timescales with both resampled precipitation and S2S precipitation forecasts at four selected watersheds in the NCEI South climate region of the United States. This chapter covers a 30-year period from 1982 to 2011. Four state-of-the-art S2S precipitation forecast products retrieved from the NMME-2 project were tested out regarding their streamflow forecast capabilities, following the raw forecast validation study by Zhang et al. (2021) over the CONUS. To overcome the coarse spatial resolution as well as the forecast bias that is commonly presented in precipitation forecast datasets, I adopted a modified version of the statistical downscaling and bias correction (BCSD) method introduced by Wood et al. (2004a). The Kling-Gupta Efficiency (KGE) was adopted for the evaluation of overall precipitation and streamflow forecast skills at different forecast lead times. The streamflow forecast skills upon daily extreme streamflow values within 75%-90%, 90%-95%, 95%-99%, and above 99% were examined with Probability of Detections (POD) and False Alarm Ratios (FAR). In addition, a few specific high-percentile events were selected and plotted as time series for visual inspections at the selected watersheds to evaluate the capability of S2S precipitation forecast in extreme streamflow predictions more directly.

The remainder of this **Chapter 3** is organized as follows: **Section 3.2** introduces the datasets used in this chapter. **Section 3.3** presents the methodology of ESP, hydrologic model, BCSD, as well as employed evaluation statistics. **Sections 3.4** and **3.5** present the result and discussion of this chapter.

3.2 Datasets and Study Regions

3.2.1 Datasets

Three measurement/reanalysis hydrometeorological datasets used in this chapter include the following (1) daily AN81d precipitation dataset from the Parameter-elevation Regressions on the Independent Slopes Model (PRISM); (2) daily potential evapotranspiration (PET) reanalysis dataset from the North American Regional Reanalysis (NARR); and (3) daily streamflow measurements from the National Water Information System at the United States Geological Survey (USGS).

In addition, S2S precipitation hindcasts (hereafter referred to as "S2S precipitation") from four contributing models to the NMME-2 data archive are used in this chapter. Four S2S precipitation datasets are from the Canadian Climate Model versions 3 and 4 (CanCM3 and CanCM4) from Environment and Climate Change Canada (Kirtman and Min, 2009; Merryfield et al., 2013), the Community Climate System Model 4.0 (CCSM4) from National Center for Atmospheric Research (Gent et al., 2011), and the Goddard Earth Observing System model version 5 (GEOS5) from NASA (Borovikov et al., 2019b). These four S2S forecast products have the same hindcast frequency and have shown relatively better predictive skills over CONUS (Zhang et al., 2021). Each of the selected S2S forecast models generates ensemble precipitation hindcasts containing 10 ensemble members with either perturbed physics strategy and/or different initial conditions (Kirtman et al., 2014). All the collected S2S

precipitation hindcasts have a monthly frequency and a daily temporal resolution. Each of the collected S2S precipitation hindcasts is initialized/issued on the 1st day of each month during the study period, providing daily hindcast values in a gridded format over CONUS with forecast lead times up to 365 days. More detailed information about the corresponding datasets is presented in Table 3.1. The S2S precipitation is later spatially downscaled to each study watershed using PRISM as the ground truth reference. Both S2S precipitation and the NARR PET datasets are transformed into the mean areal values over the study watersheds, and they serve as the inputs to the selected hydrologic model.

Table 3.1 Hydrometeorological datasets used in Chapter 3.

	Name	Spatial resolution	Sources
Measurements /Reanalysis	Precipitation (PRISM)	~0.07°	https://prism.oregonstate.edu/
	PET (NARR)	~0.3°	https://www.emc.ncep.noaa.gov/mmb/rreanl/
	Streamflow (USGS)	N/A	https://waterdata.usgs.gov/nwis
GEOS5 S2S hindcasts	Precipitation	1°	https://www.cpc.ncep.noaa.gov/products/NMME/data.html
	Temperature	1°	
	500 hPa Geopotential	1°	
	850 hPa Geopotential	1°	

3.2.2 Target Watersheds

In this chapter, four watersheds in the South climate region of the United States (per NCEI definition) are selected as the experiment watersheds. Figure 3.1 shows the locations as well as their corresponding elevation and river networks of the four selected watersheds.

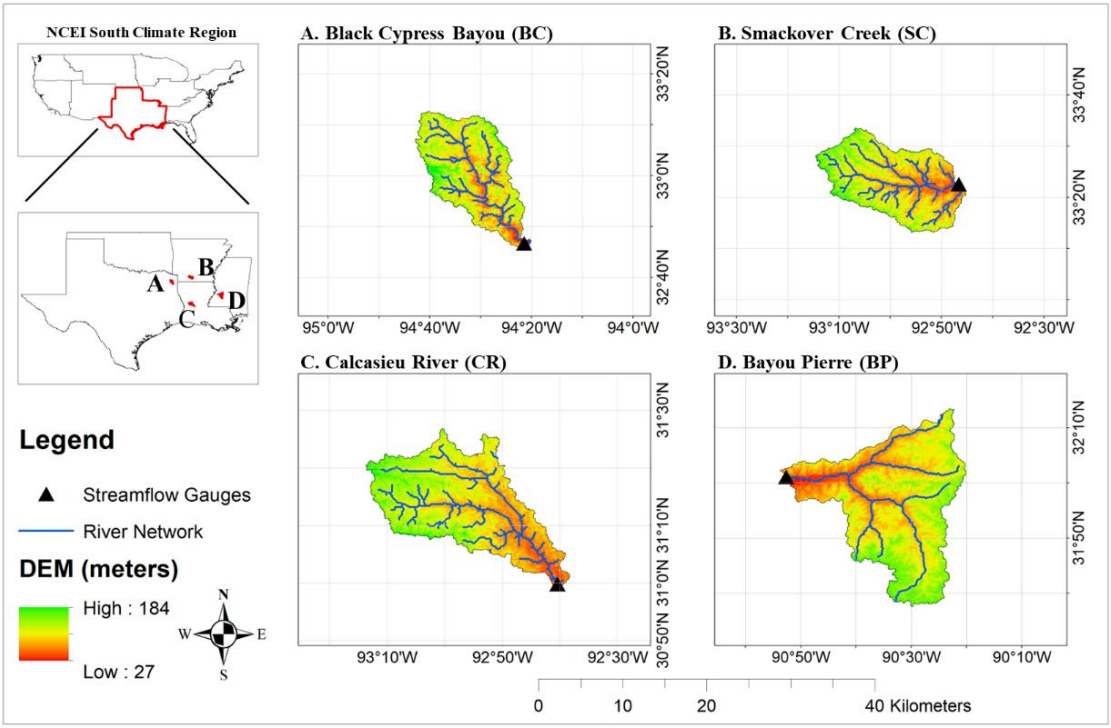

Figure 3.1 Map of the 4 study watersheds (BC is in Texas; SC is in Arkansas; CR is in Louisiana; BP is in Missouri).

The four selected watersheds are also included in the CAMELS dataset (Newman et al., 2015) and considered less affected by human activities, which makes them suitable study watersheds in the region. According to the information provided in the CAMELS dataset, these four watersheds are dry, with a major proportion of precipitation occurring in cold seasons and little snow (less than 2% of the total amount of precipitation). More detailed information on the study watersheds is presented in Table 2.2.

Table 3.2 Topography and hydrometeorology information of the study watersheds.

Watershed Names	Drainage Area (Sq-KM)	Elevation (Meters)	Aridity (PET/P)	Snow Fraction	USGS Gauge ID
Black Cypress Bayou (BC)	959.89	105.07	0.88	0.017	7346045
Smackover Creek (SC)	996.08	79.29	0.80	0.018	7362100
Calcasieu River (CR)	1293.69	81.79	0.75	0.003	8013000
Bayou Pierre (BP)	1688.73	102.05	0.82	0.004	7290650

3.3 Methodology

3.3.1 Experiment Design

This chapter aims to evaluate the hydrologic performance of S2S precipitation in streamflow forecasts at the selected watersheds in the NCEI South climate region against the resampled precipitation under the ESP framework. Therefore, I carry out a series of streamflow hindcast experiments with resampled precipitation and compared the streamflow simulation accuracy using the S2S precipitation under the ESP framework during a study period from 1981 to 2011 (30 years). The overall experiment design is depicted in Figure 3.2. Since this chapter focuses on S2S timescale, all the hindcast experiments are limited to the forecast horizon with the forecast lead times ranging from 1 to 28 days. Subjecting to the frequency of collected S2S precipitation datasets, the corresponding streamflow hindcast experiments are also initialized at the 1st days of each month during the study period (i.e., 12×30=360 hindcast experiments).

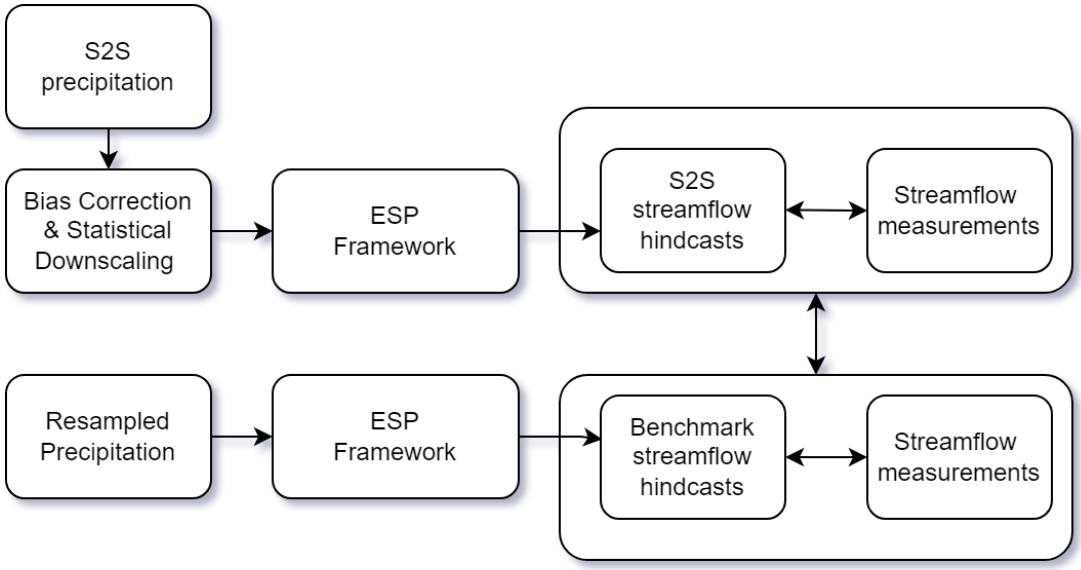

Figure 3.2 A schematic diagram of the experiment design.

To apply the ESP framework, a chosen hydrologic model needs to be calibrated at the study watersheds first. Then, the calibrated hydrologic models are forced with hydrometeorological

measurements up to the "current time" to get the best estimations of initial hydrologic conditions (IHCs) of the watersheds. This process is also known as the "spin-up". Lastly, streamflow predictions are generated through a combination of the calibrated hydrologic model, the estimated IHCs, and the S2S or resampled precipitation forecasts. I order to focus on evaluating the effectiveness of S2S precipitation in streamflow forecasts, I use the 30-year potential evapotranspiration (PET) climatology derived from the NARR reanalysis dataset for all hindcast experiments at each study watershed for the hydrologic model spin-up and streamflow hindcasts.

The selected hydrologic model for testing the ESP framework is the classical Sacramento Soil Moisture Accounting (Sac-SMA) model. The Sac-SMA model is a classical lumped hydrologic model, and it is the primary rainfall-runoff model used for operational forecasting in the National Weather Service (NWS). The Sac-SMA model has a total of 13 tunable parameters and incorporates a two-story soil moisture structure (see **Appendix A** for more detailed information). The Sac-SMA model was calibrated at the four selected watersheds with the SC-SAHEL optimization algorithm developed by Naeini et al. (2018), which included the SCE-UA algorithm (Duan et al. 1992). In all watershed studies, the first 15 years (1981-1996) of the entire data record is used for model calibration, and the remaining 15 years (1996-2011) data record is used for model validation. With calibrated Sac-SMA models at each selected watershed, 1-year "spin-up" simulations were conducted before each hindcast experiment to estimate the IHCs of watersheds at the hindcast initialization dates.

Two sets of precipitation hindcasts are prepared at each watershed for conducting the ESP experiment. The first set of precipitation hindcasts uses the resampled precipitation used to benchmark the predictive skill of streamflow at S2S timescale under the default/standard ESP

scenario. In other words, to generate month X's precipitation forecast in a certain year K, all month X's precipitation measurements from the collected 30 year's PRISM precipitation, excluding year K's, are used to create an ensemble of precipitation forecasts containing 29 ensemble members. The second set of precipitation hindcasts is the S2S precipitation from the NMME-2 S2S precipitation forecasts data archive. Due to the original coarse spatial resolution and the existing forecast bias (Zhang et al. 2021), S2S precipitation cannot be applied directly at typical USGS HUC-8 watershed scales. Therefore, in this chapter, I further apply and test out a slightly modified version of Bias Correction and Statistical Downscaling (BCSD), which was originally introduced by Wood et al. (2004a). The BCSD approach has been demonstrated to be an effective post-processing tool in removing forecast bias (Maurer and Hidalgo, 2008; Werner and Cannon, 2016; Zhang et al., 2019a). Both resampled precipitation and the S2S precipitation (after BCSD) will be transformed into mean areal values over the study watersheds and drive the Sac-SMA hydrologic model to produce streamflow forecasts.

Finally, comparisons are made between streamflow hindcasts using the resampled precipitation and the S2S precipitation forecasts at the selected watersheds. The overall predictive skills at different forecast lead times are quantified with the Kling-Gupta Efficiency (KGE). In addition, I also conduct comparison of the capability of both resampled precipitation and the S2S precipitation in predicting the occurrence of high-volume streamflow values at different percentiles segments, including 75% to 90%, 90% to 95%, 95% to 99%, and above 99%. The evaluation of extreme values is carried out with statistics of Probability of Detections (POD), and False Alarm Ratios (FAR).

3.3.2 Evaluation Statistics

3.3.2.1 Kling-Gupta Efficiency (KGE)

The Kling-Gupta Efficiency (KGE) was originally introduced by Gupta et al. (2009) and has become a widely applied performance evaluation statistic in the field of hydrology ever since. The KGE is computed with three distinct statistics of correlation coefficient (CC), Bias Ratio (BR), and relative variability (RV) altogether. The values of CC, BR, and RV can be computed with the following equations (3.1), (3.2), and (3.3):

$$CC = \frac{\sum_{i=1}^{n}\left((Q_{Sim,i}-\overline{Q}_{Sim})-(Q_{Obs,i}-\overline{Q}_{Obs})\right)}{\sqrt{\sum_{i=1}^{n}(Q_{Sim,i}-\overline{Q}_{Sim})^2 \sum_{i=1}^{n}(Q_{Obs,i}-\overline{Q}_{Obs})^2}} \quad (3.1)$$

$$BR = \frac{\overline{Q}_{Sim}}{\overline{Q}_{Obs}} \quad (3.2)$$

$$RV = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{Sim,i}-\overline{Q}_{Sim})^2}/\overline{Q}_{Sim}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{Obs,i}-\overline{Q}_{Obs})^2}/\overline{Q}_{Obs}} \quad (3.3)$$

In the equations above, n is the total number of data points, $Q_{Sim,i}$ and $Q_{Obs,i}$ are the model simulated/predicted data and observed data, and $\overline{Q}_{Sim}$ and $\overline{Q}_{Obs}$ are the means of the model simulated/predicted and observed data series, respectively. With the computed values of CC, BR, and RV, KGE can be computed with the following equation (3.4):

$$KGE = 1 - \sqrt{(CC-1)^2 + (BR-1)^2 + (RV-1)^2} \quad (3.4)$$

The values of KGE range from $-\infty$ to 1. The ideal value of KGE would be 1, indicating that model simulated/predicted values are perfectly aligned with measurements.

2.3.2.2 Probability of Detection (POD) and False Alarm Ratio (FAR)

Table 3.3 describes the categorical relationships between the predicted and the measured extreme values. In Table 3.3, the occurrence of certain types of extreme streamflow is categorized in a binary form, which either happened or did not happen. The “H” (i.e., hits) in the table represents the total cases of successful predictions on extreme streamflow. The “M”

(i.e., misses) represents the total cases of non-forecasted extreme streamflow. The "F" (i.e., false alarms) represents the total cases of false positive forecasts on the extreme streamflow.

Table 3.3 Contingency table of all possible outcomes for categorical forecasts of binary events.

Events	O	NO
P	H	F
NP	M	CR

Utilizing Table 3.3, POD and FAR can be computed with the following equation (3.5) and (3.6):

$$POD = H/(H + M) \quad (3.5)$$

$$FAR = F/(F + H) \quad (3.6)$$

The values of POD range from 0 to 1, with 1 being the ideal value indicating that all extreme streamflow is predicted. The values of FAR range from 0 to 1 as well, but with 0 being the ideal value, which suggests that there are no false alarms upon the extreme streamflow events at all.

3.4 Results

3.4.1 Calibration of Sac-SMA

Figure 3.3 presents the streamflow simulations at 4 study watersheds after the calibrations of Sac-SMA model. The red lines indicate streamflow measurements, and the blue lines indicate the simulated streamflow. The vertical dashed lines separate the entire study period into the calibration and validation periods. The KGE values within calibration and validation periods at each study watershed are labeled on the figure as well.

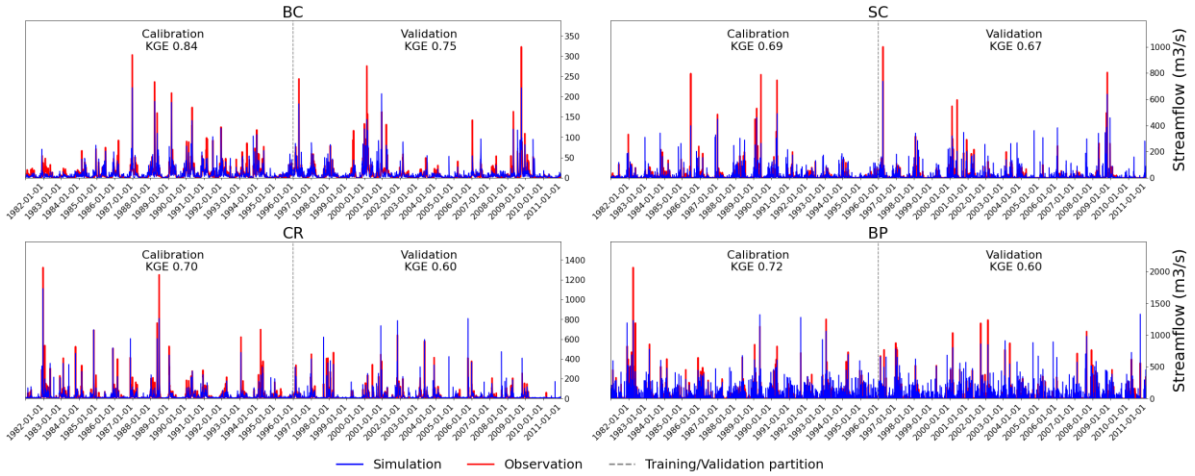

Figure 3.3 Simulated and observed daily streamflow hydrographs at the study watersheds.

From Figure 3.3, it can be observed that even after the calibration of Sac-SMA, the high-volume streamflow events seem to be consistently underestimated in all watersheds. But in general, KGE value ranges from 0.84 to 0.69 at all watersheds during the calibration period. Although the performance of hydrologic simulations drops during the validation periods, the values of KGE value still remain above 0.6 at all study watersheds.

3.4.2 Forecast Bias and Predictive Skill of Precipitation

In this chapter, I examine the forecast bias and skills of all S2S precipitation ensemble members before and after BCSD. Figure 3.4 presents the cumulative distribution functions (CDF) of mean areal precipitations over the 4 study watersheds in logarithmic scales. Each column of Figure 3.4 presents the CDFs of S2S precipitation from different forecast models over an individual study watershed. Each row of Figure 3.4 presents the CDFs of S2S precipitation from a certain forecast model but at different study watersheds. The red lines indicate the CDFs of reference precipitation. The grey lines indicate the CDF of raw S2S precipitation. The blue lines indicate the CDF of S2S precipitation after the BCSD.

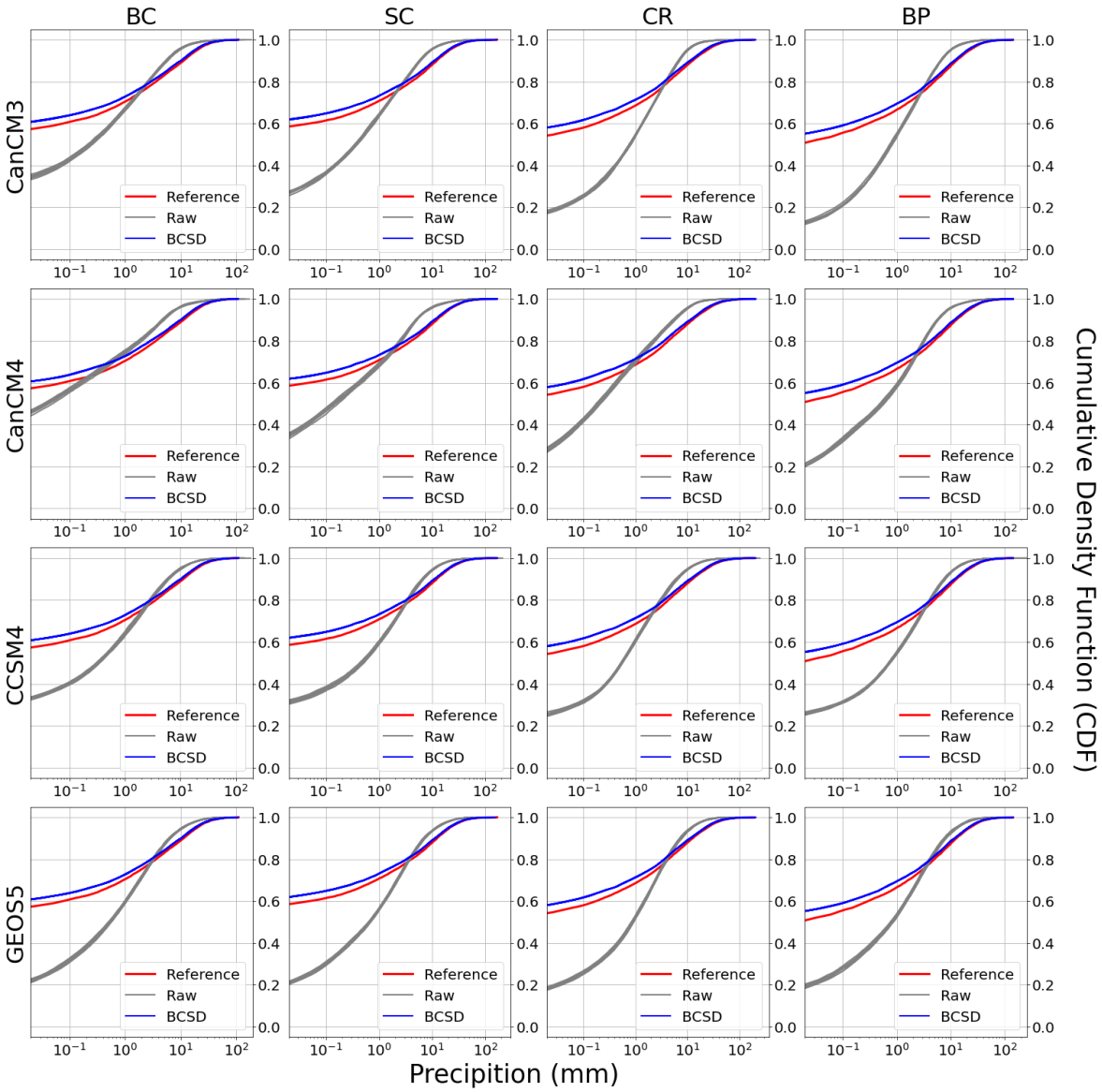

Figure 3.4 Cumulative distribution functions of the S2S precipitation before and after BCSD at four study watersheds.

In Figure 3.4, the raw precipitation (grey lines) in all panels lie below 'BCSD' or 'Reference' near 10^{-2} mm, indicating the higher occurrence of positive rain by the raw S2S precipitation values. In all examined cases, it can be observed that the CDFs of S2S precipitation become much closer to the reference CDFs after BCSD compared to that of the raw S2S precipitation, indicating the removal of much of the bias. In addition, it seems that the bias is more effectively removed for precipitation within the range of [1 mm, 10 mm] if

compared to the precipitation quantile smaller than 1 mm. However, it is difficult to further observe the bias condition of precipitation beyond 10 mm from Figure 3.4 as it is plotted in a logarithmic scale.

Therefore, to further verify the effectiveness of BCSD at high percentiles, I further plot an additional set of Quantile-Quantile plots (QQ-plots) in regular scales, as shown in Figure 3.5. In Figure 3.5, the mean areal precipitations of all ensemble members of S2S precipitation from NMME-2 over the 4 study watersheds are plotted as scatter points corresponding to the reference precipitation at the same quantile. Each column of Figure 3.5 presents the QQ-plots of S2S precipitation from different forecast models over a certain study watershed. Each row of Figure 3.5 presents QQ-plots of S2S precipitation from a certain forecast model but at different study watersheds. In Figure 3.5, grey points indicate raw S2S precipitation before BCSD and blue points indicate the S2S precipitation after BCSD. The ideal forecast values would lie perfectly on the red-colored 45-degree lines, which indicates that there's no bias of S2S precipitation at all quantiles.

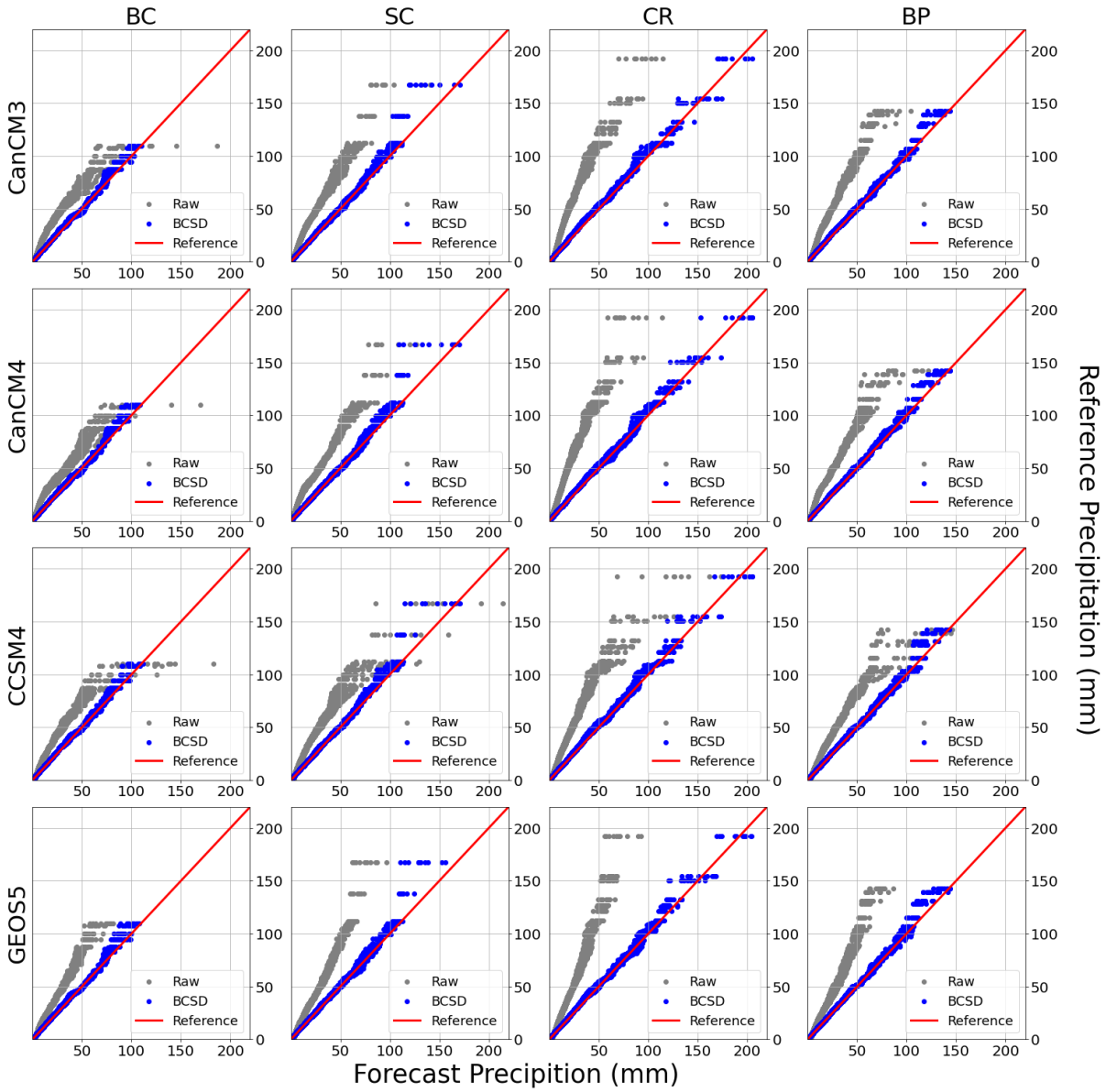

Figure 3.5 QQ-plots of the mean areal S2S precipitation before and after BCSD.

According to Figure 3.5, S2S precipitation from different forecast models tends to perform similarly over the same watersheds. However, the difference can be observed between QQ-plots at different watersheds. For example, the maximum reference precipitation at the BC watershed is no larger than 120 mm/day, whereas precipitation is measured as high as 200 mm/day at the CR watershed. But in general, at all study watersheds, the originally observed forecast bias is very nicely removed by BCSD for most of the cases. This can be observed as the S2S

precipitation scatter points lie either right on or much closer to the 45-degree lines after BCSD. Nevertheless, according to Figure 3.5, forecast bias is still preserved at high quantiles even after BCSD at certain study watersheds and for certain forecast models. For example, the scatter points of S2S precipitation at the SC watershed still deviate from the reference 45-degree lines by some distance at high-volume precipitations within the range of 100 mm to 150 mm.

In addition to the forecast bias, the predictive skills of the mean areal values of the S2S precipitation from different forecast models are presented in Figure 3.6. The predictive skill is measured by KGE over different study watersheds and at different forecast lead times. In Figure 3.6, the grey lines indicate the benchmark predictive skills of the resampled precipitation at different study watersheds. Other colored lines indicate the predictive skills of the ensemble means of S2S precipitation from various forecast models and at different forecast lead times.

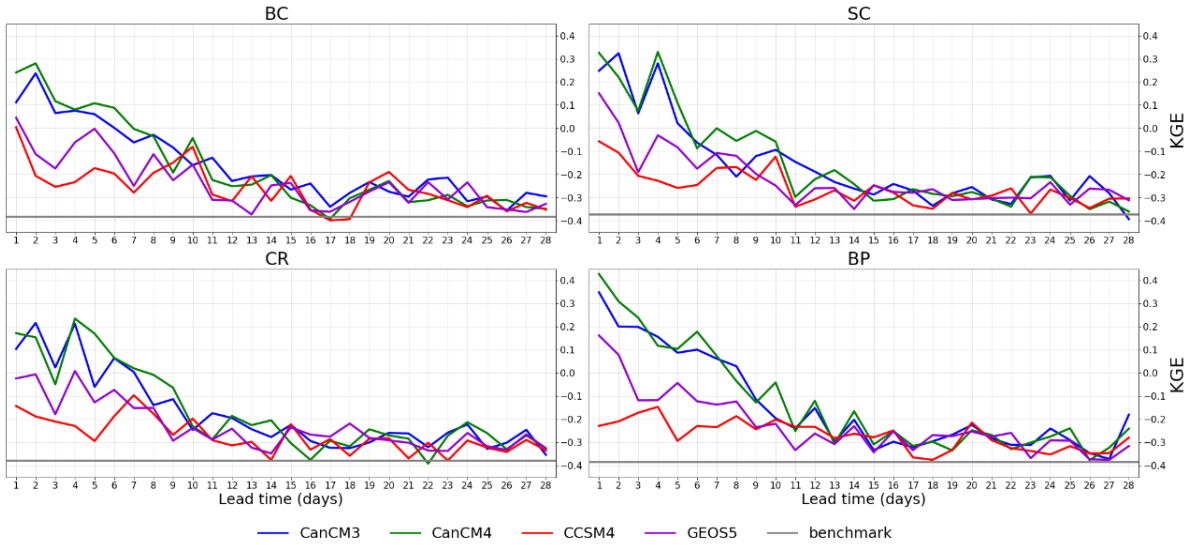

Figure 3.6 Predictive skills of S2S precipitation and resampled precipitation over forecast lead times across four study watersheds.

Figure 3.6 also shows that the skill curves of the resampled precipitation are flat over the forecast horizon as they are independent of forecast lead times. The predictive skills of the resampled precipitation are around ~0.4 (KGE) across all study watersheds, which agrees with

previous studies and presents the skill of climatology forecasts (Knoben et al., 2019). On the other hand, the predictive skills of S2S precipitation generally show a downward trend over forecast lead times. But even at the forecast lead times beyond 10 days, S2S precipitation still shows noticeably higher predictive skill in general when compared to the benchmarks. Specifically, CanCM3 and CanCM4 provide precipitation forecasts with better predictive skills at all study watersheds within ~ 14 days of forecast lead times (~ 2 weeks). But at longer lead times, there are no significant differences in terms of the overall predictive skills between S2S precipitation from different forecast models at all study watersheds.

3.4.3 Predictive Skill of Streamflow

With both resampled precipitation and BCSD-corrected S2S precipitation, streamflow hindcast experiments were carried out under the ESP framework at the selected study watersheds. The resulting overall predictive skills of streamflow hindcasts at different forecast lead times are measured with KGE and presented in **Section 2.4.3.1**. The predictive performance upon extreme streamflow values at different forecast lead times is measured with POD and FAR and presented in **Section 2.4.3.2**.

2.4.3.1 Predictive Skill of Streamflow

Figure 3.7 presents the predictive skills of ensemble means obtained from both S2S precipitation and resampled precipitation under the ESP framework at the four selected study watersheds. The grey lines are the benchmark predictive skill at different forecast lead times resulted from the resampled precipitation. Other different colored lines are the predictive skill of ensemble means resulted from the S2S precipitation from various forecast models.

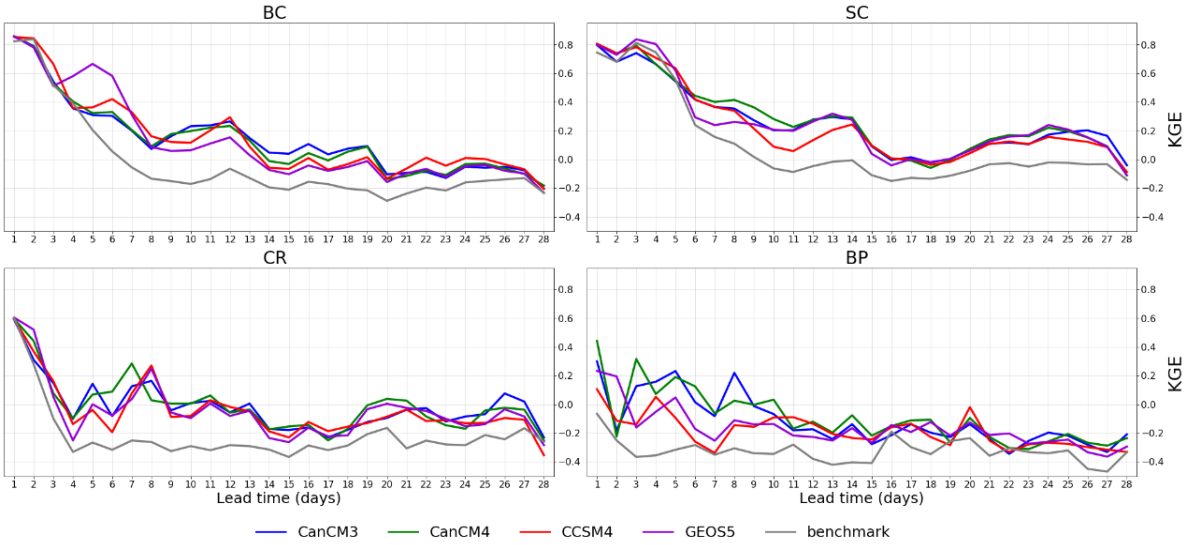

Figure 3.7 Predictive skills of streamflow with S2S precipitation and resampled precipitation over forecast lead times across four study watersheds.

The results presented in Figure 3.7 are generally consistent with previous results from Figure 3.6 (predictive skill of precipitation). It can be observed that the predictive skills of streamflow decrease over forecast lead times rapidly. However, in contrary to previous results from Figure 5 (predictive skill of precipitation), there are no significant differences between the predictive skills from the S2S and resampled precipitation at very short lead times (within ~4 days) across different study watersheds, except for at the BP watershed. Instead, the S2S precipitation leads to an overall higher predictive skill at longer forecast lead times after around 5 days. Although this advantage of the S2S precipitation over the benchmarking resampled precipitation at longer lead times is not significant, they are rather consistent across different study watersheds.

2.4.3.2 Predictive Performance on Extreme Values

Figure 3.8 presents the hindcast performance on daily extreme values within different percentile segments and over different study watersheds. In Figure 3.8, each row represents a certain study watershed, and each column represents the extreme events within a certain

percentile segment (i.e., 75%-90%, 90%-95%, 95-99%, and above 99%). Within each subplot of Figure 3.8, blue-colored scatter points indicate the performance of streamflow hindcasts associated with the usage of S2S precipitation, and grey-colored scatter points indicate the performance of streamflow hindcasts associated with the use of resampled precipitation. Different markers indicate the hindcast performance at different forecast lead times within 4 weeks. Within each subplot, each scatter point is associated with a POD value and a FAR value representing the corresponding predictive performance of S2S/resampled precipitation in predicting the extreme values within a certain quantile segment, at a certain lead time, and over a certain study watershed. The red colored boxes are used as labels if S2S precipitation outperforms resampled precipitation with not only higher POD but also smaller FAR.

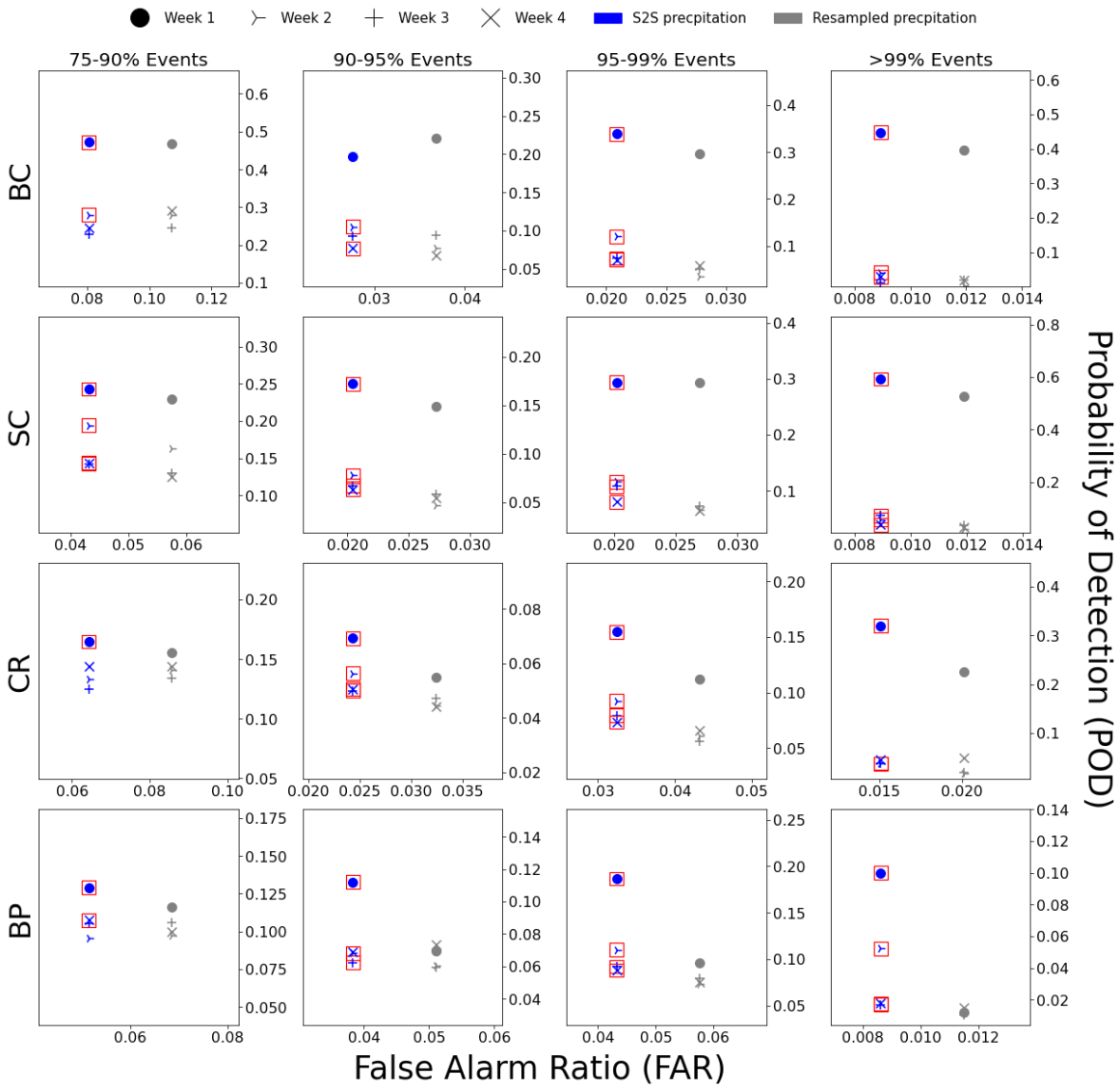

Figure 3.8 Predictive performance on daily extreme values within different quantile segments, at different lead times, and across four study watersheds. Red boxes highlight overall better performance of S2S precipitation under different scenarios.

In general, S2S precipitation forecasts provide overall better predictions of extreme streamflow events if compared to the benchmarking resampled precipitation with 52 out a total of 64 scenarios (~ 81%) being labeled with red boxes in Figure 3.8. In addition, all the blue-colored scatter points lie left to the scatter points in grey color, which indicates that the S2S precipitation led to smaller FAR under all evaluation scenarios.

However, such an advantage is less significant when predicting extreme events within the percentile segment of 75-90%. For extreme events within the percentile segment of 75%-90%, S2S precipitation can only provide overall better categorical predictions within 2 weeks (14 days) under most scenarios. Once forecast lead time exceeds 2 weeks, S2S precipitation would occasionally lead to slightly lower POD than the resampled precipitation.

But for all other high percentile values above 90%, S2S precipitation outperforms resampled precipitation in general. Only a few exceptional cases exist where S2S precipitation does not present higher POD than that of the resampled precipitation (e.g., 90%-95% at week 1 in BC watershed, above 99% in week 3 at BC watershed, and above 99% in week 4 at CR watershed). In general, such an advantage of S2S precipitation in predicting extreme events above 90% is somewhat consistent at different forecast lead times and over different study watersheds, although sometimes only by a marginal level.

3.4.4 Examination of Hydrographs from Selected Extreme Events

In addition to previous statistics employed for evaluating the forecast performance on extreme events, a few high-volume streamflow events above 90% are selected and plotted as time series for visual inspections in the following Figures 3.9 to 3.12.

The peak streamflow of selected the selected event for the BC watershed happened on the 6th day in November 2002 and is presented in Figure 3.9. Figure 3.9 contains 4 subplots, the red-colored lines in all 4 subplots indicate the streamflow measurements within the month. The 2 subplots on the left column present all streamflow forecast members generated with the resampled (benchmark) and S2S precipitation, respectively. The 2 subplots on the right column also present the forecast performance of the resampled (benchmark) and S2S

precipitation but showing the range and interquartile range (difference between 75th and 25th percentiles) with different shaded color and in logarithmic scales.

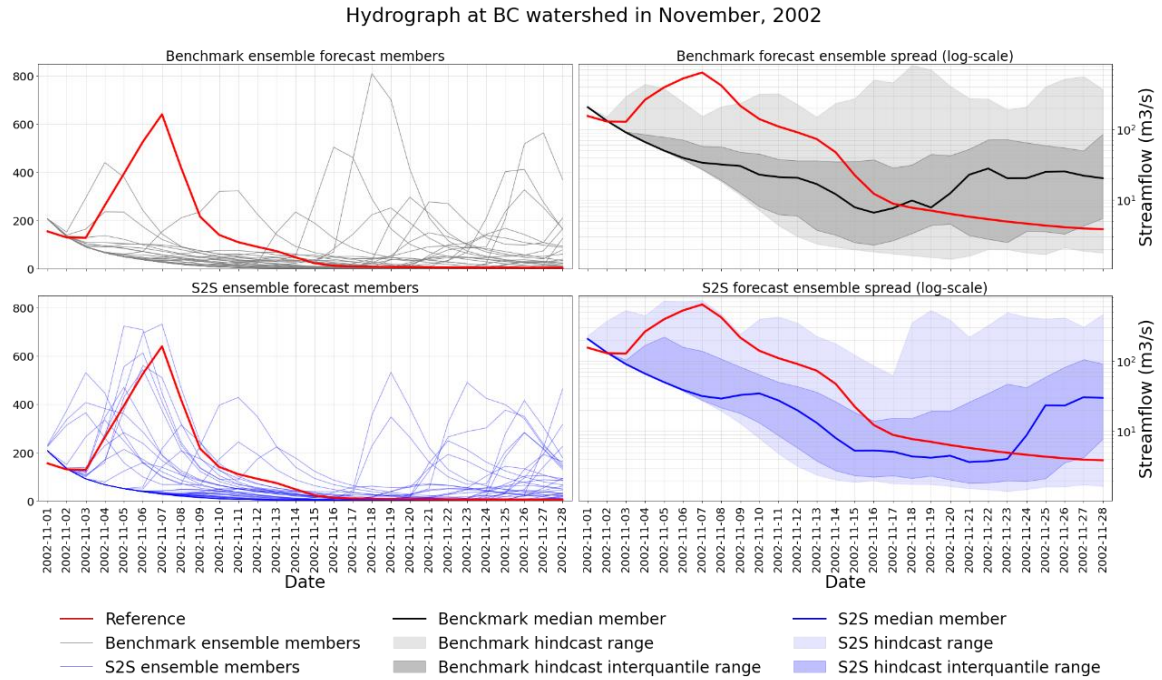

Figure 3.9 Predicted and measured daily hydrograph in November 2002 at the BC watershed.

From Figure 3.9, it can be clearly observed that multiple hindcast ensemble members obtained from the S2S precipitation have made close predictions upon the peak streamflow. As contrast, the streamflow hindcast generated with the resampled precipitation (benchmark) has totally missed the event. The prediction range plots from Figure 3.9 show very clearly that the streamflow measurements, especially the peak volume, are very nicely enveloped by the ensemble hindcasts generated with S2S precipitation in contrast to the benchmark performance.

Figure 3.10 presents the same elements as shown in Figure 3.9 but depicts the hydrograph of the SC watershed in the October of 2009. For the SC watershed, the peak of the selected extreme streamflow event happened on the 15th day of Oct 2009. For this event, only one hindcast member generated with S2S precipitation has made a prediction that is close to the peak streamflow but still with two days offset before the occurrence of the peak. However, the

streamflow hindcast generated with the resampled precipitation (benchmark) has missed the peak streamflow volume entirely. Moreover, the prediction range plots from Figure 3.10 show that the streamflow hindcasts generated with S2S precipitation have predicted higher volume streamflow with its interquartile range compared to that of the benchmark. Although the streamflow hindcasts generated with the S2S precipitation and the resampled precipitation have both missed peak streamflow, the S2S precipitation leads to closer volumetric prediction and expressed a higher chance of having large volume streamflow compared to that of the resampled precipitation.

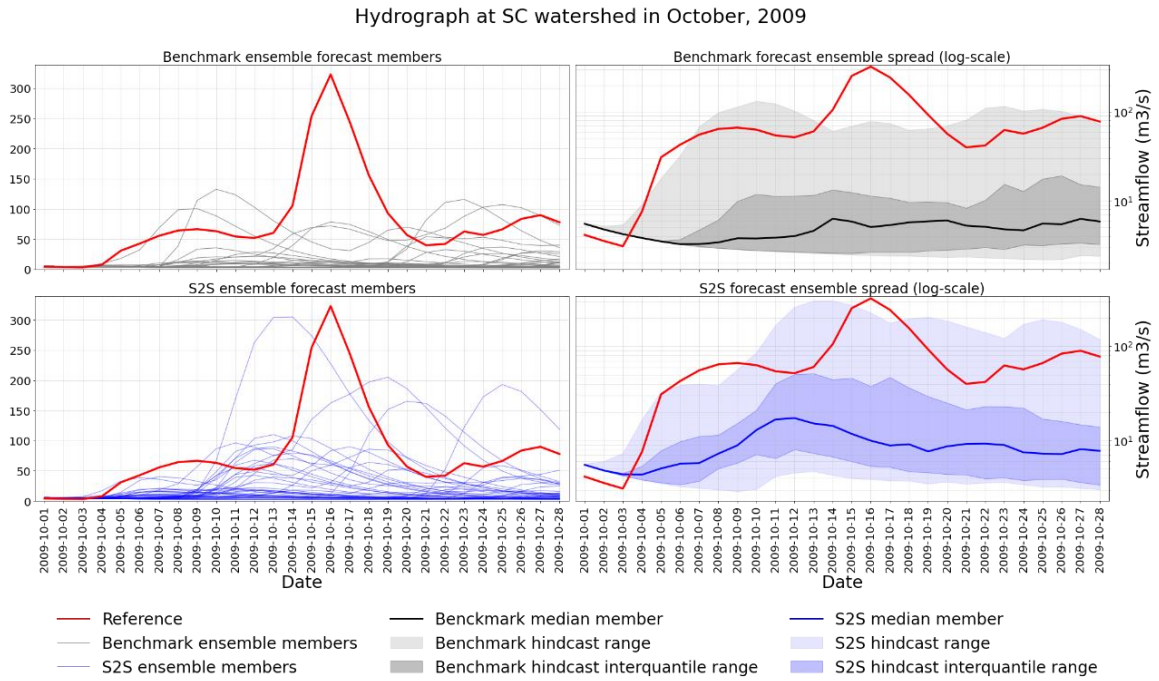

Figure 3.10 Predicted and measured daily hydrograph in October 2009 at the SC watershed.

Figure 3.11 presents the same elements as Figure 3.9, but for the hydrograph of the CR watershed in the April of 1999. The peak streamflow volume happened on the 4th day of April 1999. For this event, the streamflow hindcasts generated with both S2S precipitation and the resampled precipitation have successfully enveloped the peak streamflow volume. But the streamflow hindcast generated with the resampled precipitation has significantly overestimated

the peak volume by almost 3 times (predicted a max value of ~680 m^3/s and measured 180 m^3/s) in contrast to the streamflow hindcast generated with the S2S precipitation. The hindcast range plots from Figure 3.11 provide more details, where the reference streamflow is almost enveloped by the interquartile range of the streamflow hindcasts generated with the S2S precipitation. There is also a stronger correspondence between the measurements and the median values of the streamflow hindcasts generated with the S2S precipitation when compared to the benchmark.

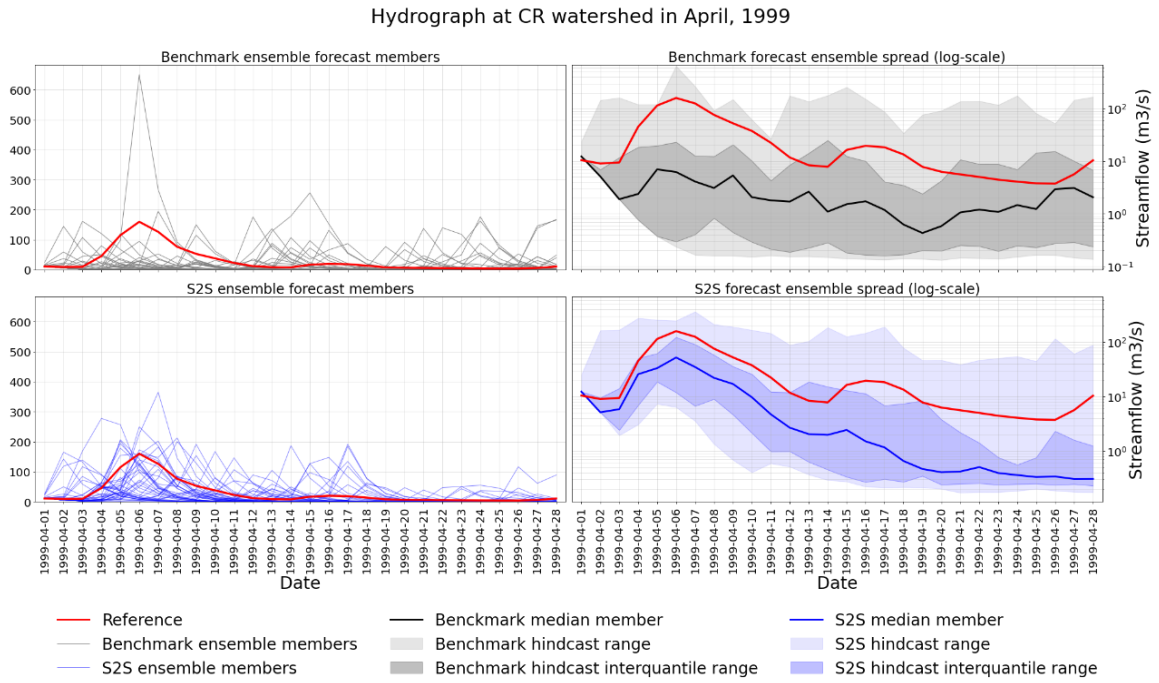

Figure 3.11 Predicted and measured daily hydrograph in April 1999 at the CR watershed.

Figure 3.12 presents the same elements as depicted in Figure 3.9, but for the hydrograph of BP watershed in the October of 1990. Within the selected month, the peak volume of the streamflow happened on the 25th day with a volume of around 800 m^3/s. Prior to the peak, there were another two smaller spikes with volumes around 280 m^3/s and 320 m^3/s happened earlier in that month. The ensemble of the streamflow hindcast generated with the S2S precipitation has successfully enveloped both spikes, whereas the benchmark streamflow simulation using

resampled precipitation has only enveloped the second spike. Hindcasts generated with both S2S precipitation and resampled precipitation missed the peak volume near the end of that month. However, there is one ensemble member generated with S2S precipitation that successfully made a close volumetric prediction with only one day offset prior to the happening of that peak. Given such a result, I reckon that the S2S precipitation is able to result an overall better streamflow prediction for the selected watershed and flooding cases.

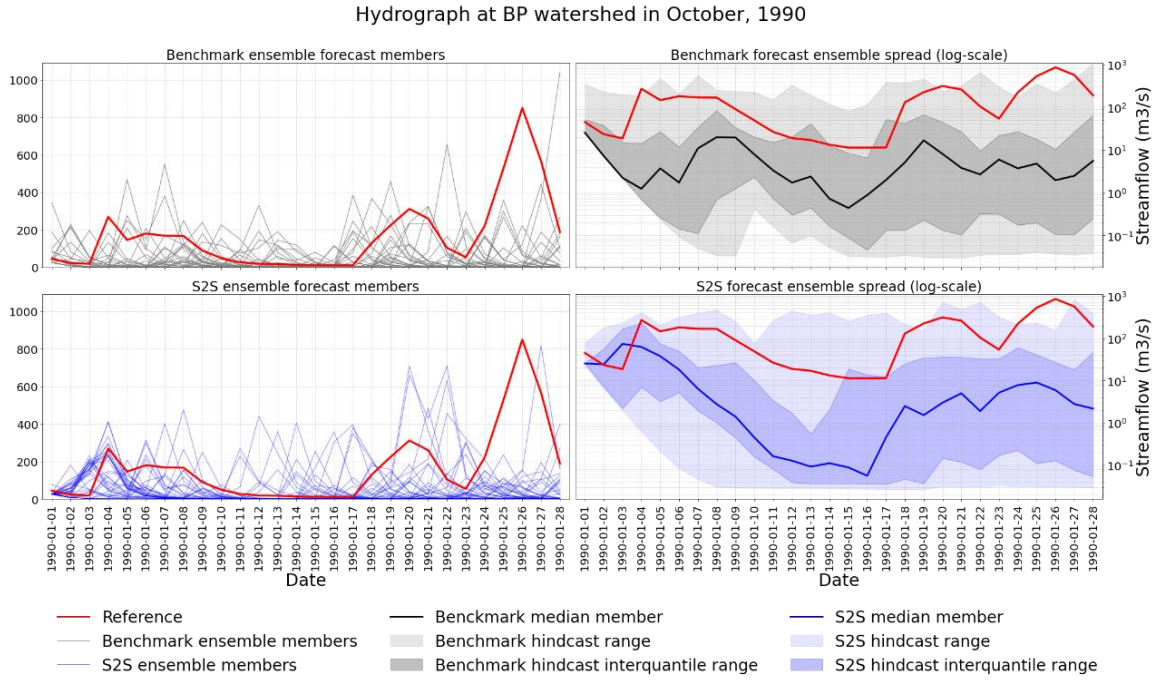

Figure 3.12 Predicted and measured daily hydrograph in January 1990 at the BP watershed.

3.5 Discussion

In this chapter, I evaluate the hydrologic performance of S2S precipitation from NMME-2 in streamflow forecasts at four selected watersheds in the Southern region of the U.S. The results show that the S2S precipitation can produce an overall better streamflow prediction capability at the Subseasonal timescales compared to the benchmarking performance of the resampled precipitation. According to my results, S2S precipitation leads to overall better statistics of KGE, POD, and FAR, which reflects the advantages of S2S precipitation in

providing not only more accurate predictions upon mid-volume streamflow but also better categorical predictions upon high percentile extreme events. The visual inspections of the selected events at the study watersheds (**Section 3.4.4**) also confirm such advantages of S2S precipitation over the resampled precipitation in a more direct manner.

The four selected watersheds are located in a climate region, in which the S2S precipitation is found to have the worst raw forecast performance and the most challenging area than some other climate regions (Zhang et al., 2021). To be specific, Zhang et al (2021) found that S2S precipitation forecasts only preserved a marginal level of skill once the forecast lead time exceeds 2 weeks and beyond. Yet, in this chapter, I find out that the S2S precipitation can still lead to an overall better streamflow hindcasts than the benchmark (i.e., streamflow hindcasts generated with the resampled precipitation). This highlights the potential broader application of S2S precipitation forecast for streamflow predictions at the sub-seasonal timescales in the future. Based on the experiment in this chapter, I further encourage other researchers to conduct similar experiments at watersheds in other climate regions of CONUS to further promote available S2S precipitation forecast products. With an overall more accurate and reliable streamflow forecast at S2S timescale, human activities, including reservoir operations, irrigation planning, proactive flood mitigations, etc., could be greatly benefited when using the S2S ensemble precipitation forecasts (Welles et al., 2007; White et al., 2017; Yang et al., 2021).

Nevertheless, the predictive skill of streamflow at longer lead times beyond 2 weeks (14 days) still remains at a relatively low statistical confidence level, which should be improved through a joint effort using new tools and emerging forecast data. Uncertainty in streamflow forecasts mainly comes from three sources including imperfect representation of hydrologic

systems, errors/imperfect estimations of IHCs at the streamflow forecast initialization time, and inaccurate S2S hydrometeorological forecasts.

In this chapter, I propose a direct implementation of S2S precipitation to develop longer streamflow prediction and benchmark the results with a rather standard ESP approach using resampled precipitation for streamflow forecasting. Another source of uncertainty I wanted to highlight is that the performance of hydrologic simulations due to model calibration is still far from perfect. As presented in **Section 4.1**, even with the perfect hydrometeorological measurements, errors in streamflow simulations are still commonly presented. A recent study by Gao et al. (2021) suggests through the application of distributed hydrologic models, the common existing non-perennial streamflow in the NCEI South climate region might be better considered, which might lead to overall better streamflow forecasts at S2S timescale. I believe that future research of such to further advance the current understandings in the fundamentals of hydrology is needed for overall better hydrologic forecasts at the Subseasonal timescales.

Research on achieving better estimations of the IHCs is another research direction that may greatly benefit the development of more accurate S2S streamflow predictions. In this chapter, I only estimate the IHC of watersheds through the standard "spin-up" processes. An alternative could be merging the large-scale hydrologic remote sensing datasets through data assimilation. Such datasets include the soil moisture estimations from the Soil Moisture Active Passive (SMAP) mission (Entekhabi et al., 2010), groundwater condition estimations from the twin Gravity Recovery and Climate Experiment (GRACE) mission (Frappart and Ramillien, 2018; Tapley et al., 2004), and/or surface water condition estimation through the recently launched Surface Water and Ocean Topography (SWOT) mission (Morrow et al., 2019), whereas these

datasets may provide intermediate and more accurate observations for estimating the IHC of the studied watersheds.

But more importantly, I reckon that the accuracy of available S2S hydrometeorological forecast products might be the current limiting factor for more skillful streamflow forecasts at S2S timescale. One limitation of this chapter is that I have been using PET climatology for all streamflow hindcast experiments to separately quantify the effectiveness of S2S precipitation in streamflow predictions. Therefore, I suspect that by adopting PET forecasts derived from other forecast variables generated by the same S2S forecast models, the quality of streamflow forecasts could be further improved. However, since the streamflow of the four selected watersheds is mainly dominated by precipitation (not snowmelt), I believe improving the quality of S2S precipitation forecasts should still be the focus for future works.

However, accurately predicting the occurrences as well as the volumes (in terms of extremes) of precipitation events at S2S timescale remains challenging. Previous evaluation studies upon different S2S precipitation forecast products generally concluded that there was only a marginal level of predictive skill when forecast lead times exceed 2 weeks (~14 days). To improve the forecast skills at longer lead times, advancements need to be made to identify additional predictability sources of precipitation at S2S timescale (de Andrade et al., 2021; Huang et al., 2021; Moron and Robertson, 2020), and/or to enhance current forecast coupled GCM models (Robertson et al., 2015; Xue et al., 2021). An alternative to improve S2S precipitation forecasts could be utilizing nowadays very popular Machine Learning and Deep Learning techniques to incorporate additional forecast variables through post-processing (Akbari Asanjan et al., 2018; Vitart et al., 2022; Zhang et al., 2023b).

In summary, based on the case studies in this chapter, I believe that future studies on streamflow predictions at S2S timescale may include (1) Extending similar S2S streamflow forecast/hindcast experiments at watersheds in other climate regions over CONUS; (2) Improving available hydrologic models as well as the estimations of IHCs for a better representation of hydrologic systems; (3) Further improving the quality of S2S precipitation forecasts by identifying additional predictability sources and/or enhancing available S2S forecast models; (4) applying machine learning or deep learning techniques for more accurate and reliable precipitation and/or streamflow forecasts at S2S timescale.

3.6 Conclusion

In this chapter, the S2S precipitation forecasts from 4 contributing forecast models to the NMME-2 data archive were adopted for streamflow hindcast experiments under the ESP framework at 4 study watersheds in the Southern region of the U.S. The performance of streamflow hindcasts generated with S2S precipitation were evaluated and compared against the benchmark streamflow hindcasts generated with resampled precipitation. The overall predictive skills of both S2S precipitation and resampled precipitation were quantified with KGE at different forecast lead times. The performance of streamflow hindcasts in terms of high-volume extreme events within different percentile segments were evaluated with POD and FAR. In addition, a few high-volume extreme events at the study watersheds were selected and plotted for visual inspections to verify the effectiveness of S2S precipitation in streamflow predictions in a more direct manner. My results highlight future applications of S2S precipitation in streamflow forecasts. The major findings and conclusions of this chapter are summarized below.

1. The predictive skill of S2S precipitation is generally higher than that of the resampled precipitation over the study watersheds, as indicated by the KGE values at different forecast lead times. However, such an advantage is only apparent at shorter forecast lead times within around 2 weeks (~14 days). Once the forecast lead times exceed 2 weeks, S2S precipitation only presents a marginal advantage over the resampled precipitation in terms of predictive skill.

2. S2S precipitation can lead to an overall better streamflow hindcasts at all forecast lead times as compared to the resampled precipitation inputs. However, in contrast to the first conclusion, the advantage of streamflow hindcasts generated with S2S precipitation is more apparent at longer lead times after ~5 days. Such advantage is consistent across different study watersheds.

3. Evaluation of high-volume streamflow values suggests that S2S precipitation leads to better categorical predictions upon extreme events. For daily extreme streamflow events within 4 different percentile segments (i.e., 75%-90%, 90%-95%, 95%-99%, and above 99%), at different forecast lead times within 4 weeks, and over 4 study watersheds (a total of $4 \times 4 \times 4 = 64$ scenarios), S2S precipitation presents overall better performance under 52 scenarios with not only lower FAR but also higher POD values.

4. Visual inspections of a few selected high-volume streamflow events at the study watersheds further confirm the advantages of S2S precipitation in streamflow forecast in a more direct manner. For the selected events, the ensemble streamflow hindcasts generated with S2S precipitation have better enveloped the peak flows than those generated with resampled precipitation. In general, the streamflow hindcasts generated with S2S precipitation present either closer volumetric predictions or smaller temporal offsets of the peak volume.

Acknowledgments

The material is based upon work supported by the National Science Foundation (NSF) under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3, and the Oklahoma EPSCoR Research Seed Grand supported by NSF. This work is also partially supported by U.S. Department of Defense, Army Corps of Engineers (DOD-COR) Engineering With Nature (EWN) Program (Award No. W912HZ-21-2-0038).

References

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J. and Peng, Q. 2018. Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. Journal of Geophysical Research: Atmospheres 123(22), 12,543-512,563.

Borovikov, A., Cullather, R., Kovach, R., Marshak, J., Vernieres, G., Vikhliaev, Y., Zhao, B. and Li, Z. 2019. GEOS-5 seasonal forecast system. Climate Dynamics 53(12), 7335-7361.

Cao, Q., Shukla, S., DeFlorio, M.J., Ralph, F.M. and Lettenmaier, D.P. 2021. Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in coastal Western US watersheds. Journal of Hydrometeorology 22(6), 1535-1552.

Day, G.N. 1985. Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management 111(2), 157-170.

de Andrade, F.M., Young, M.P., MacLeod, D., Hirons, L.C., Woolnough, S.J. and Black, E. 2021. Subseasonal precipitation prediction for Africa: Forecast evaluation and sources of predictability. Weather and Forecasting 36(1), 265-284.

Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J. and Johnson, J. 2010. The soil moisture active passive (SMAP) mission. Proceedings of the IEEE 98(5), 704-716.

Frappart, F. and Ramillien, G. 2018. Monitoring groundwater storage changes using the Gravity Recovery and Climate Experiment (GRACE) satellite mission: A review. Remote Sensing 10(6), 829.

Gao, S., Chen, M., Li, Z., Cook, S., Allen, D., Neeson, T., Yang, T., Yami, T. and Hong, Y. 2021. Mapping dynamic non-perennial stream networks using high-resolution distributed hydrologic simulation: A case study in the upper blue river basin. Journal of Hydrology 600, 126522.

Gent, P.R., Danabasoglu, G., Donner, L.J., Holland, M.M., Hunke, E.C., Jayne, S.R., Lawrence, D.M., Neale, R.B., Rasch, P.J. and Vertenstein, M. 2011. The community climate system model version 4. Journal of climate 24(19), 4973-4991.

Gupta, H.V., Kling, H., Yilmaz, K.K. and Martinez, G.F. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology 377(1-2), 80-91.

Huang, H., Patricola, C.M., Bercos-Hickey, E., Zhou, Y., Rhoades, A., Risser, M.D. and Collins, W.D. 2021. Sources of subseasonal-to-seasonal predictability of atmospheric rivers and precipitation in the western United States. Journal of Geophysical Research: Atmospheres 126(6), e2020JD034053.

Kirtman, B.P. and Min, D. 2009. Multimodel ensemble ENSO prediction with CCSM and CFS. Monthly Weather Review 137(9), 2908-2930.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P. and Becker, E. 2014. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bulletin of the American Meteorological Society 95(4), 585-601.

Knoben, W.J., Freer, J.E. and Woods, R.A. 2019. Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences 23(10), 4323-4331.

Maurer, E.P. and Hidalgo, H.G. 2008. Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. Hydrology and Earth System Sciences 12(2), 551-563.

Merryfield, W.J., Lee, W.-S., Boer, G.J., Kharin, V.V., Scinocca, J.F., Flato, G.M., Ajayamohan, R., Fyfe, J.C., Tang, Y. and Polavarapu, S. 2013. The Canadian seasonal to interannual prediction system. Part I: Models and initialization. Monthly weather review 141(8), 2910-2945.

Miao, C., Sun, Q., Duan, Q. and Wang, Y. 2016. Joint analysis of changes in temperature and precipitation on the Loess Plateau during the period 1961–2011. Climate Dynamics 47(9), 3221-3234.

Moron, V. and Robertson, A.W. 2020. Tropical rainfall subseasonal-to-seasonal predictability types. npj climate and atmospheric science 3(1), 1-8.

Morrow, R., Fu, L.-L., Ardhuin, F., Benkiran, M., Chapron, B., Cosme, E., d'Ovidio, F., Farrar, J.T., Gille, S.T. and Lapeyre, G. 2019. Global observations of fine-scale ocean surface topography with the Surface Water and Ocean Topography (SWOT) mission. Frontiers in Marine Science 6, 232.

Naeini, M.R., Yang, T., Sadegh, M., AghaKouchak, A., Hsu, K.-l., Sorooshian, S., Duan, Q. and Lei, X. 2018. Shuffled complex-self adaptive hybrid evolution (SC-SAHEL) optimization framework. Environmental Modelling & Software 104, 215-235.

Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., Viger, R., Blodgett, D., Brekke, L. and Arnold, J. 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences 19(1), 209-223.

Pegion, K., Kirtman, B.P., Becker, E., Collins, D.C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D. and Zhu, Y. 2019. The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. Bulletin of the American Meteorological Society 100(10), 2043-2060.

Robertson, A.W., Kumar, A., Peña, M. and Vitart, F. 2015. Improving and promoting subseasonal to seasonal prediction. Bulletin of the American Meteorological Society 96(3), ES49-ES53.

Schaake, J. and Larson, L. 1998 Ensemble streamflow prediction (ESP): Progress and research needs, pp. J19-J24.

Shukla, S., Sheffield, J., Wood, E.F. and Lettenmaier, D.P. 2013. On the sources of global land surface hydrologic predictability. Hydrology and earth system sciences 17(7), 2781-2796.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S. and Hsu, K.L. 2018. A review of global precipitation data sets: Data sources, estimation, and intercomparisons. Reviews of Geophysics 56(1), 79-107.

Tapley, B.D., Bettadpur, S., Watkins, M. and Reigber, C. 2004. The gravity recovery and climate experiment: Mission overview and early results. Geophysical research letters 31(9).

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E. and Fuentes, M. 2017. The subseasonal to seasonal (S2S) prediction project database. Bulletin of the American Meteorological Society 98(1), 163-173.

Vitart, F., Robertson, A.W., Spring, A., Pinault, F., Roškar, R., Cao, W., Bech, S., Bienkowski, A., Caltabiano, N. and De Coning, E. 2022. Outcomes of the WMO Prize Challenge to Improve Subseasonal to Seasonal Predictions Using Artificial Intelligence. Bulletin of the American Meteorological Society 103(12), E2878-E2886.

Welles, E., Sorooshian, S., Carter, G. and Olsen, B. 2007. Hydrologic verification: A call for action and collaboration. Bulletin of the American Meteorological Society 88(4), 503-512.

Werner, A.T. and Cannon, A.J. 2016. Hydrologic extremes–an intercomparison of multiple gridded statistical downscaling methods. Hydrology and Earth System Sciences 20(4), 1483-1508.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J. and Murray, V. 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorological applications 24(3), 315-325.

Wood, A.W. and Lettenmaier, D.P. 2008. An ensemble approach for attribution of hydrologic prediction uncertainty. Geophysical Research Letters 35(14).

Wood, A.W., Leung, L.R., Sridhar, V. and Lettenmaier, D. 2004. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Climatic change 62(1), 189-216.

Xue, Y., Yao, T., Boone, A.A., Diallo, I., Liu, Y., Zeng, X., Lau, W.K., Sugimoto, S., Tang, Q. and Pan, X. 2021. Impact of initialized land surface temperature and snowpack on subseasonal to seasonal prediction project, phase i (ls4p-i): organization and experimental design. Geoscientific Model Development 14(7), 4465-4494.

Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D. and Peng, Q. 2021. A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region. Journal of Hydrology 602, 126723.

Yossef, N.C., Winsemius, H., Weerts, A., van Beek, R. and Bierkens, M.F. 2013. Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. Water Resources Research 49(8), 4687-4699.

Zhang, L., Kim, T., Yang, T., Hong, Y. and Zhu, Q. 2021. Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous US. Journal of Hydrology 603, 127058.

Zhang, L., Yang, T., Gao, S., Hong, Y., Zhang, Q., Wen, X. and Cheng, C. 2023. Improving Subseasonal-to-Seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous US using machine learning models. Atmospheric Research 281, 106502.

Zhang, Q., Shen, Z., Xu, C.-Y., Sun, P., Hu, P. and He, C. 2019. A new statistical downscaling approach for global evaluation of the CMIP5 precipitation outputs: Model development and application. Science of the Total Environment 690, 1048-1067.

Chapter 4. Improving Subseasonal-to-seasonal Forecasts in Predicting the Occurrence of Extreme Precipitation Events over the Contiguous U.S. Using Machine Learning Models

The major content of this chapter is from a journal article published in *Atmospheric Research*.
© Elsevier. Used with permission.

Zhang, L., Yang, T., Gao, S., Hong, Y., Zhang, Q., Wen, X., & Cheng, C. (2023). Improving Subseasonal-to-Seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous US using machine learning models. *Atmospheric Research*, *281*, 106502.

Abstract

The precipitation forecasts at the Subseasonal-to-Seasonal (S2S) scale are valuable information to assist water resources planning and decision-making at an extended range. But the raw S2S precipitation forecasts are rather limited regarding their predictive skills, which hinders further hydrological applications. Many previous studies were carried out to validate the S2S precipitation forecasts from the perspective of the ensemble "mean", while existing study, which evaluates and improves the ability of S2S forecasts in predicting extreme precipitation events with their ensemble spreads, is rarely reported. This chapter aims to improve the S2S forecasts in predicting the occurrence of weekly extreme precipitation events above 99% over the contiguous United States (CONUS). The Random Forest Classifiers (RF) were employed and additional forecasts variables (i.e., surface air temperature, geopotential heights at 500 hPa and 850 hPa) were included to post-process the raw S2S forecasts from the NASA's Goddard Earth Observation System model version five (GEOS5). Different RF training inputs and RF hyperparameter sensitivity analysis are examined. I found that (1) using S2S precipitation forecast as the only inputs to RF, the forecast quality is improved significantly only at week 1; (2) When additional forecasts variables are included in RF training, the forecast skill tend to improve slightly at longer lead times after weeks 2; (3) The tunning of the maximum tree depth

of RFs combine with the inclusion of additional forecasts variables as inputs to RF can improve the forecasts skill at all lead times over CONUS. In short, this chapter demonstrates the effectiveness of the application of RF as well as the effectiveness of additional forecast variables in improving the S2S extreme precipitation forecasts, which could be potentially useful for flood and river ensemble forecasting. Multiple statistical metrics, including the ensemble probability of detections (EPOD), ensemble false alarm ratios (EFAR), ensemble critical success index (ECSI), Brier Sill Score (BSS), and Area Under the Receiver Operating Characteristics Curves (AUROC) are employed for a comprehensive evaluation of the predictive performance of S2S extreme precipitation forecasts under different experiment scenarios.

4.1 Introduction

Riverine flooding typically comes after heavy precipitation, causing impacts on human socio-economic activities, including loss of human lives (Begum et al. 2007, Singh and Kumar 2013), water contamination (Taylor et al. 2011, Yang et al. 2022), agricultural damage (Bremond et al. 2013), and disruption of transportation systems (Suarez et al. 2005). Reliable precipitation forecasts are critical for decision-makers to adaptively change their strategies to mitigate the impacts of such natural disasters (Sorooshian et al. 2011).

Advanced weather forecasting and climate models can generate precipitation forecasts at different timescales for water resources planning. At lead times within a medium range (i.e., 2 weeks into the future), precipitation forecasts generated from Numerical Weather Prediction (NWP) models are the most widely adopted products (Kuligowski and Barros 1998, Robertson et al. 2013), which simulate the propagation and evolution of atmospheres. At longer lead times, the General Circulation Models (GCMs) coupled with dynamic components of the

atmosphere, ocean, and land surface are more suitable tools as compared to the NWP models in producing seasonal, annual, or even decadal precipitation outlooks (Vitart 2017, Xiang et al. 2019). Both NWP and GCM-generated precipitation forecasts are widely studied and applied by various mission agencies and research communities (Clark et al. 2017, Wood and Lettenmaier 2006, Yang 2018).

However, a forecast gap lies in the Subseasonal-to-Seasonal (S2S) timescale, defined explicitly as the transitional period of 10 to 30 days between weather predictions and seasonal outlooks (White et al. 2017). Previous studies found that the predictability of weather is either lost already (i.e., initial conditions of the atmosphere) or has yet to start dominating local weather (i.e., ocean-weather interactions) at S2S timescales (Vitart et al. 2017). Such a lack of predictability sources makes it extremely challenging to provide accurate S2S forecasts, especially for rainfall events (White et al. 2017).

Many efforts have been made to advance precipitation forecasts by identifying additional predictability sources of the weather at S2S timescale. Series of initial conditions of land surface and atmosphere are found to be associated with the variation of sub-seasonal weather (Asoka and Mishra 2015, Chelton and Wentz 2005, Cohen et al. 2010, Guo et al. 2011, Stockdale et al. 2015, Thomas et al. 2016). More recently, multiple atmospheric variation modes over oceans are also found to be significantly dominating the distribution and magnitude of subseasonal precipitation events (Dai and Wigley 2000, Hsiao et al. 2020, Mariotti et al. 2020, Yang 2017a, Zhang and Ling 2017).

Additional efforts have also been made by mission agencies all over the globe to provide experiment S2S forecast datasets through coupled GCM models (Kirtman et al. 2014, Pegion et al. 2019, Vitart 2014, Vitart et al. 2017, Yang et al. 2018). Available S2S forecast datasets

include European Center for Medium-Range Forecasts (ECMWF), the S2S project by the World Weather Research Program, and the North America Multi-Model Ensemble Phase II (NMME-2), etc. Although existing S2S forecast products offer a promising opportunity for seamless hydrologic predictions, it is commonly agreed that S2S precipitation forecasts suffer from a substantial amount of forecasts biases and a marginal level of predictive skills (Baker et al. 2019, de Andrade et al. 2019, de Andrade et al. 2021, King et al. 2020, Tian et al. 2017, Vigaud et al. 2017, Wang and Robertson 2019, Tao et al. 2018, Yang et al. 2015).

More importantly, existing studies were mainly focused on investigating the accuracy of ensemble means or medians of S2S precipitation forecasts, while the ability of ensemble S2S forecasts to predict extreme events with their ensemble spreads is rarely reported. Nevertheless, from a practical perspective, extreme information is critically needed as precipitation forecasts generally cannot be applied deterministically at S2S ranges in hydrology (Day 1985). From the standpoint of water infrastructure operation (i.e., reservoirs and long-distance water transferring systems), decision-makers would desire the model-generated ensemble forecasts to envelop and include future extreme precipitation events as much as possible so that the potential risks of associated natural disasters could be considered beforehand and avoid infrastructure failure (VanBuskirk et al. 2021, Wu et al. 2020, Yuan et al. 2015, Yang et al. 2020).

As mentioned in previous **Chapter 3**, there is only a limited number of studies focused on S2S extreme precipitation forecasts over CONUS. Among some recent studies, Cao et al. (2021) examined the hydrologic performance of the ensemble means of S2S forecasts in forecasting flood events at three watersheds in the western U.S. In addition, Zhang et al. (2021) studied and examined the capability of S2S precipitation forecasts from NMME-2 in capturing the pattern of extreme rainfall with the entire spread of the ensembles. However, according to

Zhang et al. (2021), the skills of the S2S precipitation forecasts in predicting the occurrence of weekly extreme precipitation events are marginal after two weeks into the future.

The poor quality of S2S precipitation forecasts in predicting extreme events hinders a broader application of S2S forecasts in flood predictions and sustainable water infrastructure operations. Various statistical-based post-processing techniques are available, which can help correct the forecast bias and improve S2S precipitation forecast accuracy. When using these statistical-based post-processing tools, though the forecast biases can often be removed nicely, many studies found that the forecast skill tends to remain either unchanged or even deteriorate after removing the biases (Baker et al. 2020, Li et al. 2017, Manzanas et al. 2018, Zhao et al. 2017). This is because S2S precipitation forecast skill is affected by multiple factors, including coarse spatial resolutions, imperfect/unrealistic precipitation parameterization schemes, and the computational errors originated in resolving the partial differential equations of the atmosphere when generating rainfall forecast in the physical models (Davis and Goadrich 2006, Ebert and McBride 2000). Due to these reasons, existing statistical-based post-processing approaches are somewhat limited in improving the predictive skills of S2S precipitation forecasts. Therefore, more advanced tools are still needed, especially for extreme precipitation events at S2S timescale.

As an alternative to the existing statistical-based post-processing approaches, Machine Learning and Data Mining (ML&DM) techniques are promising tools for improving the S2S extreme precipitation forecasts skill scores. The ML&DM techniques have the flexibility to include an arbitrary set of input variables in the classification and regression process (Yang et al. 2017a and 2017b). They can effectively identify the complex relationships between selected input variables and target variables, which may not be directly related to each other. Many

researchers have successfully applied ML&DM for precipitation forecast adaptations. For example, Miao et al. (2019), Pan et al. (2021), Pan et al. (2019), Wang et al. (2021) applied ML&DM to post-process precipitation forecasts over different study regions and reported overall improvements in forecast skills. However, some other studies reported that ML&DM tends to underestimate and limit the reproduction of extreme values (Akbari Asanjan et al. 2018, Baño-Medina et al. 2020, Kim et al. 2022, Sadeghi et al. 2020). Nevertheless, given the sensitive dynamics of extreme precipitation events (Faridzad et al. 2018, Nie et al. 2020, Pendergrass 2020, Srinivas et al. 2018), I expect that the triggering of extreme events should be easier to identify when it is compared to regular precipitation events. Therefore, a popular ML&DM model, i.e., the Random Forest or RF classifier, is applied in this chapter to predict the occurrence of extreme precipitation events over different regions of CONUS. The RF is an ensemble-based tree algorithm. It can handle correlated conditional variables and is robust against overfitting with the presence of high-level noise in the training data (Breiman 2001, Strobl and Zeileis 2008). As one of the most popular ML algorithms, the RF has been widely applied in a variety of hydrometeorological studies, including statistical downscaling (He et al. 2016, Tao et al. 2018), reservoir release predictions (Yang et al. 2015, 2020, and 2021), and the post-processing of precipitation forecasts (Herman and Schumacher 2018a, b, Loken et al. 2019).

On top of using advanced ML techniques, adding additional forecast variables into the ML training may further increase the predictive skill of precipitation forecasts. Precipitation forecasts from dynamic models are generated through a process called “parameterization.” Instead of computing numerical values by resolving the partial differential equations in forecast models, this “parameterization” process generates quantitative precipitation forecasts

empirically based on the importance of other explicit atmospheric variables (i.e., variables computed directly by resolving physical equations, such as temperature, pressure, etc.). The reason for adopting such "parameterization" schemes in dynamic models is because precipitation is formed through complex micro-physical and chemical processes that surpass the resolution and capability of current models (Stensrud 2009). Such "parameterization" schemes are generally considered of high uncertainty and less reliable compared to the computations of explicit atmospheric variables (Bader and Roach 1977, Best et al. 2011, Betts et al. 1998, Bukovsky and Karoly 2007, Tian et al. 2017, Vitart 2004, Yang 2017 a, b). Thus, including additional forecast variables together with a robust ML algorithm may lead to a higher chance of identifying the dynamic, varying patterns of precipitation events over space and time.

Therefore, in this chapter, multiple S2S forecast variables are included in the training of RF to predict the occurrence of weekly extreme precipitation events over CONUS. The input forecast variables include each individual and a combination of (1) precipitation, (2) surface air temperature, (3) 500 hPa geopotential heights, and (4) 850 hPa geopotential heights. The surface air temperature, 500 hPa, and 850 hPa geopotential heights are related to the formation of precipitation events. These meteorological variables have been commonly used in previous studies to downscale and post-process precipitation forecasts (Li et al. 2022s, Miao et al. 2019, Pan et al. 2019). In this chapter, I chose 99% as the threshold to identify weekly extreme precipitation events without a universal standard.

Previously, Zhang et al. (2021) reported that the capability of NMME-2 S2S forecast in capturing the occurrence of extreme precipitation events is marginal after week 2. Following Zhang et al. (2021)'s work, this chapter develops a prototype model to improve the capability of S2S forecasts in predicting the occurrence of extreme precipitation events. The experiment S2S

forecast dataset used in this chapter is from one contributing forecast model of NMME-2, i.e., the NASA Goddard Earth Observing System version 5 model (GEOS5). Building off of many existing studies that focused on the ensemble means, I further consider the entire ensemble of S2S forecasts. A few existing statistical metrics were modified for the evaluation of ensemble forecasts. The contribution of this chapter also includes sensitivity tests on one RF hyperparameter, and different model input scenarios are also examined for comparison.

In summary of the goal of this chapter, I intend to address the following research questions: 1***) Can the predictions on the occurrence of extreme precipitation events be improved at the S2S ranges through ML techniques over CONUS? 2) Can additional atmospheric forecast variables improve S2S extreme precipitation forecasts over CONUS? 3) Does tunning of ML hyperparameter significantly affect the quality of extreme precipitation forecasts over CONUS? And 4) Do the ML-enabled forecasts perform consistently in different regions over CONUS?***

The rest of this **Chaper 4** is organized as follows: In **Section 4.2**, I first present data and study region. The experiment settings and evaluation metrics are described in **Section 4.3**. **Section 4.4** presents the results. The discussions and main conclusions are presented and summarized in **Section 4.5** and **Section 4.6**, respectively.

4.2 Data and Study Regions

The datasets used in this chapter include (1) the daily S2S forecast variables from GEOS5 and (2) a ground truth daily precipitation observation dataset for validation, the AN81d daily precipitation dataset from the Parameter-elevation Regressions on Independent Slopes Model (PRISM). Both forecast and reference datasets are collected from 01/01/1982 to 12/31/2011 to cover a 30-year study period.

The GEOS5 produces 10-member ensemble forecasts on the first day of each month during the study period. Each member of the GEOS5 dataset provides daily forecasts of multiple hydrometeorological variables with a lead time of up to 274 days. The 10-member ensemble forecasts are generated with perturbed initial conditions (Borovikov et al. 2019). Raw GEOS5 forecasts have spatial resolutions of 1° (~100 km). Four GEOS5 forecast variables are used in this chapter, including precipitation (P), surface air temperature (T), and geopotential height at 500 hPa (G500) and 850 hPa (G850).

The PRISM is a gridded dataset that covers the entire CONUS with a spatial resolution of ~0.04° (4 km). The PRISM combines a digital elevation model with both surface and Radar precipitation measurements (Daly and Bryant 2013). It is a reliable reference for validating satellite precipitation estimation and precipitation forecast products (Gowan et al. 2018, Mizukami and Smith 2012, Prat and Nelson 2015).

The collected S2S and PRISM datasets are pre-processed as follows. All collected S2S forecasts are truncated to 28 days (4 weeks) to focus on S2S timescale specifically. The truncated forecasts are labeled with individual lead times (1-28 days). Both S2S forecasts and PRISM were then resampled to 0.25-degree pixels to match each other. Specifically, the S2S forecasts were downscaled using the nearest neighbor approach. The PRISM was upscaled through aggregations (i.e., areal averages of all origin PRISM pixels fall in the corresponding 0.25-degree pixels) in order to match the coarse spatial resolution of GOES5 data.

To validate the experiment results, I conducted both pixel-based and regional evaluations of the raw and improved forecasts to better observe spatial patterns over CONUS. The regional assessment is performed based on the nine climate regions defined by the National Centers for

Environmental Information (NCEI) (Karl and Koss 1984) shown in previous Figure 4.1. A few NCEI regions include mountainous terrains. Specifically, the Rocky Mountains span the Northwest, West, West North Central, and Southeast regions; the Appalachian Mountainous regions are included in parts of the Northeast and Southeast regions.

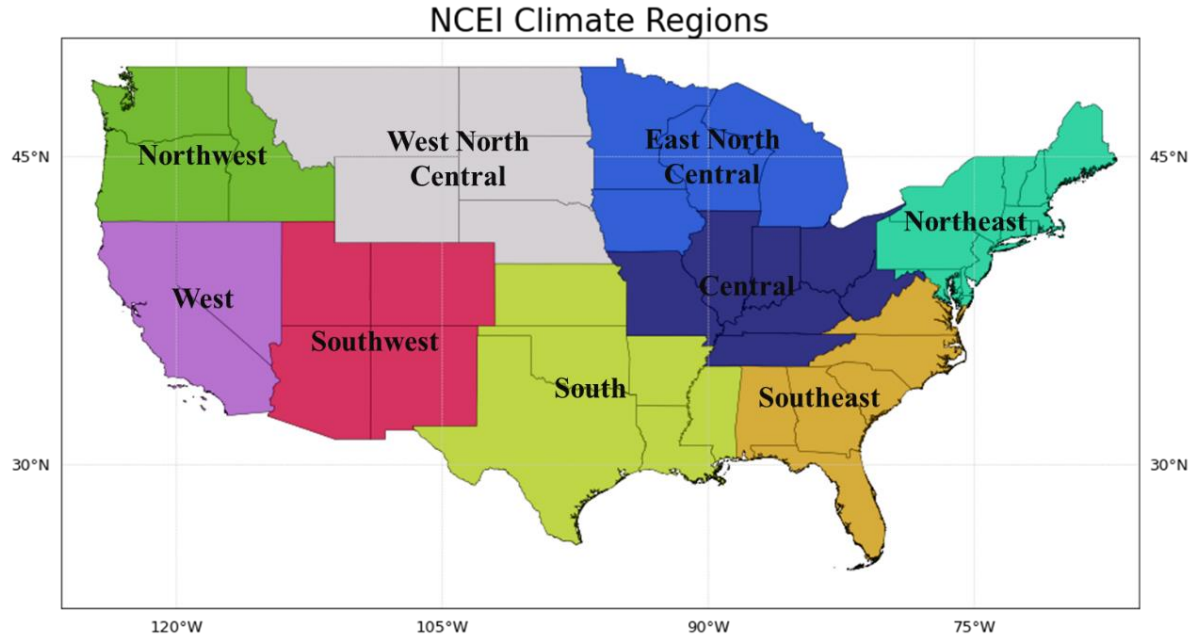

Figure 4.1 Climate region partitioning across CONUS per NCEI.

4.3 Experiment Setting and Evaluation Metrics

4.3.1 Experiment Setting

In this chapter, the RF classifiers are individually trained at each 0.25-degree pixel over CONUS to identify weekly extreme precipitation events beyond 99% using the 10-member ensemble S2S forecasts. The employment of RF is based on the open-source package in Python. The "leave-10-years-out" model across-validation strategy was adopted to avoid overfitting of the ML model over the entire study period. This strategy was commonly applied in the fields of hydrometeorology and climate change for forecast correction and statistical downscaling (Li et al. 2019, Manzanas et al. 2018). To perform the "leave-10-years-out" cross-validation, the whole study period was divided into three 10-year periods (i.e., 1982-1991,

1992-2001, and 2002-2011). When a particular 10 years are selected as the targeting period for forecast corrections, the remaining 20 years' data will be used to train the RF model.

The training of RF was carried out at different lead times as well. Taking week 1 forecasts as an example, the daily S2S forecasts with lead times smaller than 1 week (lead times 1 to 7 days) from the reference period are used as the inputs to train RF. The model training target is set to be the weekly extreme events observed in the PRISM dataset. The trained RF model is further used to produce categorial predictions upon extreme events at the same lead time for the target period. I repeated this training process at each 0.25-degree pixel over CONUS for week 2 (day 8 to 14), week 3 (day 15 to 21), and week 4 (day 22 to 28).

I expect that the S2S extreme precipitation forecasts can be improved via 1) the inclusion of additional atmospheric forecast variables as inputs to RF, and 2) the tunning of the max tree depth of RF. To test my hypothesis, a total of ten experiment scenarios were designed and shown in Table 4.1. In Table 4.1, "P" refers to precipitation forecasts, "T" refers to temperature forecasts, and "G" refers to geopotential height forecasts.

Table 4.1 Designed experiment scenarios.

Experiment Scenarios	E1 (Raw)	E2	E3	E4	E5	E6	E7	E8	E9	E10
Input variable(s)	N/A	P	P T	P G500	P G850	P T G500 G850	P T G500 G850	P T G500 G850	P T G500 G850	P T G500 G850
Max RF Tree depth	N/A	3	3	3	3	3	6	9	12	15

Among all scenarios presented in Table 4.1, I first benchmark the performance of the raw S2S precipitation forecasts in predicting the occurrence of extreme events (E1). The weekly extreme events over CONUS are identified by aggregating PRISM into weekly-averaged values and sorting out weeks with averaged values above the 99% threshold. Similarly, for the raw

precipitation forecasts for all ensemble members of GEOS5, I first aggregate them into weekly values. Then, the positive predictions from GEOS5 upon the occurrence of extreme precipitation events are identified by sorting out weeks with averaged forecast values above 99% according to their own statistics.

For experiment scenarios 2-5 (E2-E5) (Table 4.1), the hyperparameters of RF are controlled and remain unchanged, whereas the "max_depth" parameter is set to be 3, "max_features" is set to be 0.6, "n_estimators" is set to 150, and other hyperparameters remain as default for the RF model. The scenarios E2-E5 only differ in the variables used as inputs to RF, where the same P, P and T, P and G500, or P and G850 are used as the input variables under different scenarios E2-E5, respectively.

For experiment scenarios 6-10 (E6-E10), the input variables used to train RF are controlled and remain unchanged, whereas all collected S2S forecast variables are used as inputs to RF. However, under E6-E10, the "max_depth" of RF is set to be 3, 5, 7, 9, 12, and 15, respectively, while other hyperparameters remain unchanged (i.e., max_features=0.6, n_estimators=150)

4.3.1 Pixel-based and Event-based Forecast Evaluation Metrics

This chapter used three sets of evaluation metrics to compare the corrected forecast results. The first set of metrics is pixel-based evaluation metrics, including three popular categorical evaluation metrics of Probability of Detection (POD), False Alarm Ratio (FAR), and Critical Success Index (CSI). In the second set of metrics, an event-based probabilistic evaluation metric termed the Brier Skill Score (BSS), is employed. The last set of metrics is called the Area Under the Receiver Operating Characteristics Curves (AUROC), which is a graphical index used for spatial skill score evaluation for different climate regions over the CONUS.

To apply the POD, FAR, and CSI, I further modified these metrics to consider the entire ensemble of S2S forecasts and named the modified metrics as Ensemble Probability of Detection (EPOD), Ensemble False Alarm Ratio (EFAR), and Ensemble Critical Success Index (ECSI); The BSS and the AUROC are two widely applied metrics to quantify the model performance in the field of Hydrometeorology (Wilks 2011) and Machine Learning (Davis and Goadrich 2006), respectively. In this chapter, I choose BSS and AUROC to verify the performance of RF in addition to EPOD, EFAR, and ECSI. The following sections introduce the logic and detailed calculation of the employed metrics.

4.3.1.1 Mathematical Definitions of EPOD, EFAR, and ECSI

Table 4.2 describes a so-called contingency table, which illustrates the categorical relationship between the forecast and reference datasets. In this contingency table, the occurrence of extreme precipitation events is represented in the form of binary events with the possibility of two scenarios, either "$True$" or "$False$". The “H” in the contingency table means "hits,” and it refers to the number of successful predictions upon the occurrence of extreme events. Similarly, “M” means “misses,” which refers to the number of non-forecasted extreme events. “F” means “false alarms,” and it refers to the number of false positive forecasts upon the occurrence of extreme events. Based on the contingency table, three evaluation metrics of POD, FAR, and CSI are computed as follows: POD measures the fraction of accurate detection of events, and its calculation equation is H/(H+M); FAR measures the fraction of erroneous detection of events and its equation becomes F/(F+H), and CSI determines the fraction of correct event detection after ignoring the correct negative events, and the equation of CSI is H/(H+F+M).

Table 4.2 Contingency table of all possible outcomes for categorical forecasts of binary events.

Events	O	NO
P	H	F
NP	M	CR

However, precipitation forecasts generally cannot be applied deterministically on sub-seasonal timescales due to their uncertainties, and it is not informative to use the original PDO, FAR, and CSI metrics to evaluate S2S precipitation forecasts as they do not consider the entire ensemble spreads from different model outputs. Therefore, the use of EPOD, EFAR, and ECSI as the evaluation metrics will be more realistic and inclusive. The definitions of EPOD, EFAR, and ECSI generally follow the definitions of traditional POD, FAR, and CSI but consider all ensemble members of forecasts. By adopting modified EPOD, EFAR, and ECSI, I would like to quantify if extreme precipitation events are well-enveloped by the spreads of S2S forecasts.

To compute EPOD, EFAR, and ECSI, I define that successful categorical prediction made by any ensemble members upon extreme events will be counted as a "hit." Similarly, an unsuccessful categorical prediction made by all ensemble members will be counted as a "miss." Following the same logic, if none of the ten ensemble members has forecasted an extreme event above 99% while such event did not happen, it will be counted as a "correct negative". Finally, if all ensemble members have forecasted an extreme event above 99% while such an event did not happen, it will be counted as a "false alarm". With such modified definitions of "hit," "miss," and "false alarm," the EPOD, EFAR, and ECSI can be computed as follows:

$$EPOD = \frac{\sum_{i=1}^{n} I\left[\wedge_{j=1}^{m} (x_{ij}\ AND\ y_i)\right]}{\sum_{i=1}^{n} I(y_i)} \tag{4.1}$$

$$EFAR = \frac{\sum_{i=1}^{n} I\left\{\vee_{j=1}^{m}\left[x_{ij}\ AND\ (NOT\ y_i)\right]\right\}}{\sum_{i=1}^{n} I\left[\wedge_{j=1}^{m} (x_{ij})\right]} \tag{4.2}$$

$$ECSI = \frac{\sum_{i=1}^{n} I\left[\vee_{j=1}^{m}(x_{ij}\ AND\ y_i)\right]}{\sum_{i=1}^{n} I(y_i)+\sum_{i=1}^{n} I\left\{\vee_{j=1}^{m}\left[x_{ij}\ AND\ (NOT\ y_i)\right]\right\}} \tag{4.3}$$

In the equations above, n is the total number of categorical forecasts with the same lead time; and m quantifies the full ensemble members of the GEOS5 (i.e., 10); x_{ij} is the categorical prediction (i.e., either "$True$" or "$False$") made by the jth ensemble member of GEOS5 for a particular week. Similarly, y_i is the reference categorical indicator of whether an extreme event above 9the 9% threshold has happened or not at a particular week. The symbol $\wedge$ and $\vee$ represent large logical operations of OR and AND respectively. I is the indicator function where $I(True) = 1$ and $I(False) = 0$.

4.3.1.2 Event-based Brier Skill Score (BSS)

The Brier Skill Score (BSS) is a probabilistic evaluation metric that has been widely applied in the field of meteorology and atmospheric science. The BSS describes the quality of categorical probabilistic forecasts (Wilks 2011) and it quantifies the extent to which a forecast strategy improves predictions with respect to a reference forecast. The BSS is defined by the following equation (4.4):

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{4.4}$$

where BS and BS_{ref} are the Brier Scores (Brier 1950) for the forecasts and reference for a sample of n binary events, respectively. In this chapter, I use climatology as the reference forecast to quantify the relative performance of GEOS5 forecasts and the BS value is calculated from the of GEOS5 forecasts (i.e., E1 to E10 in my experiments). In other words, the BS and BS_{ref} can be computed with the following equation (4.5):

$$BS = \frac{1}{n}\sum_{t=1}^{n}(f_t - O_t)^2 \tag{4.5}$$

where f_t is the predicted probability of the event at time t, and O_t is equal to 1 or 0, depending on whether the extreme event subsequently occurred or not. The BSS metric has

been also reported to falsely inflate the skill of forecasts when adopted for evaluations upon extreme events (Wilks 2010). However, these extreme events are naturally rare in the entire data record. Therefore, BSS may inflate the actual skill of forecasts with many negative predictions upon the occurrences of rare extreme events. To derive a more realistic skill score of the forecasts, different weights are assigned to different categorical events during the calculation of BS in the following equation (4.6):

$$BS = \frac{1}{n}\left(\sum_{t=1}^{n} W_{O_t} * (f_t - O_t)^2\right) \quad (4.6)$$

In equation (4.6), $W_{O_t} = 0.99$ when O_t is equal to 1 and $W_{O_t} = 0.01$ when O_t is equal to 0, based on their probabilities in climatology. In this chapter, when computing the BS of GEOS5 forecast following equation (4.6), f_t is computed based on number of the positive categorical predictions among all 10 ensemble forecast members at time t. For example, when six out of ten GEOS5 members gave positive predictions upon extreme events, f_t would equal to 0.6. On the other hand, the f_t of reference forecast would constantly be equal to 0.01, given the climatology probability threshold of extreme events is 99%.

4.3.1.3 Regional-based Area Under a Receiver Operating Characteristics Curve (AUROC)

The last set of evaluation metrics, i.e., AUROC, is defined as the Area Under a Receiver Operating Characteristics (ROC) curve. Note that due to the huge amount of computation at all 0.25-degree pixels across the entire CONUS, the ROC curves at each pixel and with different probability thresholds of RF cannot be plotted and presented recursively. Instead, I pooled the categorical predictions obtained from all ensemble forecasts under different experiment scenarios at different NCEI climate regions to calculate the area based AUROC. In other words, the ROC curves and their corresponding AUROC were plotted and computed at nine NCEI climate regions with the default probability threshold of RF of 50%.

The POD and POFD are needed to plot ROC curves and to compute AUROC. According to the contingency Table 4.2, the calculation of POD follows H/(H+M) and calculation of POFD follows F/(F+CR). For example, assuming a classifier has made series of predictions upon a certain "yes/no" type of event and resulted in POD of 0.8 and POFD of 0.3, then, such a classifier's corresponding ROC curve can be plotted as Figure 4.2. In Figure 4.2, the red line is the ROC curve. The blue dotted diagonal line indicates the skill of a classifier constantly making random guesses of yes or no. The red-shaded area below the ROC curve is defined as the AUROC. The ideal value of AUROC would be equal to one (i.e., POD=1 and POFD=0). In this chapter, I only present the NCEI regional AUROC for conciseness. The individual ROC curves at each NCEI climate region are presented in **Appendix B**.

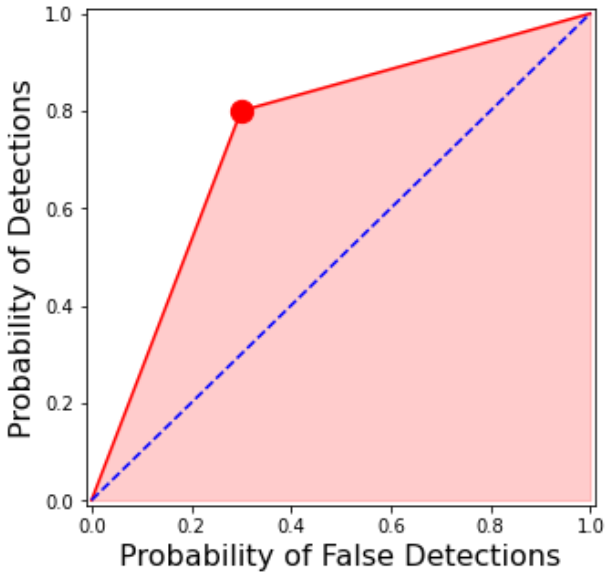

Figure 4.2 An Illustrating figure of ROC and AUROC.

4.4 Results

4.4.1 Model's Sensitivity on Input Variables (Comparison of Scenarios E1-E6)

In this **Section 4.4.1**, I first present model's sensitivity analysis on different combinations of inputs based on the results obtained from scenarios E1-E6. Then, in **Section 4.4.2**, I further analyze how model's correction performance changes over different ML hyperparameters based on the results from scenarios E1, and E6-E10. Here, E1 benchmarks the performance of raw precipitation forecasts in predicting the occurrences of weekly extreme events above 99%. E2-E6 are the RF-generated categorial predictions with different input variables, and E6-E10 differs on the "max-depth" parameter of the RF model with the same input combination of using all meteorological predictors. The evaluations and results are summarized in the order of previously defined EPOD, EFAR, ECSI, BSS, and AUROC metrics for both sensitivity analysis cases.

Figure 4.3 presents the EPOD under E1-E6 with different forecast lead times. The ideal value of EPOD is 1, indicating all extreme events have been enveloped by the spreads of ensemble forecasts. In Figure 4.3, darker red colors indicate higher EPOD. The average EPOD over the entire CONUS is computed and presented in red in each subplot for comparison.

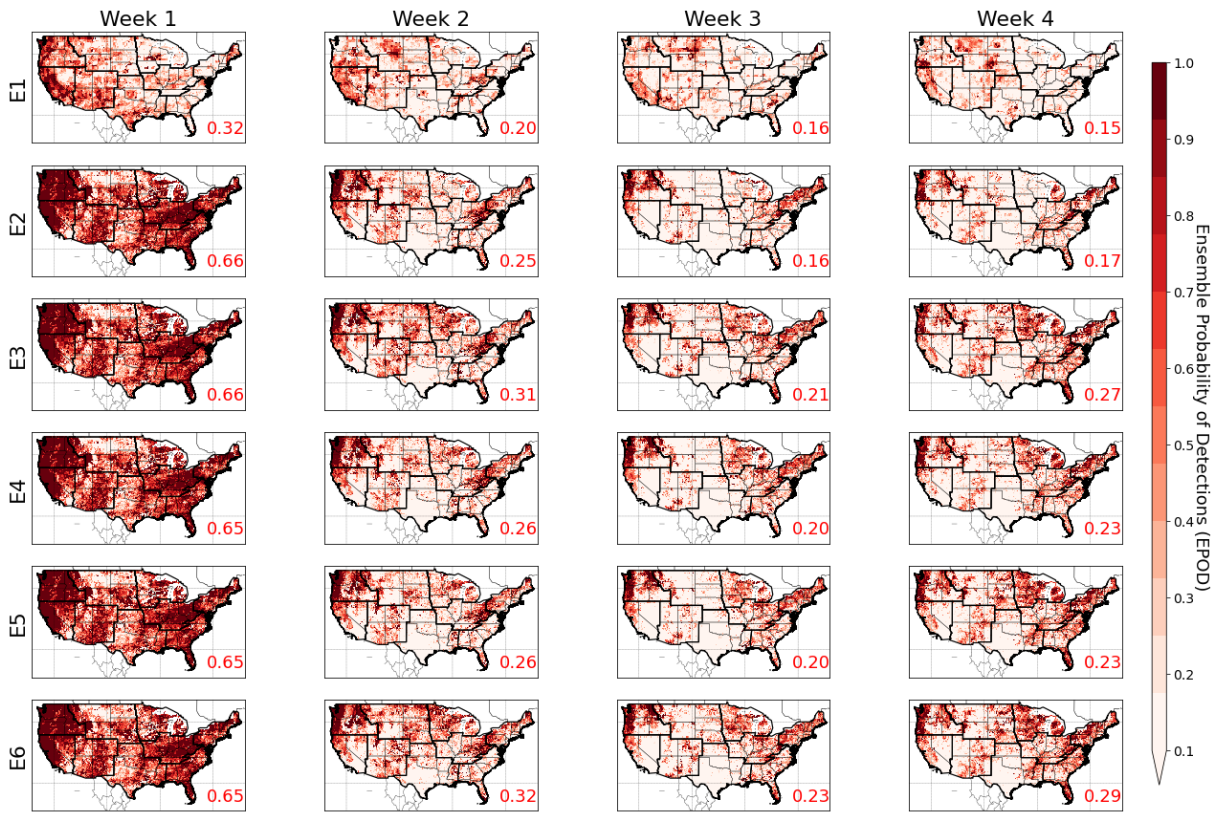

Figure 4.3 The pattern of EPOD from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4).

Compared to the raw forecast performance (E1), the RF-generated forecasts with P as the only input (E2) show significantly higher EPOD at the shortest lead time of week 1 and slightly higher EPOD after week 1 (Figure 4.3). The inclusion of additional forecast variables of T, G500, or G850 (E3-E5) can further increase the EPOD slightly at week 2, week 3, and week 4, as it is compared to the case where only P is used to train RF (E2). When all forecast variables of P, T, G500, and G850 are included together as input to train RF model (E6), additional improvements of EPOD at weeks 2-4 can be observed when comparing to the rest of the cases where only using one or two forecast variables as inputs to train RF (E2- E5). However, without additional tunning of the hyperparameter of the RF model, the inclusion of additional forecast variables (E2- E5) has slightly decreased the EPOD after week 1 compared to the raw GEOS5 forecasts (E1).

Spatially, the raw precipitation forecasts (E1) deliver higher EPOD values at the West Coastal regions of CONUS. But this advantage decreases rapidly over lead times. Different spatial patterns are also observed between experiment scenarios using P, P and T, P and G500, or P and G850 as the only input to the RF model (E2- E5). While using P as the only input to the RF model (E2), higher EPOD can be observed in West Coastal regions and Appalachian Mountainous regions. While including T as an additional input to the RF model (E3), higher EPOD can be observed at most of the NCEI climate regions except for the South at week 2, week 3, and week 4. As a contrast, when including G500 or G850 as an additional input to the RF model (E4 and E5), noticeable higher EPOD values are only observed in the NCEI climate regions of East North Central, Central, as well as Southeast at week 2, week 3, and week 4. Finally, when using all available S2S forecast variables as inputs to RF classifiers (E6), the obtained EPOD values are increased at most of the locations over CONUS at week 2, week 3, and week 4, compared to the cases of using only one or two variables (E2-E5).

Figure 4.4 presents the EFAR under E1-E6 at different forecast lead times. The ideal value of EFAR is 0, indicating the ensemble spreads of forecasts are not under-dispersed and lead to ensemble false alarms. In Figure 4, darker blue colors indicate higher EFAR. The average EFAR over the entire CONUS is computed and presented in red in each subplot for comparison.

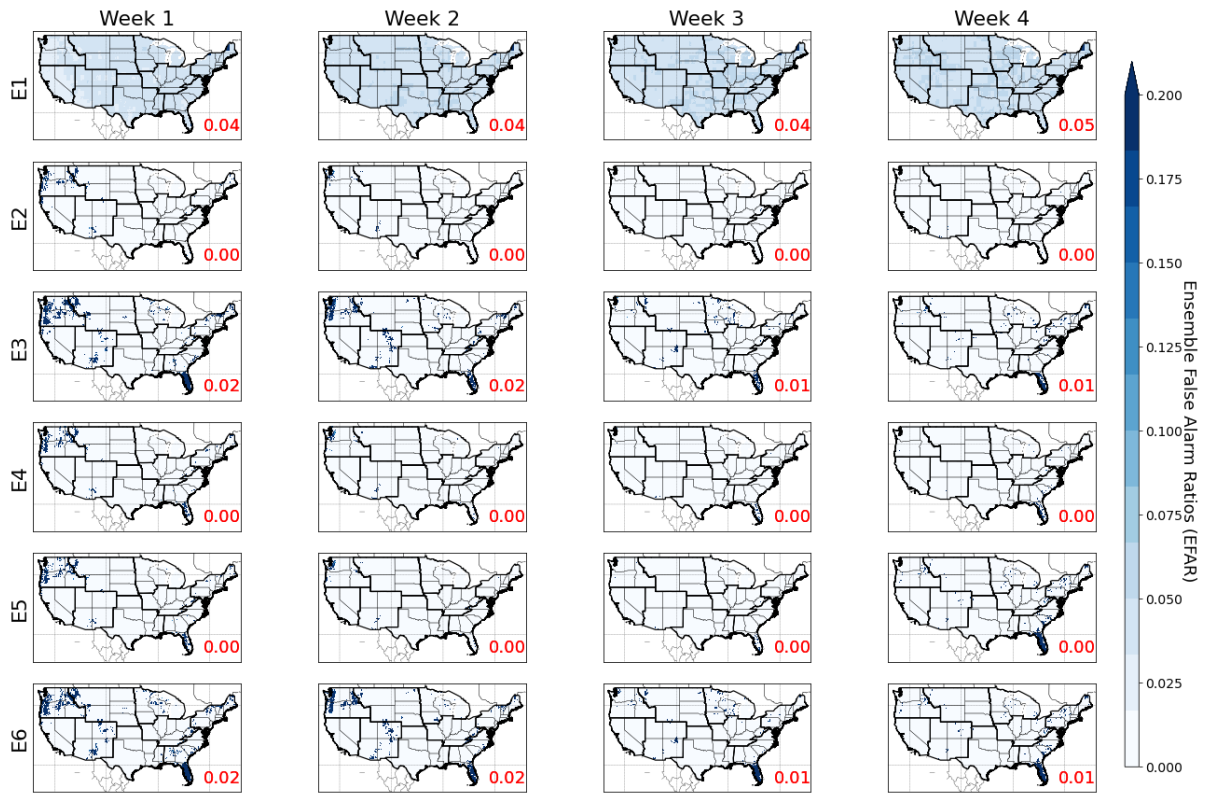

Figure 4.4 The pattern of EFAR from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4).

According to Figure 4.4, the overall small EFAR values (< 0.1) of raw forecast seem evenly presented over CONUS without any noticeable high-value spots. Further, the utilization of the RF model to correct extreme rainfall (E2-E5) has further reduced the overall EFAR values over CONUS. However, the application of the RF model seems to increase EFAR at some exceptional locations over CONUS. For example, when including T as an additional input to the RF model (E3), a few bluish pixels can be observed in the Rocky Mountain regions and the Florida Peninsula. This pattern has become more apparent when utilizing all available forecast variables as inputs to the RF model (E6), as more blue-colored pixels appeared in the Northwestern regions of CONUS, Rocky Mountainous regions, and in the Florida Peninsula.

Figure 4.5 presents the ECSI under E1-E6 at different forecast lead times. The ECSI considers the number of ensemble detections and ensemble false alarms at the same time. The

ideal value of ECSI is 1, indicating all forecasts have been enveloped by the spreads of ensemble forecasts while no false alarms were issued due to the under-dispersion of ensemble forecasts. In Figure 4.5, darker green colors indicate higher ECSI. The average ECSI over the entire CONUS is computed and presented in red in each subplot for comparison.

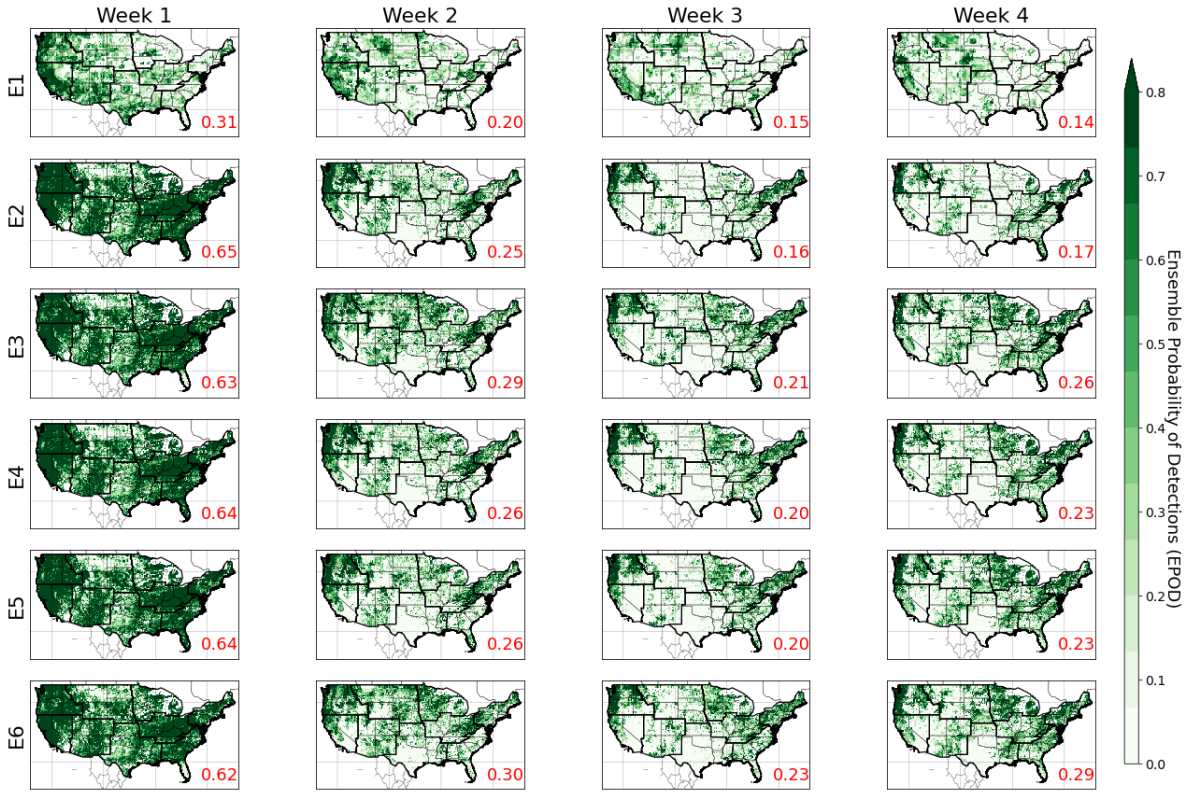

Figure 4.5 The pattern of ECSI from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4).

Comparing different experiment scenarios E1-E6 across the entire CONUS, it is obvious to notice that when using P as the only input to RF (E2), significantly higher skill over the raw forecasts (E1) can be observed at week 1 lead time (Figure 4.5). However, such advantages become marginal at longer lead times of weeks 3 and 4. When including additional variables of T, G500, or G850 as input to the RF model (E3-E5), I found that higher ECSI values can be observed at week 2, week 3, and week 4 lead times. However, the inclusion of additional forecast variables (E3-E5) leads to a marginal level of decrease of ECSI at the shortest lead

time of week 1, compared to the case where only P is used as input the RF model (E2). When all forecast variables are included as inputs to the RF model (E6), further skill improvements at longer lead times of weeks 2-4 are observed over CONUS. But the inclusion of all variables still leads to a marginal level of ECSI decrease at lead time of week 1 as it is compared to cases of using less input variables (E2-E5).

Figure 4.6 presents the BSS under E1-E6 at different forecast lead times. The BSS measures the agreement between forecast probability and measured events. The ideal value of BSS is 1. In Figure 4.6, darker purple colors indicate higher BSS. The average BSS values over the entire CONUS is computed and presented in red in each subplot for comparison.

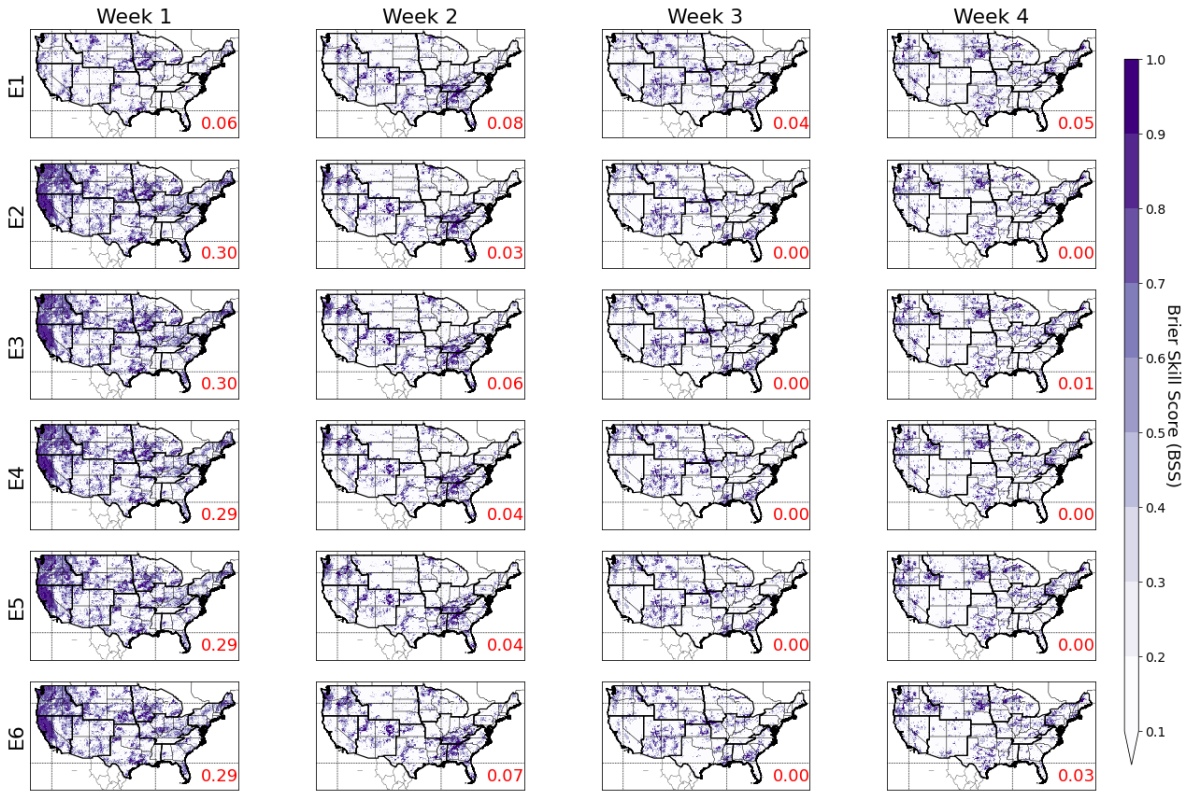

Figure 4.6 The pattern of BSS from experiment scenarios E1-E6, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4).

The application of the RF model with P as the only input (E2) has significantly improved BSS over CONUS compared to the raw GEOS5 forecast (E1) at week 1 lead time (Figure 4.6).

However, the BSS decreased over CONUS at weeks 2 to 4 with the applications of the RF model. The inclusion of T as an additional input variable to the RF model (E3) could slightly increase the BSS over CONUS at lead times of weeks 2 to 4, as it is compared to the case of using P as the only input to the RF model (E2). The inclusion of G500 or G850 as an additional input variable to the RF model (E4 and E5) could only improve the BSS over CONUS marginally at week 2 lead time, as it is compared to the case of using P as the only input to the RF model (E2). When all available forecast variables are used as inputs to the RF model, only a marginal level of additional improvement of BSS can be observed at lead times of week 2 and week 4, as it is compared to the case of using P as the only input to the RF model (E2).

The spatial patterns of BSS over CONUS appear to be similar among all experiment scenarios from E1 to E6. At lead time of week 1, higher BSS values are observed along the western coastal regions over CONUS. At lead time of week 2, higher BSS values are observed along the Rocky mountainous regions and the Appalachian mountainous regions over CONUS. At week 3 and week 4, higher BSS values are only observed at some scattered spots in the NCEI climate regions of Southwest, West North Central, Southeast, and Northeast, given an overall marginal level improvement of BSS values over CONUS.

Figure 4.7 presents the regional AUROC under E1 to E6 at different lead times. The ideal value of AUROC is 1, indicating all ensemble forecast members have performed perfectly with all extreme events detected while making no false detections at all. In Figure 4.7, darker red colors represent higher AUROC. The fractional values in each colored box are the computed AUROC values. Note that the “left-to-right” sequence of 9 NCEI regions in Figure 4.7 corresponds to the real-world geographical layout of these climate regions in the U.S., i.e., the

columns from left to right of Figure 9 correspond to the western coast to the eastern coast of CONUS.

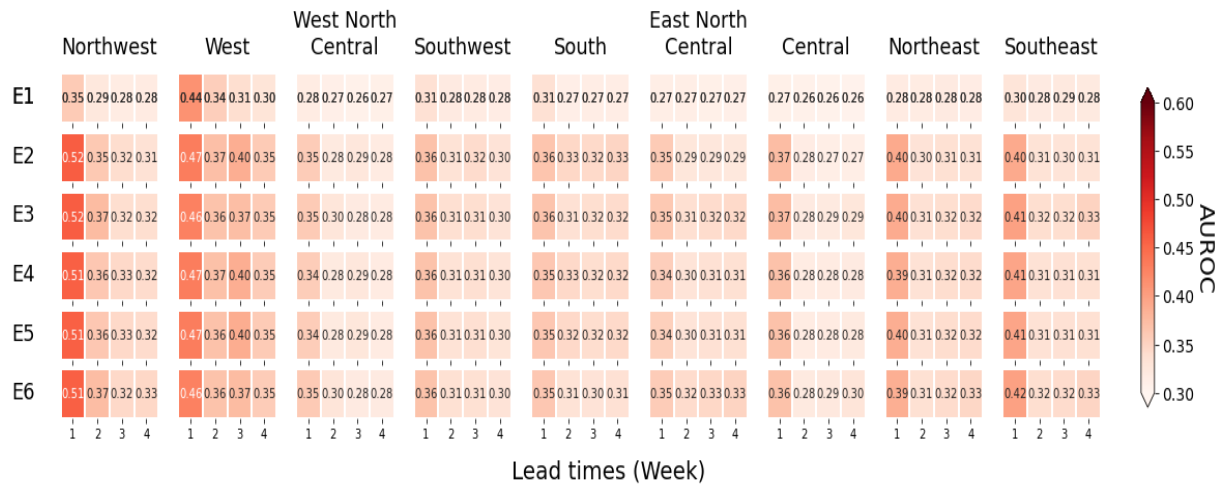

Figure 4.7 Spatially averaged AUROC values at 9 NCEI climate regions from experiment scenarios E1-E6, over CONUS and at different lead times (week 1 to 4).

A major difference in AUROC results is the observed statistics between the raw forecast (E1) and the remaining experiment scenarios with the applications of the RF model (E2-E6) (Figure 4.7). For example, using P as the only input to the RF model can bring noticeable improvement of AUROC at week 1 lead time at all NCEI climate regions over CONUS. However, such improvement becomes marginal at longer lead times after week 2, especially in the central regions of CONUS.

Compared to only using P as input to the RF model (E2), the inclusion of additional forecast variables brings marginal improvement of AUROC at longer lead times after week 2. The inclusion of T as input to the RF model (E3) has slightly increased the AUROC at longer lead times after week 2 at NCEI climate regions of Northwest, East North Central, Central, Northeast, and Southeast. The inclusion of G500 or G850 as an additional variable to the RF model has increased the AUROC after week 2 lead time at NCEI climate regions of Northwest, East North Central, Central, and Northeast. However, with the default hyperparameter settings

of RF, the inclusion of all available forecast variables (E6) does not show overall superior AUROC values over the entire CONUS. Compare to using P as the only input to the RF model (E2), the inclusion of all available forecasts variables (E6) is able to generate higher AUROC values at lead times after week 2 over NCEI climate regions of Northwest, West North Central, East North Central, Northeast, and Southeast.

4.4.2 Model's Sensitivity on Hyperparameter (Comparison of Scenarios E1 and E6-E10)

In this **Section 4.4.2**, I present the results from E1, and E6-E10, where E1 evaluates the raw precipitation forecast in predicting the occurrence of weekly extreme events above 99% and scenarios E6-E10 evaluate RF-generated categorial prediction upon weekly extreme precipitation events above 99%. In addition to EPOD, EFAR, ECSI, and BSS, I also present the regional AUROC over CONUS. The individual ROC curves at NCEI climate region are provided in **Appendix B**.

Figure 4.8 presents the EPOD values under E1, and E6-E10 at different forecast lead times. The ideal value of EPOD is 1, indicating all extreme events have been enveloped by the spreads of ensemble forecasts. In Figure 4.8, darker red colors indicate higher EPOD. The average EPOD over the entire CONUS are computed and presented in red in each subplot for comparison.

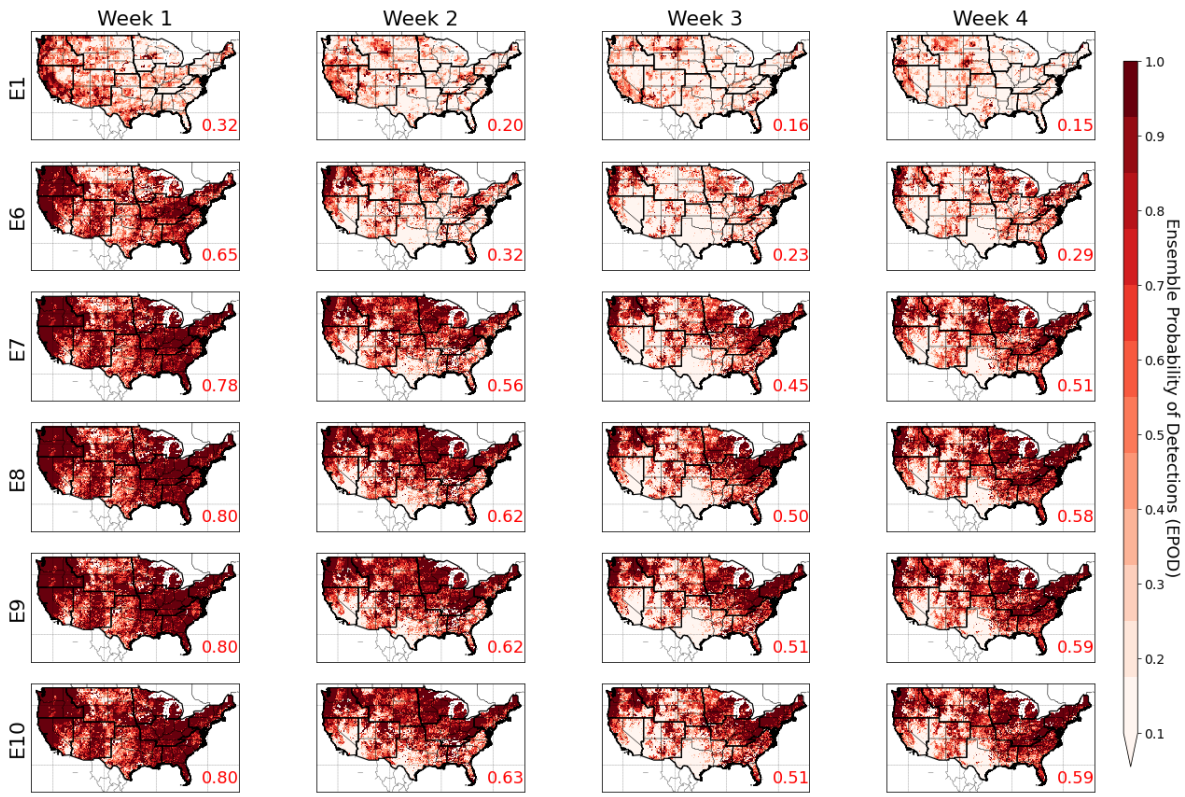

Figure 4.8 The pattern of EPOD from experiment scenarios E1, and E6-E10, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4).

Drastic differences are observed between the performance of raw forecast (E1) and RF-generated forecasts from other experiment scenarios (E6- E10) (Figure 4.8). This result agrees with previous EPOD values presented in Figure 4.3, indicating that the application of RF to incorporate additional forecast variables can significantly improve the EPOD values over CONUS.

Spatial differences are observed between experiment scenarios E6-E10, in which the maximum tree depth of RF varies from 3-15, respectively. The EPOD values steadily increase over CONUS with a larger max tree depth used in the RF model (Figure 4.8). But after the tree depth is over 9 (E8), the improvement of EPOD becomes marginal. Spatially, although a larger max tree depth leads to higher EPOD over CONUS in general, the EPOD values in some

regions remain unchanged. For example, at lead time of weeks 2-4, the EPOD over Southern California and Texas is not improved at all with a larger number of max tree depth.

Figure 4.9 presents the EFAR under E1, and E6-E10 at different forecast lead times. The ideal value of EFAR is 0, indicating that the ensemble spreads of forecasts are not under-dispersed and lead to ensemble false alarms. In Figure 4.9, darker blue colors indicate higher EFAR. The average EFAR over the entire CONUS is computed and presented in red in each subplot for comparison.

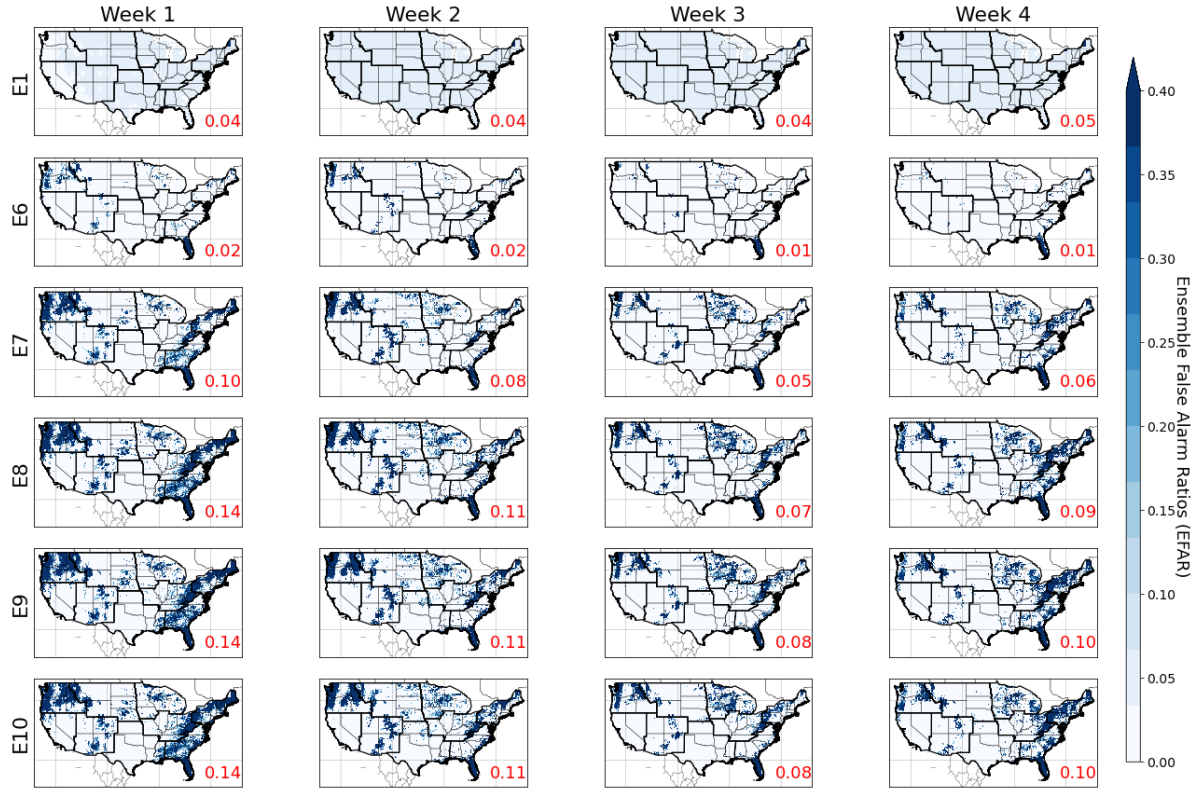

Figure 4.9 The pattern of EFAR from experiment scenarios E1, and E6-10, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4).

The EFAR of the raw forecasts are spatially uniform over CONUS, and they are without any noticeable high-value spots at all lead times, when compared with other experiment scenarios (E6- E10) (Figure 4.9). The utilization of RF with a max tree depth of 3 (E6) reduced the EFAR values at most of the regions over CONUS. However, higher EFAR values are

observed in Northwestern regions and Florida Peninsula at all lead times from E6, when it is compared to the raw forecasts. I also notice that if the tree depth is further increased (E7- E10), the EFAR will become larger at the Northwestern regions of CONUS, the Florida Peninsula, as well as the Appalachian Mountainous regions. Similarly, according to Figure 4.9, the increase of tree depth seems to only increase the EFAR at certain regions (i.e., Northwestern regions, Appalachian Mountainous regions, Florida Peninsula, as well as some scattered locations lies in the Central regions of CONUS), while the EFAR over other regions seems to be not sensitive to the increase of tree depth.

Figure 4.10 presents the ECSI under E1, and E6-E10 at different forecast lead times. The ideal value of ECSI is 1, indicating all forecasts have been enveloped by the spreads of ensemble forecasts while ensemble forecasts are not under-dispersed and lead to false alarms. In Figure 4.10, darker green colors indicate higher ECSI. The average ECSI over the entire CONUS is computed and presented in red in each subplot for comparison.



Figure 4.10 The pattern of ECSI from experiment scenarios E1, and E6-E10, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4).

Results from Figure 4.10 show an overall similar pattern to that of Figure 4.7. The major difference resides between the ECSI plots of raw forecast (E1) and remaining experiments E6-E10. According to the CONUS-averaged ECSI, the application of RF with multiple S2S forecast variables (E6-E10) has significantly improved the overall skill in predicting the occurrence of extreme precipitation events at all lead times. The results obtained under scenarios E6-E10 also show some degree of differences, which are mainly due to the increase of max tree depth used in the RF model. It is obvious that as the max tree depth increases, the overall ECSI over CONUS steadily increases. However, when tree depth exceeds 9 (E8- E10), the improvement of ECSI becomes marginal over CONUS. Among all experiment scenarios, the improvements of ECSI values appear to be the most significant when the tree depth is set to 9 (E8). Compared to the baseline E1 of raw precipitation forecasts, the relative ECSI

improvements under Scenarios E6-10 are 120% (0.30 to 0.66), 155% (0.20 to 0.51), 187% (0.15 to 0.43), and 250% (0.14 to 0.49) at lead times of weeks 1-4, respectively. Spatially, I observe that a larger number of the max tree depth will likely result in larger ECSI values over most of the regions across CONUS. But no improvements in ECSI can be observed at some specific locations, e.g., Southern California, parts of Nevada, and Texas.

Figure 4.11 presents the BSS under E1 and E6-E10 at different forecast lead times. The BSS measures the agreement between forecast probability and recorded extreme events. The ideal value of BSS is 1. In Figure 4.5, darker purple colors indicate higher BSS. The average BSS values over the entire CONUS are computed and presented in red in each subplot for comparison.

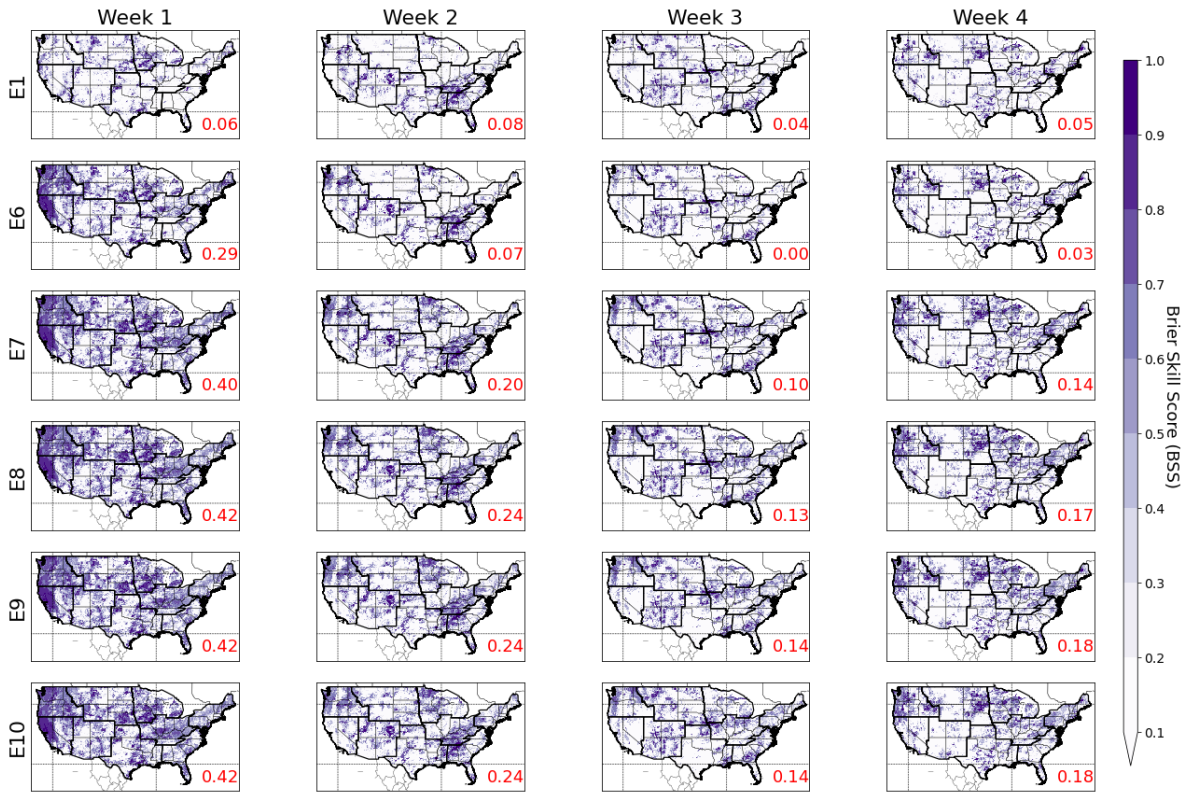

Figure 4.11. The pattern of BSS from experiment scenarios E1, and E6-E10, in predicting extreme precipitation events over the CONUS at different lead times (week 1 to 4).

Compared to the raw GEOS5 forecasts (E1), the utilization of the RF model with a max tree depth of 3 (E6) significantly increases the BSS at the lead time of week 1, but slightly decreases the BSS at lead times of weeks 2 to 4 over the entire CONUS (Figure 4.11). However, as the maximum tree depth increases (E7-E10), the post-processed forecasts present an overall higher BSS at all lead times compared to the raw GEOS5 forecasts (E1) over CONUS. Among all experiment scenarios, the improvement of BSS is the most significant when the tree depth is set to 12 or 15 (E9 or E10).

For all RF-involved experiment scenarios (E6-10), the spatial patterns of BSS are similar to that of Figure 4.5 (i.e., BSS from E1-E6). At the lead time of week 1, the BSS values at parts of the NCEI regions of West, Southwest, South, and West North Central are consistently lower than at other regions. At week 2 lead time, higher BSS values are only observed at locations in the Rocky mountainous regions and Appalachian mountainous regions over CONUS. At lead times of week 3 and week 4, a few scattered spots show higher BSS values in the NCEI climate regions of Southwest, West North Central, Southeast, and Northeast. In general, the tunning of the maximum tree depth of RF does not improve BSS at certain locations and lead times over CONUS.

Figure 4.12 presents the regional AUROC under E1, and E6-E10 at different lead times. The ideal value of AUROC is 1, indicating all ensemble forecast members have performed perfectly with all extreme events detected while making no false detections at all. In Figure 4.12, darker red colors represent higher AUROC. The fractional values that appeared in each colored box are the computed AUROC values. Note that the “left-to-right” sequence of 9 NCEI regions in Figure 4.12 corresponds to the real-world geographical layout of these climate

regions in the U.S., i.e., the columns from left to right of Figure 9 correspond to the western coast to the eastern coast of CONUS.

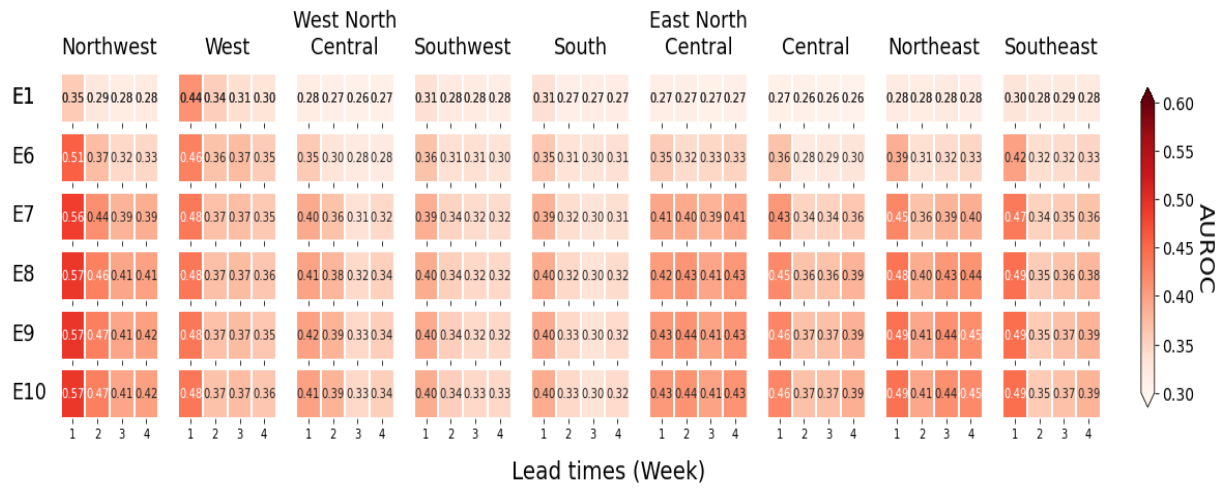

Figure 4.12 Spatially averaged AUROC values at 9 NCEI climate regions from experiment scenarios E1, and E6-E10, over CONUS and at different lead times (week 1 to 4).

Similar to previous results, I also compare the AUROC values RF-generated forecasts from E6-E10 with the raw forecast performance (E1). It is apparent that the application of the RF model along with additional atmospheric variables (E6-E10) resulted in improved values of AUROC in all regions over CONUS and at all lead times (Figure 4.12), which were similar to previously presented results of EPOD, EFAR, and ECSI.

Figure 4.12 also shows that the larger the tree depths of the RF, the higher AUROC over CONUS in general. However, some regions show less significant improvements of the AUROC as the max tree depth increases. These NCEI regions include West North Central, Southwest, and South (Figure 4.2). Within these regions, the improvement of AUROC seems to be neglectable after the tree depth exceeds 9 (E8 to E10). Among all NCEI climate regions, the AUROC values are significantly higher in the western and eastern coastal regions of U.S. than that over other regions in the middle of the continent. The results of AUROC values indicate that the employed RF model performed the best in the NCEI region of Northwest (Figure 4.2).

4.5 Discussion

The results from **Section 4.4.1** (E1-E6) show that without the tunning of the hyperparameter of the RF model, the post-processed S2S precipitation forecasts present mixed performance that are subject to different evaluation statistics. With the inclusion of additional forecast variables of T, G500, and/or G850, the evaluation metrics of EPOD, EFAR, and ECSI were improved in general, which were also consistent through all lead times over CONUS. However, the BSS and AUROC only showed a marginal level of improvement after week 2 and sometimes slightly deteriorated at week 1 lead time. I reckon such mixed performance of the post-processed forecasts show that the inclusion of additional meteorological variables is still informative, especially at longer lead times after week 2. However, the tunning of the hyperparameter of the RF model is critical when additional forecast variables are included as inputs to the RF model.

The BSS and the regional AUROC results from **Section 4.4.2** (E6-E10) further confirm the effectiveness of the inclusion of additional variables as well as my proposed RF model in post-processing S2S extreme precipitation forecast over CONUS. With a proper tunning scheme for the hyperparameter of the RF model, all evaluation statistics have shown noticeable improvements over CONUS. However, I also notice that once the max maximum tree depth of the RF model exceeds 12, the overall improvements of S2S forecasts become neglectable, and the EFAR values at some locations even slightly deteriorated. Then, I further suspect this indicates an overfitting of the RF model. Although the RF model does not overfit when the number of trees of the RF model is large enough (Breiman 2001), I did not test it in this particular study for conciseness. On the other hand, when the depth of trees of RF model becomes too deep but trained with relatively too few input variables, it is also possible that the RF model can no longer generalize over unseen points in the test dataset (Tang et al. 2018).

Nevertheless, combining all evaluation statistics from both **Section 4.4.1** and **Section 4.4.2**, I reckon that the S2S extreme precipitation forecasts can be improved through the employment of RF when (1) additional atmospheric forecast variables other than precipitation are included as inputs to the RF model; and with (2) the hyperparameter of RF classifiers is manually tunned to allow the model better to capture the spatial and temporal patterns of the extreme precipitation events.

The combined use of the ML model with the inclusion of additional forecast variables to improve S2S precipitation forecasts has a potential value in assisting flood predictions and river forecasts, especially at an extended range. Before the emergence of available S2S forecasts, the classical approach to predicting streamflow at S2S timescale is to create an ensemble of multiple precipitation time series by randomly resampling historical rainfall measurements. Although S2S forecasts provide an alternative to the classical approach, the accuracy and reliability of the raw S2S forecasts are rather limited, as indicated by existing research and present studies in prior chapters, especially in predicting extreme precipitation events. Many practitioners argue that the traditional way of resampling hydrometeorological measurements is less computationally expensive and could easily gain a higher probability of enveloping extreme events by creating a large ensemble size through resampling. But this chapter demonstrated that the employment of the RF model can improve S2S extreme precipitation forecasts without increasing the size of ensemble forecasts.

In this chapter, I demonstrate that additional atmospheric information, i.e., T, G500 and G850, could greatly benefit extreme precipitation forecasts at S2S timescale. Given the fact that heavy precipitation events could be triggered through different mechanisms of convection, orographic lifts, and/or large-scale synoptic systems, it is reasonable to expect additional

information could further improve the extreme precipitation forecasts at the S2S scale. I also suspect that additional slope and DEM information might be helpful in identifying orographic precipitations. Additional atmospheric forecast variables (e.g., wind direction and speed, specific humidity, sea level pressure, etc.) might be able to improve the prediction of the stratiform precipitation events brought by synoptic systems. The convective precipitation events at the Subseasonal timescales are sometimes considered as unpredictable given currently limited understanding in the geographical distribution of their predictability (Moron and Robertson 2020). Therefore, I can only hope that through the inclusion of multiple additional variables, AI&DM models may be able to identify of some patterns of convective precipitation given different local information (e.g., land cover types) as well as the projected atmospheric conditions at the Subseasonal timescales.

In this chapter, I also conduct a sensitivity test on the maximum tree depth of the RF model (**Section 4.4.2**). The sensitivity test result shows that the occurrence of extreme precipitation events can be better predicted by slightly tunning one RF model hyperparameter. According to my results, a larger tree depth of the RF model can increase the overall predictive performance on the occurrence of weekly extreme events. However, such improvement becomes marginal after the tree depth reaches 9. Further increase of the tree depth does not bring significant improvement of the forecast statistical measures. For decision tree models, there are many other hyperparameters left un-tested in this chapter, such as the number of leaves, maximum data samples in one leaf, splitting criteria, etc. In addition, I did not compare the predictive performance of RF with other popular ML models, such as Supportive Vector Machine, K-mean clusters, etc. There are some existing studies comparing popular ML models for interested readers in the field of hydrometeorology and water resources management (Hess and Boers

2022, Moon et al. 2019, Nayak and Ghosh 2013, Yang et al. 2016, Yang et al. 2021, Zhang et al. 2018). It will be a future effort to identify which ML is more capable of improving the S2S extreme precipitation forecasts together with a larger-scale model sensitivity study.

In this chapter, different spatial patterns of the categorial predictions are observed over CONUS. According to the regional AUROC results, as well as the obtained BSS values from **Section 4.4.2**, I can confidently conclude that, in general, the extreme precipitation events that happened in the NCEI climate regions of Southwest, West North Central, and South tend to be harder to predict even with a larger tree depth of the RF model. This spatial pattern is probably because extreme precipitation events at different regions over CONUS are dominated by different meteorological mechanisms. Extreme precipitation events in the coastal regions of CONUS are largely due to or associated with synoptic-scale events, such as extratropical cyclones and atmospheric rivers (Chen et al. 2018, Kunkel et al. 2012, Mahoney et al. 2016). These types of extreme events have enormous spatial (horizontal length > 100 km) and temporal (multiple days) extent and thus might be easier for GCMs to consider with their coarse spatial resolutions. In contrast, mesoscale convective systems, such as thunderstorms that occur more frequently in the NCEI regions of West North Central, Southwest, South, and Central regions (Kunkel et al. 2012), are generally smaller in spatial (smaller than 100 km horizontally) and temporal (often sub-daily) scales. Thus, these convective systems are very difficult for GCMs to simulate and thus challenging for the AI & DM models to capture their spatial and temporal variabilities.

One limitation of this chapter is that only precipitation events above 99% threshold were examined. However, heavy precipitation events at other quantiles (e.g., above 95%, 90%, 75%, etc.) are also capable of causing floods, which should be further studied separately. Due to the

limited length of this chapter, it will be a future effort to examine the available S2S precipitation forecasts in predicting heavy precipitation events at other quantiles.

Another follow-on study could be devoted to restoring the positive categorical predictions upon extreme events back to numerical values, i.e., rain rates. After all, categorical predictions cannot be directly used for hydrologic simulations to provide information on the magnitude of future streamflow. One simple way to resolve this issue is utilizing popular distribution-based approaches, such as quantile mapping (Cannon et al. 2015, Maraun 2013), to conduct bias corrections upon S2S precipitation forecasts. Specifically, the RF-categorized extreme and non-extreme events will be corrected/restored to numerical values according to historical records and thus ready for next-step hydrologic forecasts. However, I expect that treating extreme values might be extremely challenging since a few recent studies identified that the extreme precipitation events over the globe exhibited significant changes in frequency and magnitudes under global climate change (Fan et al. 2021, Kunkel et al. 2003, Madsen et al. 2014, Sun et al. 2017). Thus, when bias-correcting the extreme precipitation forecasts from S2S models, considerations of the "non-stationary" of climate and its impacts are also needed. (AghaKouchak et al. 2011, Cheng et al. 2014, Tao et al. 2018).

Finally, applying S2S precipitation forecast products in hydrologic forecasts at watershed scales is still an ongoing effort due to the raw forecasts' coarse spatial resolutions. I encourage practitioners to further evaluate and apply S2S precipitation forecasts for riverine flood predictions at a daily time step (Cao et al. 2021, Li et al. 2017, McInerney et al. 2020, Quedi and Fan 2020, Kim et al. 2021). This will help overcome the pitfall of the current study that only evaluates the S2S forecasts at a limited number of statistical metrics on a weekly basis. I encourage future research to be devoted to answering the questions of how post-processed S2S

ensemble precipitation forecasts perform in predicting floods induced by extreme precipitation events.

In summary, future research could be devoted to (1) investigating how additional variables further help ML models to better post-process S2S forecasts and identify the intrinsic physical dynamics related to extreme precipitation over CONUS; (2) quantifying the influences of different ML hyperparameters on the predictive performance upon extreme precipitation events at S2S timescale; (3) capturing the structural and random errors associated with different S2S forecast models at different spatial and temporal domains; and (4) linking the S2S forecasts to ensemble hydrologic forecasting and further investigating the usefulness of S2S forecasts in assisting ensemble streamflow forecasting of riverine floods.

4.6 Conclusion

In the current research, I conduct comprehensive evaluations regarding the performance of GEOS5 S2S forecasts with a specific focus on weekly extreme precipitation events over CONUS. I develop a proof-of-concept forecast adaptation framework using the Random Forest (RF) classifier to post-process the raw S2S forecasts at each 0.25-degree pixel over CONUS. Four S2S forecast variables are used, including precipitation, surface air temperature, and geopotential height at 500 hPa and 850 hPa. I also examine the performance of RF with respect to different maximum tree depths over CONUS. A total of ten different experiment scenarios are created to identify the ideal input variable combination and maximum tree depth that leads to the best predictions upon extreme precipitation events.

To evaluate the skill and demonstrate the potential hydrologic effectiveness of S2S ensemble forecast, I employ modified categorical metrics of ensemble probability of detections (EPOD), ensemble false alarm ratios (EFAR), and ensemble critical success index (ECSI). The

probabilistic evaluation metric of Brier Skill Score (BSS), as well as the regional Area Under the Receiver Operating Characteristics Curves (AUROC) are computed to confirm the improvements of the S2S precipitation forecast. The improved S2S forecasts will be useful for future hydrologic studies in different spatial and temporal domains. My major research conclusions in this chapter are listed as follows:

1. The application of RF can significantly improve S2S forecasts in terms of predicting the occurrence of weekly extreme precipitation events. The improvement is most significant at the lead time of week 1 and deteriorates rapidly after week 2 lead time. I found out that by including additional S2S forecast variables as the RF inputs, the forecast performance is further improved at longer lead times (i.e., weeks 2-4). Compared to the raw forecasts, RF could improve ECSI up to 116% (0.31 to 0.67), 165% (0.20 to 0.53), 200% (0.15 to 0.45), and 257% (0.14 to 0.50), at weeks 1-4, respectively.
2. I also found that the extreme precipitation events in West North Central, Southwest, and South regions, per NCEI definition, are harder to predict than that in other regions. I speculate such differences are due to different dominating precipitation mechanisms, which may result in different spatial and temporal scales of extreme precipitation events.
3. Sensitivity analysis indicates that increasing the maximum tree depth of RF would result in an overall better forecast at all lead times over CONUS. However, the improvements become neglectable once the tree depth exceeds nine in my study. I suspect this indicates an overfitting of the RF model. The obtained BSS and AUROC show consistent conclusions as indicated by EPOD, EFAR, and ECSI, which confirms the improvement of the S2S forecast is not subject to the definitions of modified evaluation metrics.

Acknowledgments

The material is based upon work supported by the National Science Foundation (NSF) under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3, and the Oklahoma EPSCoR

Research Seed Grand supported by NSF. This work is also partially supported by U.S. Department of Defense, Army Corps of Engineers (DOD-COR) Engineering With Nature (EWN) Program (Award No. W912HZ-21-2-0038); This research was also supported by a grant from the Research Council of the University of Oklahoma Norman Campus. The computing for this project was performed at the OU Supercomputing Center for Education & Research (OSCER) at the University of Oklahoma (OU).

References

AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K. and Amitai, E. (2011) Evaluation of satellite-retrieved extreme precipitation rates across the central United States. Journal of Geophysical Research: Atmospheres 116(D2).

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J. and Peng, Q. (2018) Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. Journal of Geophysical Research: Atmospheres 123(22), 12,543-512,563.

Alam, M.S. and Vuong, S.T. (2013) Random forest classification for detecting android malware, pp. 663-669, IEEE.

Asoka, A. and Mishra, V. (2015) Prediction of vegetation anomalies to improve food security and water management in India. Geophysical Research Letters 42(13), 5290-5298.

Bader, M. and Roach, W. (1977) Orographic rainfall in warm sectors of depressions. Quarterly Journal of the Royal Meteorological Society 103(436), 269-280.

Baker, S.A., Wood, A.W. and Rajagopalan, B. (2019) Developing subseasonal to seasonal climate forecast products for hydrology and water management. JAWRA Journal of the American Water Resources Association 55(4), 1024-1037.

Baker, S.A., Wood, A.W. and Rajagopalan, B. (2020) Application of postprocessing to watershed-scale subseasonal climate forecasts over the contiguous United States. Journal of Hydrometeorology 21(5), 971-987.

Baño-Medina, J., Manzanas, R. and Gutiérrez, J.M. (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. Geoscientific Model Development 13(4), 2109-2124.

Begum, S., Stive, M.J. and Hall, J.W. (2007) Flood risk management in Europe: Innovation in policy and practice, Springer Science & Business Media.

Best, M., Pryor, M., Clark, D., Rooney, G., Essery, R., Ménard, C., Edwards, J., Hendry, M., Porson, A. and Gedney, N. (2011) The Joint UK Land Environment Simulator (JULES), model description–Part 1: energy and water fluxes. Geoscientific Model Development 4(3), 677-699.

Betts, A.K., Viterbo, P. and Wood, E. (1998) Surface energy and water balance for the Arkansas–Red River basin from the ECMWF reanalysis. Journal of Climate 11(11), 2881-2897.

Borovikov, A., Cullather, R., Kovach, R., Marshak, J., Vernieres, G., Vikhliaev, Y., Zhao, B. and Li, Z. (2019) GEOS-5 seasonal forecast system. Climate Dynamics 53(12), 7335-7361.

Breiman, L. (2001) Random forests. Machine learning 45(1), 5-32.

Bremond, P., Grelot, F. and Agenais, A.-L. (2013) Economic evaluation of flood damage to agriculture–review and analysis of existing methods. Natural Hazards and Earth System Sciences 13(10), 2493-2512.

Bukovsky, M.S. and Karoly, D.J. (2007) A brief evaluation of precipitation from the North American Regional Reanalysis. Journal of Hydrometeorology 8(4), 837-846.

Cannon, A.J., Sobie, S.R. and Murdock, T.Q. (2015) Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? Journal of Climate 28(17), 6938-6959.

Cao, Q., Shukla, S., DeFlorio, M.J., Ralph, F.M. and Lettenmaier, D.P. (2021) Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in coastal Western US watersheds. Journal of Hydrometeorology 22(6), 1535-1552.

Chelton, D.B. and Wentz, F.J. (2005) Global microwave satellite observations of sea surface temperature for numerical weather prediction and climate research. Bulletin of the American Meteorological Society 86(8), 1097-1116.

Chen, X., Leung, L.R., Gao, Y., Liu, Y., Wigmosta, M. and Richmond, M. (2018) Predictability of extreme precipitation in western US watersheds based on atmospheric river occurrence, intensity, and duration. Geophysical Research Letters 45(21), 11,693-611,701.

Cheng, L., AghaKouchak, A., Gilleland, E. and Katz, R.W. (2014) Non-stationary extreme value analysis in a changing climate. Climatic change 127(2), 353-369.

Clark, R.T., Bett, P.E., Thornton, H.E. and Scaife, A.A. (2017) Skilful seasonal predictions for the European energy industry. Environmental Research Letters 12(2), 024002.

Cohen, J., Foster, J., Barlow, M., Saito, K. and Jones, J. (2010) Winter 2009–2010: A case study of an extreme Arctic Oscillation event. Geophysical Research Letters 37(17).

Dai, A. and Wigley, T. (2000) Global patterns of ENSO-induced precipitation. Geophysical Research Letters 27(9), 1283-1286.

Daly, C. and Bryant, K. (2013) The PRISM climate and weather system—an introduction. Corvallis, OR: PRISM climate group 4.

Davis, J. and Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves, pp. 233-240.

Day, G.N. (1985) Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management 111(2), 157-170.

de Andrade, F.M., Coelho, C.A. and Cavalcanti, I.F. (2019) Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. Climate Dynamics 52(9), 5451-5475.

de Andrade, F.M., Young, M.P., MacLeod, D., Hirons, L.C., Woolnough, S.J. and Black, E. (2021) Subseasonal precipitation prediction for Africa: Forecast evaluation and sources of predictability. Weather and forecasting 36(1), 265-284.

Ebert, E. and McBride, J. (2000) Verification of precipitation in weather systems: Determination of systematic errors. Journal of Hydrology 239(1-4), 179-202.

Fan, X., Miao, C., Duan, Q., Shen, C. and Wu, Y. (2021) Future climate change hotspots under different 21st century warming scenarios. Earth's Future 9(6), e2021EF002027.

Faridzad, M., Yang, T., Hsu, K., Sorooshian, S., & Xiao, C. (2018). Rainfall frequency analysis for ungauged regions using remotely sensed precipitation information. Journal of hydrology, 563, 123-142.

Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R. (2004) Random forest classification of multisource remote sensing and geographic data, pp. 1049-1052, IEEE.

Gowan, T.M., Steenburgh, W.J. and Schwartz, C.S. (2018) Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. Weather and forecasting 33(3), 739-765.

Guo, Z., Dirmeyer, P.A. and DelSole, T. (2011) Land surface impacts on subseasonal and seasonal predictability. Geophysical Research Letters 38(24).

Hess, P. and Boers, N. (2022) Deep learning for improving numerical weather prediction of heavy rainfall. Journal of Advances in Modeling Earth Systems 14(3), e2021MS002765.

Hsiao, W.T., Maloney, E.D. and Barnes, E.A. (2020) Investigating recent changes in MJO precipitation and circulation in multiple reanalyses. Geophysical Research Letters 47(22), e2020GL090139.

Karl, T. and Koss, W.J. (1984) Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983.

Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., ... & Hong, Y. (2021). Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study of four watersheds with different hydro-climatic regions across the CONUS. Journal of Hydrology, 598, 126423.

Kim, T., Yang, T., Zhang, L. and Hong, Y. (2022) Near real-time hurricane rainfall forecasting using convolutional neural network models with Integrated Multi-satellitE Retrievals for GPM (IMERG) product. Atmospheric Research 270, 106037.

King, A.D., Hudson, D., Lim, E.P., Marshall, A.G., Hendon, H.H., Lane, T.P. and Alves, O. (2020) Sub‐seasonal to seasonal prediction of rainfall extremes in Australia. Quarterly Journal of the Royal Meteorological Society 146(730), 2228-2249.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P. and Becker, E. (2014) The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bulletin of the American Meteorological Society 95(4), 585-601.

Kuligowski, R.J. and Barros, A.P. (1998) Experiments in short-term precipitation forecasting using artificial neural networks. Monthly weather review 126(2), 470-482.

Kunkel, K.E., Easterling, D.R., Kristovich, D.A., Gleason, B., Stoecker, L. and Smith, R. (2012) Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States. Journal of Hydrometeorology 13(3), 1131-1141.

Kunkel, K.E., Easterling, D.R., Redmond, K. and Hubbard, K. (2003) Temporal variations of extreme precipitation events in the United States: 1895–2000. Geophysical Research Letters 30(17).

Li, W., Chen, J., Li, L., Chen, H., Liu, B., Xu, C.-Y. and Li, X. (2019) Evaluation and bias correction of S2S precipitation for hydrological extremes. Journal of Hydrometeorology 20(9), 1887-1906.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W. and Di, Z. (2017) A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. Wiley Interdisciplinary Reviews: Water 4(6), e1246.

Li, W., Pan, B., Xia, J. and Duan, Q. (2022) Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. Journal of Hydrology 605, 127301.

Madsen, H., Lawrence, D., Lang, M., Martinkova, M. and Kjeldsen, T. (2014) Review of trend analysis and climate change projections of extreme precipitation and floods in Europe. Journal of Hydrology 519, 3634-3650.

Mahoney, K., Jackson, D.L., Neiman, P., Hughes, M., Darby, L., Wick, G., White, A., Sukovich, E. and Cifelli, R. (2016) Understanding the role of atmospheric rivers in heavy precipitation in the southeast United States. Monthly Weather Review 144(4), 1617-1632.

Manzanas, R., Lucero, A., Weisheimer, A. and Gutiérrez, J.M. (2018) Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? Climate Dynamics 50(3), 1161-1176.

Mao, Y. and Sorteberg, A. (2020) Improving radar-based precipitation nowcasts with machine learning using an approach based on random forest. Weather and forecasting 35(6), 2461-2478.

Maraun, D. (2013) Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. Journal of Climate 26(6), 2137-2143.

Mariotti, A., Baggett, C., Barnes, E.A., Becker, E., Butler, A., Collins, D.C., Dirmeyer, P.A., Ferranti, L., Johnson, N.C. and Jones, J. (2020) Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. Bulletin of the American Meteorological Society 101(5), E608-E625.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N. and Kuczera, G. (2020) Multi-temporal hydrological residual error modeling for seamless subseasonal streamflow forecasting. Water Resources Research 56(11), e2019WR026979.

Miao, Q., Pan, B., Wang, H., Hsu, K. and Sorooshian, S. (2019) Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network. Water 11(5), 977.

Mizukami, N. and Smith, M.B. (2012) Analysis of inconsistencies in multi-year gridded quantitative precipitation estimate over complex terrain and its impact on hydrologic modeling. Journal of Hydrology 428, 129-141.

Moon, S.-H., Kim, Y.-H., Lee, Y.H. and Moon, B.-R. (2019) Application of machine learning to an early warning system for very short-term heavy rainfall. Journal of Hydrology 568, 1042-1054.

Nayak, M.A. and Ghosh, S. (2013) Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. Theoretical and Applied Climatology 114(3), 583-603.

Nie, J., Dai, P. and Sobel, A.H. (2020) Dry and moist dynamics shape regional patterns of extreme precipitation sensitivity. Proceedings of the National Academy of Sciences 117(16), 8757-8763.

Pan, B., Anderson, G.J., Goncalves, A., Lucas, D.D., Bonfils, C.J., Lee, J., Tian, Y. and Ma, H.Y. (2021) Learning to correct climate projection biases. Journal of Advances in Modeling Earth Systems 13(10), e2021MS002509.

Pan, B., Hsu, K., AghaKouchak, A. and Sorooshian, S. (2019) Improving precipitation estimation using convolutional neural network. Water Resources Research 55(3), 2301-2321.

Pegion, K., Kirtman, B.P., Becker, E., Collins, D.C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D. and Zhu, Y. (2019) The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. Bulletin of the American Meteorological Society 100(10), 2043-2060.

Pendergrass, A.G. (2020) Changing degree of convective organization as a mechanism for dynamic changes in extreme precipitation. Current climate change reports 6(2), 47-54.

Prat, O. and Nelson, B. (2015) Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). Hydrology and Earth System Sciences 19(4), 2037-2056.

Quedi, E.S. and Fan, F.M. (2020) Sub seasonal streamflow forecast assessment at large-scale basins. Journal of Hydrology 584, 124635.
Robertson, D., Shrestha, D. and Wang, Q. (2013) Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrology and Earth System Sciences 17(9), 3587-3603.
Sadeghi, M., Nguyen, P., Hsu, K. and Sorooshian, S. (2020) Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. Environmental Modelling & Software 134, 104856.
Singh, O. and Kumar, M. (2013) Flood events, fatalities and damages in India from 1978 to 2006. Natural hazards 69(3), 1815-1834.
Sorooshian, S., AghaKouchak, A., Arkin, P., Eylander, J., Foufoula-Georgiou, E., Harmon, R., Hendrickx, J.M., Imam, B., Kuligowski, R. and Skahill, B. (2011) Advancing the remote sensing of precipitation. Bulletin of the American Meteorological Society 92(10), 1271-1272.
Speiser, J.L., Miller, M.E., Tooze, J. and Ip, E. (2019) A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications 134, 93-101.
Srinivas, C., Yesubabu, V., Prasad, D.H., Prasad, K.H., Greeshma, M., Baskaran, R. and Venkatraman, B. (2018) Simulation of an extreme heavy rainfall event over Chennai, India using WRF: Sensitivity to grid resolution and boundary layer physics. Atmospheric Research 210, 66-82.
Stensrud, D.J. (2009) Parameterization schemes: keys to understanding numerical weather prediction models, Cambridge University Press.
Stockdale, T.N., Molteni, F. and Ferranti, L. (2015) Atmospheric initial conditions and the predictability of the Arctic Oscillation. Geophysical Research Letters 42(4), 1173-1179.
Suarez, P., Anderson, W., Mahal, V. and Lakshmanan, T. (2005) Impacts of flooding and climate change on urban transportation: A systemwide performance assessment of the Boston Metro Area. Transportation Research Part D: transport and environment 10(3), 231-244.
Sun, Q., Miao, C., Qiao, Y. and Duan, Q. (2017) The nonstationary impact of local temperature changes and ENSO on extreme precipitation at the global scale. Climate Dynamics 49(11), 4281-4292.
Tang, C., Garreau, D. and von Luxburg, U. (2018) When do random forests fail? Advances in neural information processing systems 31.
Tao, Y., Yang, T., Faridzad, M., Jiang, L., He, X., & Zhang, X. (2018). Non-stationary bias correction of monthly CMIP5 temperature projections over China using a residual-based bagging tree model. International Journal of Climatology, 38(1), 467-482.
Taylor, J., man Lai, K., Davies, M., Clifton, D., Ridley, I. and Biddulph, P. (2011) Flood management: prediction of microbial contamination in large-scale floods in urban environments. Environment international 37(5), 1019-1029.
Thomas, J.A., Berg, A.A. and Merryfield, W.J. (2016) Influence of snow and soil moisture initialization on sub-seasonal predictability and forecast skill in boreal spring. Climate Dynamics 47(1), 49-65.
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(3), 1477-1490.

VanBuskirk, O., Ćwik, P., McPherson, R.A., Lazrus, H., Martin, E., Kuster, C. and Mullens, E. (2021) Listening to Stakeholders: Initiating Research on Subseasonal-to-Seasonal Heavy Precipitation Events in the Contiguous United States by First Understanding What Stakeholders Need. Bulletin of the American Meteorological Society 102(10), E1972-E1986.

Vigaud, N., Robertson, A.W. and Tippett, M.K. (2017) Multimodel ensembling of subseasonal precipitation forecasts over North America. Monthly Weather Review 145(10), 3913-3928.

Vitart, F. (2004) Monthly forecasting at ECMWF. Monthly Weather Review 132(12), 2761-2779.

Vitart, F. (2014) Evolution of ECMWF sub-seasonal forecast skill scores. Quarterly Journal of the Royal Meteorological Society 140(683), 1889-1899.

Vitart, F. (2017) Madden—Julian Oscillation prediction and teleconnections in the S2S database. Quarterly Journal of the Royal Meteorological Society 143(706), 2210-2220.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E. and Fuentes, M. (2017) The subseasonal to seasonal (S2S) prediction project database. Bulletin of the American Meteorological Society 98(1), 163-173.

Wang, F., Tian, D., Lowe, L., Kalin, L. and Lehrter, J. (2021) Deep learning for daily precipitation and temperature downscaling. Water Resources Research 57(4), e2020WR029308.

Wang, L. and Robertson, A.W. (2019) Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. Climate Dynamics 52(9), 5861-5875.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J. and Murray, V. (2017) Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorological applications 24(3), 315-325.

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

Wood, A.W. and Lettenmaier, D.P. (2006) A test bed for new seasonal hydrologic forecasting approaches in the western United States. Bulletin of the American Meteorological Society 87(12), 1699-1712.

Wu, W., Emerton, R., Duan, Q., Wood, A.W., Wetterhall, F. and Robertson, D.E. (2020) Ensemble flood forecasting: Current status and future opportunities. Wiley Interdisciplinary Reviews: Water 7(3), e1432.

Xiang, B., Lin, S.J., Zhao, M., Johnson, N.C., Yang, X. and Jiang, X. (2019) Subseasonal week 3–5 surface air temperature prediction during boreal wintertime in a GFDL model. Geophysical Research Letters 46(1), 416-425.

Yang, T., Gao, X., Sellars, S. L., & Sorooshian, S. (2015). Improving the multi-objective evolutionary optimization algorithm for hydropower reservoir operations in the California Oroville–Thermalito complex. Environmental Modelling & Software, 69, 262-279.

Yang, T., Gao, X., Sorooshian, S. and Li, X. (2016) Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. Water Resources Research 52(3), 1626-1651.

Yang, T., Asanjan, A. A., Welles, E., Gao, X., Sorooshian, S., & Liu, X. (2017a). Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resources Research, 53(4), 2786-2812.

Yang, T., Asanjan, A. A., Faridzad, M., Hayatbini, N., Gao, X., & Sorooshian, S. (2017b). An enhanced artificial neural network with a shuffled complex evolutionary global optimization with principal component analysis. Information Sciences, 418, 302-316.

Yang, T., Tao, Y., Li, J., Zhu, Q., Su, L., He, X., & Zhang, X. (2018). Multi-criterion model ensemble of CMIP5 surface air temperature over China. Theoretical and Applied Climatology, 132(3), 1057-1072.as

Yang, T., Liu, X., Wang, L., Bai, P., & Li, J. (2020). Simulating hydropower discharge using multiple decision tree methods and a dynamical model merging technique. Journal of Water Resources Planning and Management, 146(2), 04019072.

Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D. and Peng, Q. (2021) A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region. Journal of Hydrology 602, 126723.

Yang, Z., Li, B., Xia, R., Ma, S., Jia, R., Ma, C., Wang, L., Chen, Y. and Bin, L. (2022) Understanding China's industrialization driven water pollution stress in 2002–2015—A multi-pollutant based net gray water footprint analysis. Journal of Environmental Management 310, 114735.

Yuan, X., Wood, E.F. and Ma, Z. (2015) A review on climate - model - based seasonal hydrologic forecasting: physical understanding and system development. Wiley Interdisciplinary Reviews: Water 2(5), 523-536.

Zhang, C. and Ling, J. (2017) Barrier effect of the Indo-Pacific Maritime Continent on the MJO: Perspectives from tracking MJO precipitation. Journal of Climate 30(9), 3439-3459.

Zhang, D., Lin, J., Peng, Q., Wang, D., Yang, T., Sorooshian, S., Liu, X. and Zhuang, J. (2018) Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm. Journal of Hydrology 565, 720-736.

Zhang, L., Kim, T., Yang, T., Hong, Y. and Zhu, Q. (2021) Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous US. Journal of Hydrology 603, 127058.

Zhao, T., Bennett, J.C., Wang, Q., Schepen, A., Wood, A.W., Robertson, D.E. and Ramos, M.-H. (2017) How suitable is quantile mapping for postprocessing GCM precipitation forecasts? Journal of Climate 30(9), 3185-3196.

Chapter 5. Combining Machine Learning with a Distribution-based Technique to Adapt Subseasonal-to-Seasonal (S2S) Precipitation Forecasts at Watershed Scale for Hydrologic Ensemble Streamflow Prediction (ESP)

The major content from this chapter will be submitted to the *Journal of Hydrology* as

Zhang, L., Yang, T., Gao, S. & Hong, Y. (2023). Combining Machine Learning with a distribution-based technique to adapt Subseasonal-to-Seasonal (S2S) precipitation forecasts at watershed scale for hydrologic Ensemble Streamflow Prediction (ESP). *In preparation*.

Abstract

Accurate and reliable precipitation predictions made by dynamical forecast models could provide crucial information for human socioeconomic activities by enabling hydrologic forecasts at a Subseasonal-to-Seasonal (S2S) timescale. To utilize available S2S precipitation predictions for hydrologic forecasting, post-processing techniques have been applied to adapt the raw S2S precipitation to local watersheds. However, conventional statistical-based post-processing techniques are more focused on correcting the forecast bias, which is rather limited in improving the temporal predictive skill of available S2S precipitation forecasts. In this chapter, I combine the Random Forest classifiers (RF) with the Bias Correction and Spatial Disaggregation (BCSD) to adapt the raw S2S precipitation forecast from the NASA Goddard Earth Observing System model version 5 (GEOS5) to 4 watersheds located in the NCEI South climate region of the United States. The adapted S2S precipitation is further applied for streamflow forecast experiments under the Ensemble Streamflow Prediction (ESP) framework. Two evaluation statistics of Kling-Gupta Efficiency (KGE) and Continuous Ranked Probability Skill (CRPS) are employed to evaluate the predictive skill of precipitation and streamflow both deterministically and probabilistically. My results indicate that the application of RF can improve the predictive skill of the raw S2S precipitation from GEOS5 at a typical watershed

scale, while the forecast bias is also effectively removed. Further, the added predictive skill of S2S precipitation propagates into streamflow forecasts through the ESP framework, predominantly at longer forecast lead times exceeding 10 days. Overall, my experiment results highlight the success of using a data-driven approach to adapt the raw S2S precipitation forecasts to local watersheds, achieving more accurate and reliable streamflow prediction at S2S timescale than the conventional baseline ESP method.

5.1 Introduction

Accurate and reliable streamflow predictions at a Subseasonal-to-Seasonal (S2S) timescale, defined explicitly as the transitional period of 10 to 30 days between weather predictions and seasonal outlooks (Vitart et al., 2017), could provide crucial information benefiting human socio-economic activities, including reservoir operation, irrigation planning, and proactive flood control (Chiew et al., 2003; White et al., 2017; Yang et al., 2021).

The Ensemble Streamflow Prediction (ESP) framework is commonly adopted for streamflow predictions (Day, 1985; Schaake and Larson, 1998). There are mainly two key steps to apply ESP for hydrologic forecasts at S2S timescale. First, a calibrated hydrologic model is forced with measured hydrometeorological information up to the "current" time step to estimate the initial hydrologic conditions (IHCs). Then, with the estimated IHCs, the calibrated hydrologic model is further forced with multiple randomly resampled historical precipitation measurements to predict streamflow volumes probabilistically over the forecast horizon of interest. While exploiting the current understanding of hydrology through the application of a chosen hydrologic model, the ESP considers the predictability of streamflow from both the IHCs (i.e., initial soil moisture and/or snow condition) and future hydrometeorological conditions. By incorporating an ensemble of resampled historical precipitation measurements,

probabilistic predictions on future extreme events (i.e., floods) can be achieved (Delaney et al., 2020; Harrigan et al., 2018; Troin et al., 2021).

However, one common challenge persists when predicting streamflow at S2S timescale under the ESP framework. That is, when the forecast lead time is long enough (typically beyond 5-10 days), the predictability of streamflow coming from the IHCs would be lost, leaving the volumetric streamflow prediction dominated by inaccurate resampled precipitation forecasts. As a result, the corresponding streamflow predictions usually present little deterministic predictive skill (Cao et al., 2021; Shukla et al., 2013; Wood and Lettenmaier, 2008).

Advancements in weather/climate predictions at S2S timescale have provided an alternative to the resampled precipitation in the ESP framework. Many dynamical S2S precipitation forecast products have been made available by different mission agencies over the globe in the recent decade (Kirtman et al., 2014; Pegion et al., 2019; Richter et al., 2022; Vitart et al., 2017). These dynamical S2S precipitation forecast products are generated by General Circulation Models (GCMs) coupled with ocean and land surface components with a consideration of the "real-time" condition of the Earth System. Therefore, available S2S precipitation forecasts should be more reliable and accurate compared to the randomly resampled precipitation, which could lead to overall better hydrologic predictions at S2S timescale under the ESP framework.

However, according to several previous evaluation studies, the raw S2S precipitation forecasts are generally associated with a substantial amount of forecast bias and only present a marginal level of predictive skill when forecast lead time exceeds 2 weeks (de Andrade et al., 2021; Tian et al., 2017; Zhang et al., 2021). Combining the fact that the raw S2S precipitation forecasts often come with a spatial resolution of around 1 degree (~100 Km), post-processing

techniques need to be applied to adapt the raw S2S precipitation to a typical watershed scale before any potential hydrologic applications.

To further verify the effectiveness of available S2S precipitation in hydrologic forecasts under the ESP framework, Zhang et al. (2023a) applied a popular and rather standard post-processing technique, termed Bias Correction and Spatial Disaggregation (i.e., BCSD, originally developed by Wood et al. (2004)), to adapt the raw S2S precipitation forecast from the North America Multi-Model Ensemble Phase II (NMME-2) to 4 USGS Hydrologic Unit Code 8 (HUC-8) level watersheds in the NCEI South climate region (i.e., Oklahoma, Texas, Arkansas, and Louisiana) for streamflow hindcast experiments. The results show that the BCSD-adapted S2S precipitation leads to overall better streamflow forecasts than the resampled precipitation does under the ESP framework. However, such an advantage tends to be marginal once the forecast lead time exceeds ~10 days. Zhang et al. (2023a) also highlighted that novel post-processing techniques should be developed to further improve the predictive skill of the raw S2S precipitation for more accurate S2S streamflow predictions.

Machine Learning (ML)-based techniques are promising alternatives to conventional post-processing techniques such as BCSD to further improve the predictive skill of S2S precipitation forecasts from a post-processing perspective. Most recently, Zhang et al. (2023b) applied the Random Forest classifiers (RF) to incorporate temperature and geopotential height forecasts when post-processing the raw S2S precipitation forecasts from NASA Goddard Earth Observing System Model Version 5 (i.e., GEOS5, one contributing model to the NMME-2 project) over CONUS. According to Zhang et al. (2023b), the proposed RF can improve the predictive skill of the S2S precipitation in predicting the occurrences of weekly extreme precipitation events above 99% over the entire contiguous United States (CONUS). This

highlights the potential success of various data-driven ML applications for more skillful S2S precipitation forecasts, which could enable more accurate and reliable streamflow forecasts at S2S timescale.

However, the RF framework proposed by Zhang et al. (2023b) is due for a verification of its effectiveness at typical watershed scale and under the ESP framework, for primarily two reasons. Firstly, Zhang et al. (2023b) carried out over the entire CONUS, which exceeds the conventional hydrologic scale. Further, in Zhang et al. (2023b), only categorical information was provided with the application of RF but without any volumetric information (which is critically needed for hydrologic forecasting). It is thus unclear whether the proposed RF can be applied at local watersheds for S2S streamflow predictions under the ESP framework.

Therefore, in this chapter, I aim to verify the effectiveness of ML in adapting raw S2S precipitation at a typical watershed scale for daily S2S streamflow predictions under the ESP framework. To be more specific, I intend to address the following research questions: ***(1) Can an ML technique improve the predictive skill of daily S2S precipitation forecasts at typical watershed scales through post-processing? (2) Can ML lead to overall more skillful streamflow prediction under the ESP framework, in contrast to a baseline post-processing technique? And (3) If ML can indeed improve the predictive skill of S2S precipitation forecasts, how does the added skill of the S2S precipitation forecast propagate into streamflow predictions under the ESP framework at different forecast lead times (i.e., augmented or compressed)?***

To answer the above-mentioned research questions and also extend two previous works by Zhang et al. (2023a) and Zhang et al. (2023b), I further apply the RF model at a typical watershed scale for streamflow forecasting under the ESP framework. Specifically, the RF

framework proposed by Zhang et al. (2023b) is jointly employed with BCSD to adapt the raw S2S precipitation from GEOS5 to the same 4 watersheds studied by Zhang et al. (2023a). The 4 study watersheds are located in the NCEI South climate region, where S2S precipitation appears to perform consistently worse than it is in other climate regions. The underlying logic of selecting these 4 study watersheds in the NCEI South climate region is that if the proposed approach is proven to be effective at the selected watersheds, then it should be more likely to show at least the same level of effectiveness when transferred to watersheds located in other climate regions. The adapted S2S precipitation is further passed to the ESP framework for streamflow forecasting experiments at the selected study watersheds to verify its hydrologic performance. Finally, the "elasticities" between the added skill of precipitation and the corresponding added skill of streamflow forecasts brought by RF are computed for quantification and interpretation. This chapter covers a period from 1982 to 2011 (30 years).

In this chapter, the RF is employed to give categorical corrections on the occurrence of high-percentile extreme precipitation events only, leaving the volumetric information of the S2S precipitation to be further corrected by BCSD. There are primarily two reasons for such a design. The first reason is that previous works show ML techniques tend to underestimate high percentile values (Akbari Asanjan et al., 2018; Kim et al., 2022). Another and perhaps more important reason is that I do not have a sufficient number of input-measurement pairs but only a relatively limited number of hindcast experiments. I believe that, under this circumstance, the application of RF could be prone to overfitting if directly targeted on numerical values during its training (Caruana et al., 2000; Karras et al., 2020). Therefore, I complete the training of RF models to predict the occurrence of high-percentile extreme events first and then correct the volumetric information of precipitation forecast with the classical BCSD method.

The rest of this chapter is organized as follows: In **Section 5.2**, I first introduce data and study watersheds. The experiment settings, methodologies, and evaluation metrics are described in **Section 5.3**. **Section 5.4** presents the results. The discussions and main conclusions are presented and summarized in **Section 5.5** and **Section 5.6**, respectively.

5.2 Datasets and Study Regions

5.2.1 Datasets

Three measurement/reanalysis hydrometeorological datasets and one S2S forecast product are used in this chapter, including (1) the daily AN81d precipitation dataset from the Parameter-elevation Regressions on the Independent Slopes Model (PRISM), (2) Daily potential evapotranspiration (PET) reanalysis dataset from the North American Regional Reanalysis (NARR), (3) Daily streamflow measurements from the National Water Information System at the United States Geological Survey (USGS), and (4) Multiple S2S hindcast variables from one contributing model to the NMME-2 data archive, namely the NASA Goddard Earth Observing System Model version 5 (GEOS5). All datasets used in this chapter cover a common period from 01/01/1982 to 12/31/2011, constituting a 30-year study period.

Multiple S2S hindcasts variables from GEOS5 are used in this chapter, including precipitation, temperature, and geopotential height at 500 hope and 850 hPa. The GEOS5 produces 10-member ensemble forecasts/hindcasts on the first day of each month throughout the study period. Each member of the GEOS5 dataset provides daily hindcasts of multiple hydrometeorological variables with a lead time of up to 274 days. The 10-member ensemble hindcasts are generated with perturbed initial conditions (Borovikov et al., 2019). In addition, all collected S2S hindcasts variables are truncated to 28 days (4 weeks) for this chapter is focus

on S2S timescale specifically. More details about the datasets used in this chapter are presented in Table 5.1.

Table 5.1 Hydrometeorological datasets used in Chapter 5.

	Name	Spatial resolution	Sources
Measurements /Reanalysis	Precipitation (PRISM)	~0.07°	https://prism.oregonstate.edu/
	PET (NARR)	~0.3°	https://www.emc.ncep.noaa.gov/mmb/rreanl/
	Streamflow (USGS)	N/A	https://waterdata.usgs.gov/nwis
GEOS5 S2S hindcasts	Precipitation	1°	https://www.cpc.ncep.noaa.gov/products/NMME/data.html
	Temperature	1°	
	500 hPa Geopotential	1°	
	850 hPa Geopotential	1°	

5.2.2 Study Watersheds

In this chapter, the same four watersheds in the South climate region of the United States (per NCEI definition) studied by Zhang et al. (2023a) are selected as the experiment watersheds. Figure 1 shows the locations, elevation, and river networks of the selected watersheds. The selected watersheds are considered less affected by human activities, which makes them suitable study watersheds in the region (Newman et al., 2015).

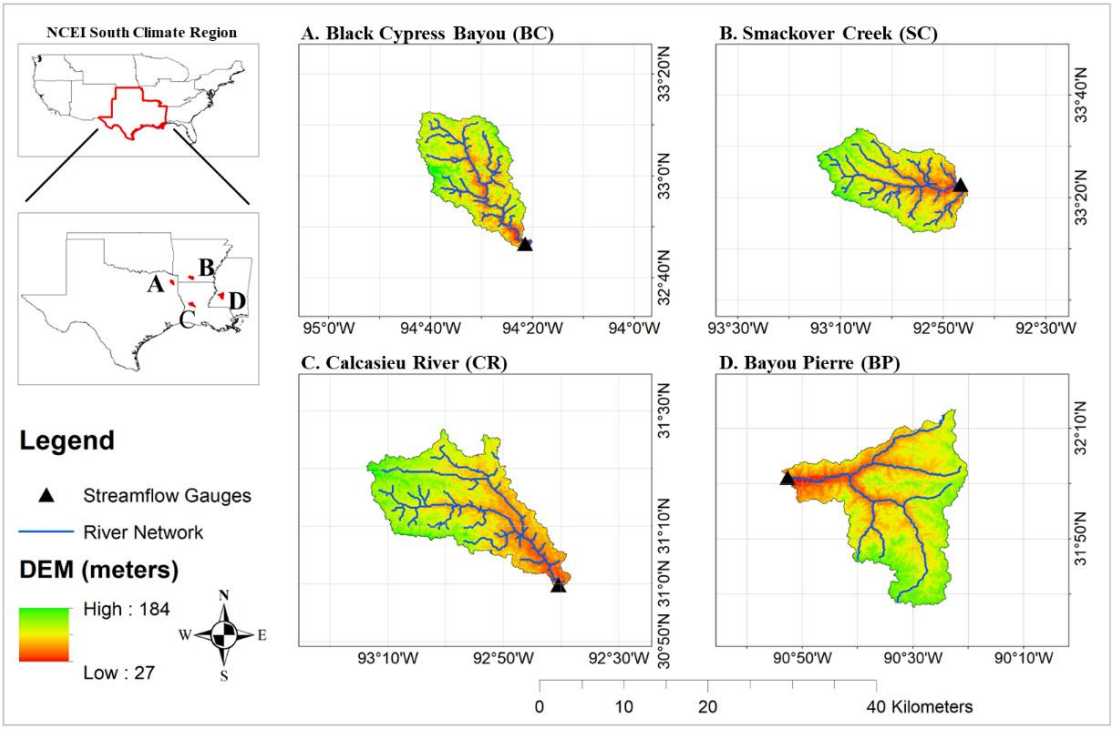

Figure 5.1 Map of the 4 study watersheds (BC is in Texas; SC is in Arkansas; CR is in Louisiana; BP is in Missouri).

The 4 selected watersheds are dry, with a major proportion of precipitation occurring in cold seasons and little snow (less than 2% of the total amount of precipitation). More detailed information about the study watersheds is presented in Table 5.2.

Table 5.2 Topography and hydrometeorology information of the study watersheds in Chapter 5.

Watershed Names	Drainage Area (Sq-KM)	Elevation (Meters)	Aridity (PET/P)	Snow Fraction	USGS Gauge ID
Black Cypress Bayou (BC)	959.89	105.07	0.88	0.017	7346045
Smackover Creek (SC)	996.08	79.29	0.80	0.018	7362100
Calcasieu River (CR)	1293.69	81.79	0.75	0.003	8013000
Bayou Pierre (BP)	1688.73	102.05	0.82	0.004	7290650

5.3 Experiment Setting and Evaluation Metrics

5.3.1 General Experiment Setting

This chapter employs two post-processing schemes to adapt the raw S2S precipitation forecasts from GEOS5 at the 4 selected watersheds. The first post-processing scheme (i.e., the

baseline scheme) is the BCSD post-processing. To apply BCSD, the raw precipitation forecasts are bilinearly interpolated into 0.07-degree pixels first (i.e., the spatial resolution of PRISM). Then, bias corrections are conducted at each interpolated pixel using Multi-Segmented Quantile Mapping (i.e., MSQM, originally introduced by Grillakis et al. (2013)). The second post-processing scheme is a combination of BCSD and RF (hereafter referred to as "RF-BCSD"). Under the second post-processing scheme, the raw precipitation forecasts are also bilinearly interpolated into 0.07-degree pixels first. Only this time, the RF is applied to correct the interpolated precipitation forecasts at each interpolated pixel first before the application of MSQM to remove forecast bias. Precipitation adapted under both schemes is eventually transformed into mean areal values over the study watersheds for evaluation as well as next-step events. More details about the two post-processing schemes of BCSD and RF-BCSD are introduced in the following **sections 5.3.2** and **5.3.3**.

The adapted S2S precipitation is then passed to the ESP framework for a series of streamflow hindcast experiments. The resulting ESP streamflow hindcasts are evaluated against the proxies of streamflow measurements (i.e., streamflow simulations generated by forcing the calibrated hydrologic model with precipitation and PET measurements). The usage of such streamflow proxies instead of measured streamflow during the evaluation is for later quantification of the propagation of the added skill from precipitation to streamflow, without mixing the effects coming from other sources, e.g., the imperfect structure of the hydrologic model, uncertainty arises from parameters of the hydrologic model, etc. The overall experiment design is depicted in Figure 5.2.

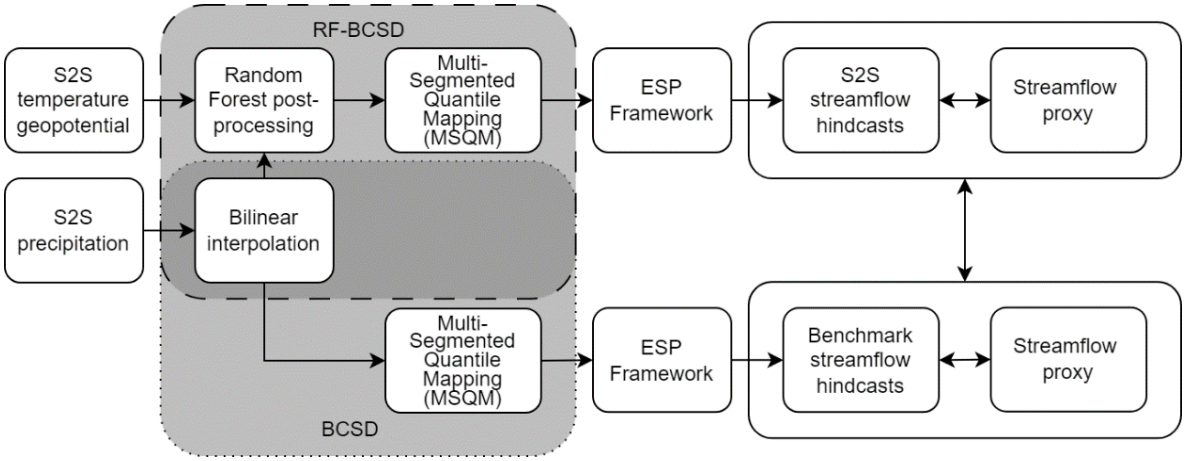

Figure 5.2 A schematic diagram of the experiment design.

In this chapter, I choose the Sacramental Soil Moisture Accounting (Sac-SMA) model as the hydrologic component of the ESP framework. The Sac-SMA is calibrated at each study watershed first using the SC-SAHEL algorithm developed by Naeini et al. (2018). To apply ESP, the calibrated Sac-SMA needs to be forced with hydrometeorological measurements up to the "current time" to get the best estimation of the "initial hydrological conditions (IHCs)" of the watershed. Then, an ensemble of streamflow forecasts is generated through a combination of the calibrated Sac-SMA, the estimated IHCs, as well as the post-processed S2S precipitation. Since there are two groups of S2S precipitation from two post-processing scenarios, a total number of 720 streamflow hindcast experiments, i.e., 2 groups of precipitation × 12 months × 30 years, are conducted at each of the study watersheds during the entire study period. Such ESP streamflow hindcasts are initialized on the 1st day of each month during the study period. Note that in this chapter, the actual PET reanalysis was used for the calibration and spin-up of the Sac-SMA, while the PET climatology derived for each month was used for streamflow forecasts to focus on evaluating the effectiveness of different post-processing schemes of the S2S precipitation.

The evaluation of both precipitation and streamflow forecasts in this chapter is carried out with the Kling-Gupta Efficiency (KGE) and the Continuous Ranked Probability Skill (CRPS) at different forecast lead times and across different study watersheds. More details about KGE and CRPS are described in **Section 5.3.4**.

5.3.2 S2S Forecast Adaptation Schemes

For both post-processing schemes, the "leave-10-years-out" cross-validation strategy is employed. The whole study period was divided into three 10-year periods (i.e., 1982-1991, 1992-2001, and 2002-2011). When a particular 10 years are selected as the targeting period for forecast adaptation, the remaining 20 years' data will be used to train the post-processing/correction model. Through such reference/correction period portioning, the entire data record of S2S precipitation is cross-corrected. During the post-processing/correction under both schemes, the data samples with similar forecast lead times are pooled together when training the post-processing/correction models. To be more specific, the S2S data and PRISM from the training period are divided into 4 groups based on 4 weeks of forecast lead times (i.e., 1 to 7 days, 8 to 14 days, 15 to 21 days, and 22 to 28 days). The post-processed S2S precipitation under both schemes is eventually transformed into mean areal values for evaluation and next-step streamflow hindcast experiments.

5.3.2.1 Bias Correction and Spatial Disaggregation (BCSD)

In this section, the first post-processing scheme of BCSD is introduced. As briefly mentioned in the previous section, there are generally two steps when applying BCSD to post-process precipitation forecasts. Firstly, the raw S2S precipitation is first bilinearly interpolated into 0.07° pixels (i.e., the spatial resolution of PRISM). Then, bias corrections are conducted at each interpolated 0.07° pixel using the Multi-Segmented Quantile Mapping (MSQM).

To apply MSQM at a certain interpolated pixel, PRISM and S2S precipitation from the reference period will be sorted first. After the sorting of the datasets, the $i\text{-}th$ rank of PRISM precipitation is identified by counting the total number of rainy days (precipitation above zero). The corresponding value at the same $i\text{-}th$ rank of the S2S precipitation is then identified as the dry threshold since GCM normally generates lots of extremely small non-zero precipitation values. During the correction period, all the S2S precipitation values smaller than the dry threshold value will be set to zeros to make sure forecast precipitation will have the same precipitation frequency compared to the reference precipitation. Then, another $j\text{-}th$ rank of the PRISM precipitation is identified by finding out the threshold value above 97% according to the statistics of all non-zero values of the PRISM. The corresponding value at the same j-th rank of the S2S precipitation is seen as the extreme value threshold. A first-order linear regression relationship is built by utilizing both S2S precipitation and reference precipitation above the extreme value threshold. All the S2S precipitation values above the extreme value threshold during the correction period will be corrected utilizing the generalized linear relationship. All the remaining non-zero values below the extreme value threshold from both the PRISM and S2S precipitation during the reference period are then divided equally into 3 segments based on their rankings to fit separate Gamma Cumulative Distribution Functions (CDFs). Taking an arbitrary quantile segment as an example, non-zero values within that segment from both PRISM and S2S precipitation are used to fit two Gamma CDFs, respectively. Utilizing the 12 Gamma CDFs derived during the training period, the raw S2S precipitation is corrected by substituting values at the same quantile according to the reference CDF following equation (5.1):

$$P_{cor} = F_{obs}^{-1}[F_{raw}(P_{raw})] \quad (5.1)$$

Where $Praw$ indicates the raw S2S precipitation from the correction period, $Fraw$ indicates the CDF of raw S2S precipitation derived during the training period, and $Fobs\ {-}1$ indicates the inverse of the CDF of reference precipitation derived during the training period, and $Pcor$ is the corrected S2S precipitation. When applying the MSQM correction of S2S precipitation, I have considered the seasonality as well as forecast lead times at the same time. However, to maintain sufficient sample sizes during the correction, I cannot conduct MSQM within different months and at different lead times at the same time. Instead, I chose to pool the values of S2S precipitation with similar seasonality and forecast lead times together when conducting the MSQM. To be specific, S2S precipitation was pooled together according to their forecast lead times in 4 weeks (i.e., 1-7 days, 8-14 days, 15-21 days, and 21-28 days) and in 4 seasons (i.e., December, January, February; March, April, May; June July August; and September, October, November) for the MSQM corrections. Overall, a total of 48 MSQM models are used based on the combinations of 3 segments, 4 lead time groups, and 4 seasons.

5.3.2.2 Random Forest plus Bias Correction and Spatial Disaggregation (RF-BCSD)

In this section, the second forecast adaptation scheme of RF-BCSD is introduced. Under the second scheme (i.e., the RF-BCSD approach), the RF model and the BCSD approach are jointly employed. To be more specific, the raw S2S precipitation forecast is also bilinearly interpolated into 0.07-degree pixels first. However, before the MSQM bias correction, the RF is employed to correct the raw S2S precipitation for better categorical predictions on the occurrence of high-percentile extreme events.

The RF employed under the second scheme aims to improve the capability of S2S precipitation in predicting the occurrence of daily extreme events. In other words, the RF is trained to make categorical predictions of either (1) no extreme precipitation, (2) extreme

precipitation from 97% to 98.5%, (3) extreme precipitation from 98.5 to 99.5%, (4) extreme precipitation from 99.5% to 99.9%, or (6) extreme precipitation above 99.9% (i.e., totaling 6 categories). The training/correction of RF was carried out while considering different forecast lead times and with the same cross-validation strategy as explained in the previous section. When training the RF classifier, all collected S2S variables of precipitation, temperature, and geopotential height at 500 hPa and 850 hPa are used as inputs. The employment of RF in this chapter is based on the open-source package "scikit-learn" in Python. The hyperparameters of RF are set to values recommended by Zhang et al. (2023b), where the "max_depth" of RF is set to be 9, the "max_features" is set to be 0.6, and the "n_estimators" is set to be 200.

After the RF correction, the positive corrections on extreme events are restored to numerical values based on the statistics of the raw S2S precipitation forecasts. For example, if the RF is making a positive prediction of the occurrence of extreme precipitation within the category of 99.5% to 99.9% on date X, the median value of the quantile segments from 99.5% to 99.9% of the raw S2S precipitation will be placed on date X to replace the original forecast values. Finally, the RF-corrected S2S precipitation forecasts are corrected using the MSQM described in the previous section to thoroughly remove the forecast bias.

5.3.3 Evaluation Statistics

In this chapter, two evaluation statistics are employed to evaluate the predictive skill of both precipitation and streamflow hindcasts. The deterministic Kling-Gupta Efficiency (KGE) and the probabilistic continuous ranked probability skill (CRPS) are computed to quantify the predictive skills of precipitation and streamflow. In addition, the elasticity between the added values of KGE and CRPS (of precipitation and streamflow) are computed to quantitatively measure the propagation of the added skill from precipitation to streamflow hindcasts.

5.3.3.1 Kling-Gupta Efficiency (KGE)

The Kling-Gupta Efficiency (KGE) was originally introduced by Gupta et al. (2009) and has become a widely applied performance evaluation statistic in the field of hydrology ever since. The KGE is computed with three distinct statistics of correlation coefficient (CC), Bias Ratio (BR), and relative variability (RV) altogether. The values of CC, BR, and RV can be computed with the following equations (5.2), (5.3), and (5.4):

$$CC = \frac{\sum_{i=1}^{n}\left((Q_{Sim,i}-\overline{Q}_{Sim})-(Q_{Obs,i}-\overline{Q}_{Obs})\right)}{\sqrt{\sum_{i=1}^{n}(Q_{Sim,i}-\overline{Q}_{Sim})^2 \sum_{i=1}^{n}(Q_{Obs,i}-\overline{Q}_{Obs})^2}} \quad (5.2)$$

$$BR = \frac{\overline{Q}_{Sim}}{\overline{Q}_{Obs}} \quad (5.3)$$

$$RV = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{Sim,i}-\overline{Q}_{Sim})^2}/\overline{Q}_{Sim}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{Obs,i}-\overline{Q}_{Obs})^2}/\overline{Q}_{Obs}} \quad (5.4)$$

In the equations above, n is the total number of data points, $Q_{Sim,i}$ and $Q_{Obs,i}$ are the model simulated/predicted data and observed data, and $\overline{Q}_{Sim}$ and $\overline{Q}_{Obs}$ are the means of the model simulated/predicted and observed data series, respectively. With the computed values of CC, BR, and RV, KGE can be computed with the following equation (5.5):

$$KGE = 1 - \sqrt{(CC-1)^2 + (BR-1)^2 + (RV-1)^2} \quad (5.5)$$

The values of KGE range from $-\infty$ to 1. The ideal value of KGE would be 1, indicating that model simulated/predicted values are perfectly aligned with measurements.

5.3.3.2 Continuous Ranked Probability Score (CRPS)

The Continuous Ranked Probability Score (CRPS) is a widely applied evaluation statistic for the evaluation of probabilistic forecasts of various kinds. Assuming a PDF forecast by an ensemble system X and the corresponding measurement Y, the CRPS between X and Y is defined as the following equation (5.6):

$$CRPS = \frac{1}{N}\sum_{n=1}^{N}\int[F_n(x) - F_n(y)]^2 dx \quad (5.6)$$

In equation (5.6), n represents a certain forecast case, $F_n(x)$ is the empirical PDF of the ensemble forecast system, $F_n(y)$ is the empirical PDF given a measured value. The unit of CRPS is subject to the variable to be evaluated. For example, when applying the CRPS for the evaluation of precipitation, the unit of CRPS should be in length (e.g., mm, cm, etc.). Since the CRPS measures the errors between the forecast PDF and measured PDF, it has a range from o to ∞. The ideal value of CRPS should be 0, indicating that the ensemble forecast is perfect with all its ensemble members making perfect predictions.

5.3.3.3 "Elasticities" between the added skill of precipitation and streamflow hindcasts

The "elasticities" are computed at different forecast lead times to quantity the propagation of the added skill from precipitation to streamflow hindcasts. The computation of elasticity is defined as the following equation (5.7):

$$E = \Delta_{KGE}(Precipitation)/\Delta_{KGE}(Streamflow) \quad (5.7)$$

The skill elasticities are defined as the unit change of streamflow hindcasts skill in correspondence to the unit change of precipitation hindcast skill. Note that when computing skill elasticities, only KGE was used instead of both KGE and CRPS. The reason is that CRPS of precipitation and streamflow hindcasts have different units and therefore cannot be divided by each other.

5.4 Results

5.4.1 Calibration of Sac-SMA

Figure 5.3 presents the streamflow simulations at 4 study watersheds after the calibrations of Sac-SMA. The red lines indicate streamflow measurements, and the blue lines indicate the simulated streamflow. The vertical dashed lines separate the entire study period into the

calibration and validation periods. The KGE values within calibration and validation periods at each study watershed are labeled on the figure as well.

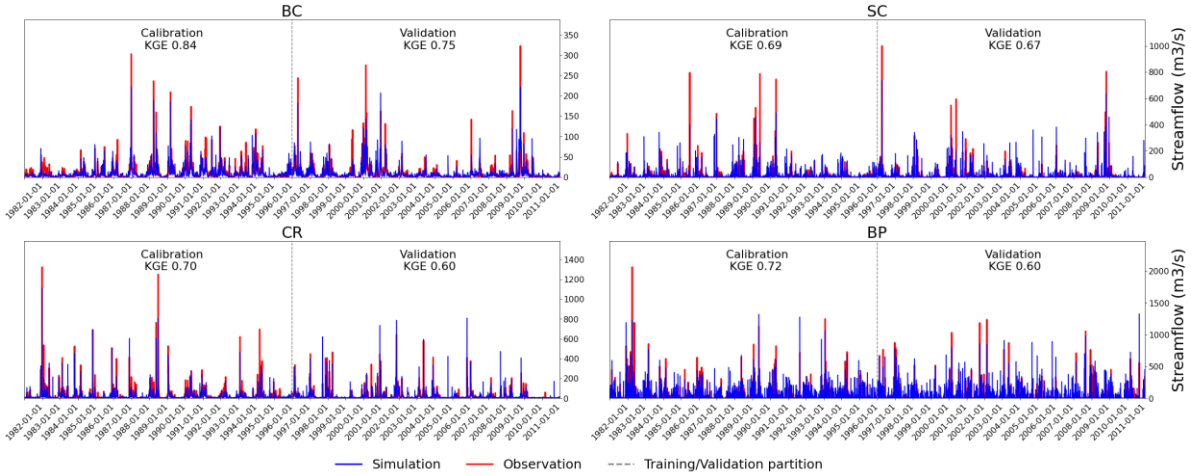

Figure 5.3 Simulated and observed daily streamflow hydrographs at the study watersheds.

From Figure 5.3, it can be observed that after the calibration of Sac-SMA, the high-volume streamflow events seem to be consistently underestimated at all watersheds. But in general, KGE ranges from 0.84 to 0.69 at all watersheds during the calibration period. Although the performance of hydrologic simulations drop during the validation periods, the values of KGE remain above 0.6 at all study watersheds.

5.4.2 Predictive Performance of Precipitation

To examine the quality of the S2S precipitation resulting from the 2 forecast adaptation schemes, both forecast bias and predictive skill are examined. The forecast bias is examined with Quantile-Quantile plots (QQ-plots), and the predictive skill is quantified by KGE and CRPS values.

5.4.2.1 Forecast Bias

In Figure 5.4, the mean areal S2S precipitation hindcasts of all ensemble members from GEOS5 over the 4 study watersheds are plotted as scatter points against the reference PRISM dataset. The 4 columns of Figure 5.4 correspond to the 4 study watersheds. The 3 rows of

Figure 5.4 correspond to the raw GEOS5 S2S precipitation hindcasts (grey-colored dots), the BCSD-adapted S2S precipitation (black-colored dots), and the RF-BCSD adapted S2S precipitation (blue-colored dots), respectively. The ideal S2S precipitation hindcast values would lie perfectly on the red-colored 45-degree lines, which indicates that there's no bias of S2S precipitation at all magnitudes.

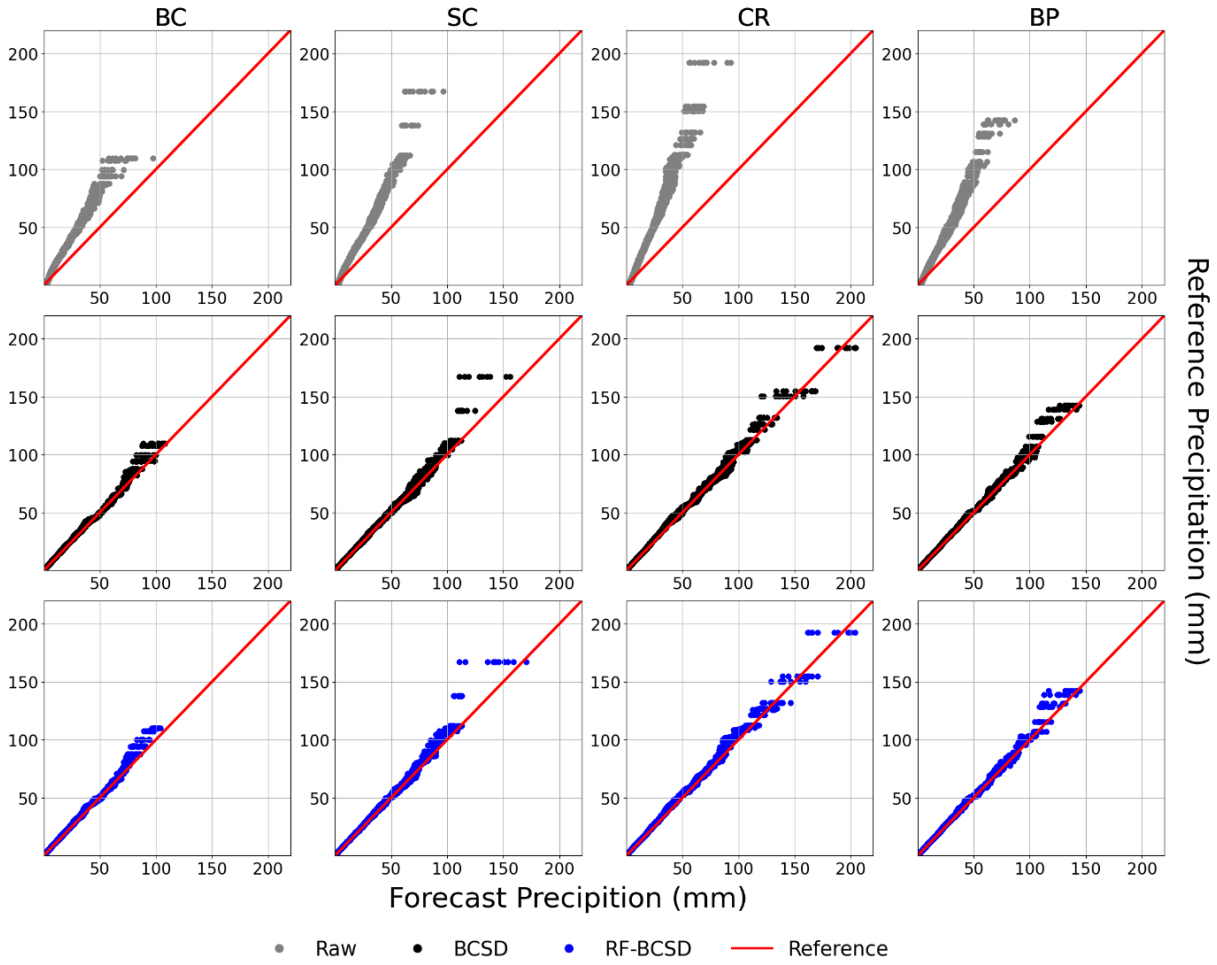

Figure 5.4 QQ-plots of the mean areal S2S precipitation in raw condition, after BCSD post-processing, and after RF-BCSD post-processing.

From Figure 5.4, differences can be observed in terms of the climatology of precipitation. The maximum precipitation ranges from around 120 mm (BC) to 200 mm (CR) across the 4 study watersheds. The raw S2S hindcasts from GEOS5 underestimate precipitation across all study watersheds, as all the grey-colored scatter points lie above the 45-degree reference lines.

The application of BCSD removed the forecast bias very nicely at all magnitudes and across all study watersheds, as most of the black-colored scatter points aligned with the 45-degree reference line much better if compared to the grey-colored scatter points. On the other hand, the joint application of RF and BCSD also removes forecast bias very effectively. No major differences can be identified if comparing second and third-row panels, which indicates that the application of RF does not affect the removal of forecast bias.

5.4.2.2 Predictive Skill

To further quantify the predictive skill of S2S precipitation resulting from BCSD and RF-BCSD adaptations, two evaluation statistics of the KGE and the CRPS are computed. Figure 5.5 presents the deterministic KGE values of the ensemble means of S2S precipitation at the 4 selected study watersheds over the forecast horizon of 28 days. The higher the KGE values, the more skillful the S2S precipitation is. In Figure 5.5, the blue-colored lines are the KGE skill of the S2S precipitation resulting from RF-BCSD adaptation, and the black-colored lines are the KGE skill of the S2S precipitation resulting from BCSD.

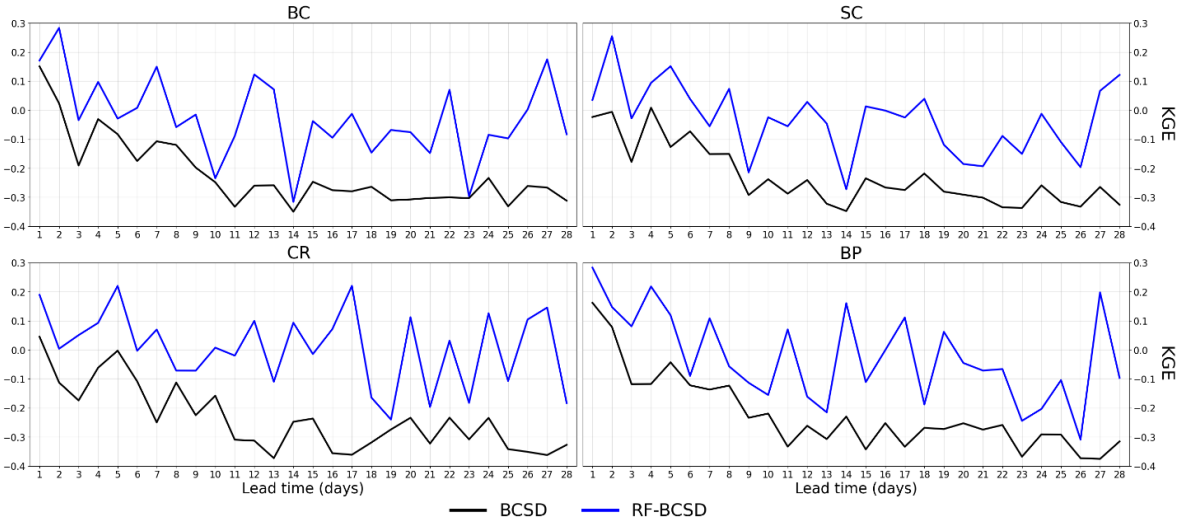

Figure 5.5 KGE skill of the mean areal S2S precipitation after BCSD and RF-BCSD adaptation at 4 study watersheds and over the 28-day forecast horizon.

From Figure 5.5, it can be observed that both S2S precipitation resulting from BCSD and RF-BCSD show a similar decreasing trend across the forecast horizon and over the 4 study watersheds. Comparing S2S precipitation resulting from BCSD and RF-BCSD adaptations, it is evident that the RF-BCSD adaptation leads to consistently higher predictive skills over the entire forecast horizon and across all study watersheds. Over the entire forecast horizon and across different study watersheds, the RF-BCSD seems to be bringing more significant improvement in KGE values at longer forecast lead times. This can be observed as RF-BCSD-associated KGE values are sometimes equivalent to the BCSD-associated KGE values at much shorter lead times (e.g., at the BC watershed, RF-BCSD at day 12 and BCSD at day 1; at the SC watershed, RF-BCSD at days 14-18 and BCSD at days 1-2; at the CR watershed, RF-BCSD at days 19-28 and BCSD at days 5-8; at the BP watershed, RF-BCSD at days 23-26 and BCSD at days 7-10).

Figure 5.6 presents the probabilistic statistics of the CRPS of the S2S ensemble precipitation at the study watersheds over the forecast horizon of 28 days. The smaller the CRPS values, the more skillful the S2S precipitation is. Similar to previous Figure 5.5, the blue-colored lines are the CRPS skill of the S2S precipitation resulting from RF-BCSD adaptation, and the black-colored lines are the CRPS skill of the S2S precipitation resulting from BCSD.

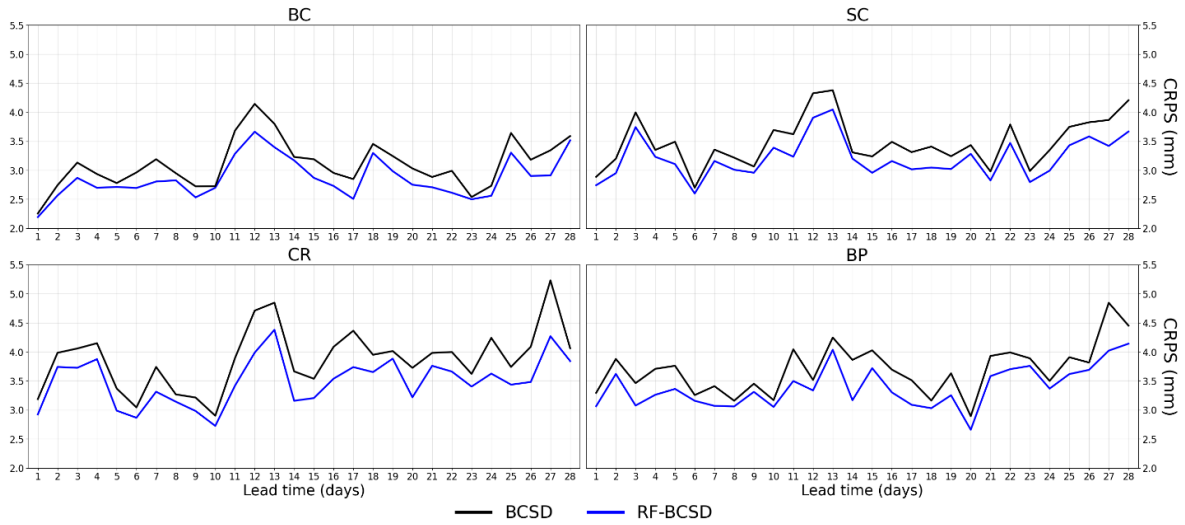

Figure 5.6 CRPS skill of the mean areal S2S precipitation after BCSD and RF-BCSD adaptation 4 study watersheds and over the 28-day forecast horizon.

The CRPS results presented in Figure 5.6 are consistent with previous KGE results in general. The values of CRPS of the S2S ensemble precipitation show similar increasing trends over the forecast horizon across different study watersheds, which indicate continuously decreasing predictive skills with longer forecast lead times. Comparing CRPS values resulting from RF-BCSD and BCSD, it can be observed that RF-BCSD brings consistently smaller CRPS values (higher skill) across different study watersheds. Similar to the KGE result from previous Figure 5.5, RF-BCSD-associated CRPS values at longer lead times are sometimes equivalent to the BCSD-associated CRPS values at much shorter lead times (e.g., at the BC watershed, RF-BCSD at days 26-27 and BCSD at days 8-9; at the SC watershed, RF-BCSD at days 25-27 and BCSD at days 4-5; at the CR watershed, RF-BCSD at days 27-28 and BCSD at days 2-3; at the BP watershed, RF-BCSD at days 21-23 and BCSD at days 4-5).

5.4.3 Predictive Performance of Streamflow

To quantify how available S2S precipitation affects the overall quality of streamflow predictions while excluding other factors (e.g., the imperfect structure of the hydrologic model, uncertainty arising from parameters of the hydrologic model, etc.), the streamflow predictions

resulting from S2S precipitation are evaluated against the proxy of the streamflow measurements (i.e., simulated streamflow from Sac-SMA using the calibrated parameter set and precipitation measurements from PRISM). Similar to the previous evaluation of the S2S precipitation, a deterministic evaluation metric KGE, and a probabilistic metric of CRPS are employed.

Figure 5.7 presents the deterministic KGE values of the ensemble means of the streamflow hindcasts resulting from RF-BCSD and BCSD-adapted S2S precipitation over the forecast horizon of 28 days and across 4 study watersheds. In Figure 5.7, the blue-colored lines are the KGE of streamflow hindcasts associated with the RF-BCSD adaptation and the black-colored lines are the KGE of streamflow hindcasts associated with the BCSD adaptation.

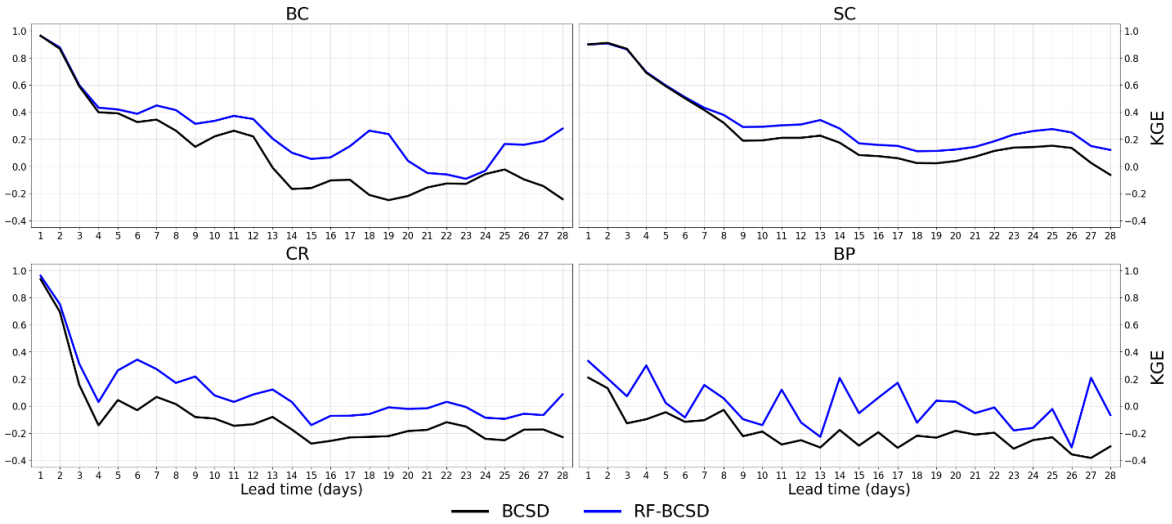

Figure 5.7 KGE skill of the streamflow hindcasts resulted from S2S precipitation after BCSD and RF-BCSD adaptation over the forecast horizon at the 4 study watersheds.

The KGE of streamflow hindcasts presented in Figure 5.7 shows a similar decreasing trend over the forecast horizon at all study watersheds. Comparing KGE at different study watersheds, similar patterns are observed. The KGE values are drastically higher (close to 1) at forecast lead times within 2-3 days than that at longer forecast lead times. However, at the BP

watershed, such a pattern appears to be less significant compared to other than at other watersheds. Comparing the KGE of streamflow associated with BCSD and RF-BCSD, RF-BCSD leads to overall higher KGE values across all study watersheds. However, unlike previous precipitation evaluation results, the advantages of KGE brought by RF-BCSD appear to be marginal at forecast lead times within ~3 to ~7 days at all watersheds except at the BP. At the BP watershed, the advantages of KGE brought by RF-BCSD are rather consistent and show fewer differences at different forecast lead times.

Figure 5.8 presents the probabilistic CRPS values of the entire ensemble of the streamflow hindcasts resulting from BCSD and RF-BCSD adapted S2S precipitation over the forecast horizon of 28 days and across 4 study watersheds. In Figure 5.8, the blue-colored lines are the CRPS skill of streamflow hindcasts associated with the RF-BCSD adaptation, and the black-colored lines are the CRPS skill of streamflow hindcasts associated with the BCSD adaptation.

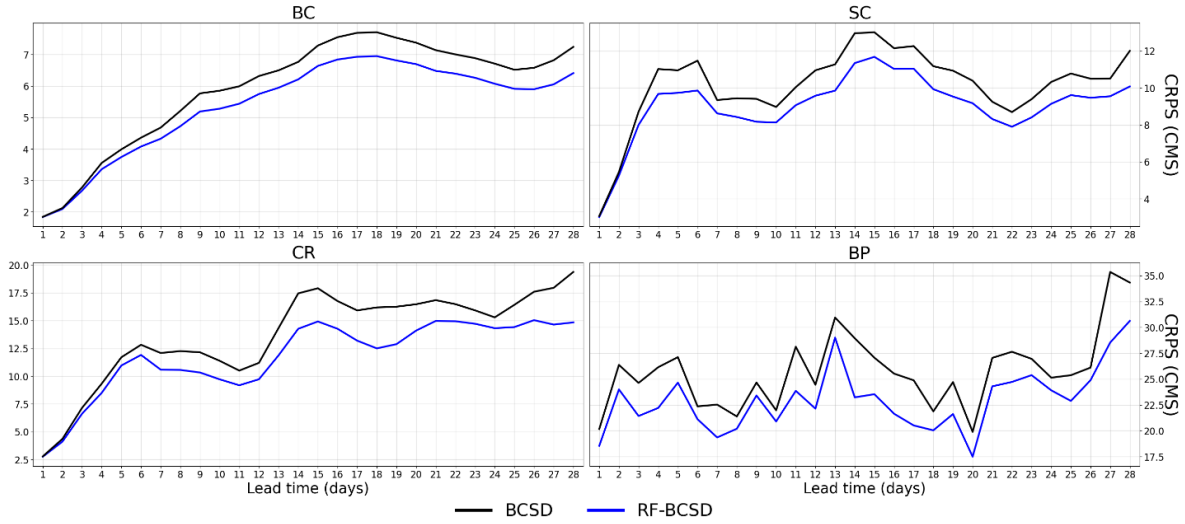

Figure 5.8. CRPS skill of the streamflow hindcasts resulted from S2S precipitation after BCSD and RF-BCSD adaptation over the forecast horizon at 4 study watersheds.

From Figure 5.8, it can be observed that the CRPS results are consistent with the KGE results presented in Figure 5.7. The CRPS at different study watersheds shows overall

increasing trends over forecast lead times, indicating the decrease of predictive skill in streamflow over forecast lead times across all study watersheds. The application of RF-BCSD brings overall smaller CRPS of the streamflow hindcasts in contrast to the case where BCSD was applied. Similar to the KGE results presented in Figure 5.7, the differences between the CRPS associated with RF-BCSD and BCSD appear to be marginal at very short forecast lead times within ~3 to ~5 days but become more apparent and significant as forecast lead time increases. The only exception is still the BP watershed where the improvement of CRPS brought by the application of RF seems more consistent at different forecast lead times.

5.4.4 The "Elasticity" of the Added Skill Between Streamflow and Precipitation

To further quantify the changes in the predictive skill of streamflow hindcasts corresponding to the changes in the predictive skill of precipitation, the "elasticities" of KGE at different forecast lead times are computed and presented in Table 5.3.

Table 5.3 KGE "elasticities" at different forecast lead times and at different study watersheds.

Study watersheds	Forecast lead times (days)						
	1 to 4	5 to 8	9 to 12	13 to 16	17 to 20	20 to 24	25 to 28
BC	0.08	0.93	2.50	2.63	2.02	1.74	1.20
SC	0.00	0.13	0.63	0.62	0.51	0.48	0.56
CR	0.32	2.24	1.03	0.60	2.01	0.85	0.83
BP	1.06	0.95	1.02	0.97	1.03	0.97	1.00

According to the "elasticities" shown in Table 5.3, the 4 study watersheds can be divided into 2 groups in general. The BC and SC watersheds are showing similar behavior. Where at the BC and SC watersheds, the "elasticities" are relatively small at very short lead times within 8 days but become drastically larger after ~8 days. On the other hand, the "elasticities" at the CR and BP watersheds show different behaviors. At the CR watershed, the elasticities fluctuate

across the entire forecast horizon without showing an obvious pattern. The elasticity at the BP watershed remains rather stable around the values of 1 across the entire forecast horizon.

5.5 Discussion

In this chapter, I employ two post-processing schemes to adapt the raw S2S precipitation at 4 local watersheds for streamflow predictions. The first post-processing scheme is a popular and standard BCSD, while another is a combination of the RF and BCSD. My results from **section 5.4.2** show that both BCSD and RF-BCSD remove the original forecast bias very effectively. Most importantly, according to the computed KGE and CRPS values, the application of RF has led to overall more skillful S2S precipitation forecasts, both deterministically and probabilistically.

Without many surprises, the added skill of the S2S precipitation brought by the application of RF has propagated into the streamflow predictions through the ESP framework. According to the computed KGE and CRPS values in **Section 4.3**, the RF-BCSD adapted S2S precipitation has led to consistently more skillful streamflow predictions over the entire forecast horizon of 28 days and across 4 different study watersheds.

To further quantify the propagation of the added skill of S2S precipitation to streamflow predictions, the "elasticities" of KGE at different forecast lead times are computed and presented in **Section 5.4.4**. I found that the streamflow forecast seems unaffected by the quality of precipitation at very short forecast lead times at the BC and SC watersheds. However, similar behavior is not observed in the CR and BP watersheds. To explain such differences observed in "elasticities" at the 4 study watersheds, I took a detailed look at the calibrated parameter sets of the Sac-SMA models at the 4 study watersheds. For BC and SC, the upper zone free water capacities (UZFWC) are 55mm and 63mm, which are much larger than upper zone tension

water capacities (UZTWC) of 2.42mm and 9.92mm. Further, the upper zone water depletion coefficient (UZK, which describes the linear relationship between the stored water in UZFWC and its contributing runoff) at BC and SC are relatively small with values of 0.17 and 0.25. In contrast, the UZFWC at CR and BP (42mm and 62mm) are much smaller than the UZTWC (140mm and 120mm). In addition, the UZK at CR and BP (0.58 and 0.68) are much larger than that at BC and SC.

Given these parameters, while considering the computation logic of the Sac-SMA, I believe the observed difference in "elasticities" is mainly due to the differences in the "hydrologic memories" at the 4 study watersheds. In Sac-SMA, the amount of input precipitation tries to fill UZTWC first before entering UZFWC. As a result, a larger UZTWC would result in less water in UZFWC. However, only the water stored in UZFWC contributes to streamflow. Further, smaller UZK values could make such lagged effects last even longer due to a slower draining speed. Therefore, at BC and SC, relatively more water in the UZFWC and relatively slower draining speed lead to relatively longer "hydrologic memories", and vice versa at the CR and BP watersheds. In other words, when forecast lead time is relatively short, the quality of streamflow prediction tends to be affected by the initial hydrologic conditions (IHCs) more significantly at BC and SC watersheds than that at CR and BP. The relatively shorter "hydrologic memories" of the CR and BP watersheds make the quality of streamflow prone to the quality of precipitation inputs, even at very short forecast lead times.

My interpretation of the different behaviors of "elasticiticity" values across study watersheds suggests that further advancements of streamflow predictions at S2S timescale perhaps require efforts in multi-aspects. The quality of streamflow predictions at S2S timescale is jointly determined by the estimation of IHCs, the available hydrologic model (along with

calibrated parameter set), and S2S precipitation forecasts altogether. My analysis of "elasticities" at different forecast lead times indicates that at shorter forecast lead times, the quality of streamflow predictions could be significantly affected by the uncertainty arising from the estimation of IHCs. Therefore, more advanced data assimilation techniques should be developed to consider the land-surface inertia of watersheds more accurately and comprehensively.

However, at longer forecast lead times exceeding ~10 days, the predictive skill of streamflow is more likely to be dominated by the quality of precipitation forecasts. Efforts have been made to advance weather/climate predictions at S2S timescale by identifying additional predictability sources, better assimilating measured atmospheric variables, and improving modeling tools (Domeisen et al., 2022; Mayer and Barnes, 2022; White et al., 2022). But due to uncertainty arising from various sources, it is still extremely challenging to provide accurate and reliable weather/climate forecasts at S2S timescale to this date (AghaKouchak et al., 2022; Krishnamurthy, 2019).

It is therefore important for hydrologists, forecasters, or end-users, to make the best of available S2S precipitation forecasts through various post-processing techniques. In this chapter, I demonstrate the effectiveness of a relatively simple and popular ML-based approach (i.e., RF) at a typical watershed scale by combining it with a more conventional distribution-based forecast adaption technique (i.e., BCSD) for streamflow forecasts at S2S timescale. Due to a limited training sample size subject to the availability of S2S precipitation hindcasts, I did not train RF for direct volumetric information. Instead, in this chapter, the RF is trained to give categorical predictions. The corresponding volumetrical information is further corrected by the BCSD. The effectiveness of such a framework suggests that more advanced data-driven

techniques could also be effective for similar tasks as long as enough training samples are provided. Most recently, Pan et al. (2021) have applied a deep learning technique, termed generative adversarial network (GAN), in correcting the bias of GCM-simulated precipitation over the entire CONUS. The successful application of GAN by Pan et al. (2021) shed some light on potential future works at typical watershed scales, as GAN is known for its prone to over-fitting with limited training data (Creswell et al., 2018; Wang et al., 2017). Therefore, I encourage future studies to further explore the effectiveness of more advanced data-driven approaches in adapting S2S precipitation at a typical watershed scale for more accurate and reliable S2S streamflow forecasting.

In summary, I believe that to advance streamflow forecast at S2S timescales, future studies could: (1) develop novel or more advanced data assimilation techniques to better consider the "hydrologic memory" of watersheds; or (2) improve weather/climate predictions from its core for more accurate and reliable S2S hydrometeorological forecasts; or (3) apply other data-driven approaches to get the best use of available S2S forecast products in hydrology through nowadays popular deep learning techniques of various kind.

5.6 Conclusion

In this chapter, I investigate the effectiveness of an ML technique, termed RF, in adapting the raw S2S precipitation forecasts from NASA GEOS5 to 4 local watersheds in the NCEI South climate region. To address the commonly presented forecast bias and improve the predictive skill of raw S2S forecast during the forecast adaptation, the RF is jointly employed with BCSD (i.e., RF-BCSD) in contrast to the benchmark scenario where only BCSD is employed. The adapted S2S precipitation under both schemes is further applied for streamflow hindcast experiments under the ESP framework. The predictive skills of S2S precipitation as

well as streamflow forecasts, are quantified with KGE and CRPS measurements at different forecast lead times and across 4 study watersheds. In addition, the propagation of the added skill from precipitation to streamflow is quantified with my pre-defined "elasticity". My obtained results highlight future applications of other data-driven ML or Deep Learning techniques of various kinds at typical watershed scales to get the best use of available S2S forecast products for more accurate streamflow forecasts. My major research conclusions are listed as follows:

1. The joint application of RF and BCSD can effectively remove the forecast bias in the raw S2S precipitation forecasts. But more importantly, the application of RF can further improve the predictive skill of the raw S2S precipitation both deterministically and probabilistically, as suggested by KGE and CRPS results.
2. The added skill of S2S precipitation associated with the usage of RF could propagate into the streamflow predictions through the ESP framework. The S2S precipitation adapted by RF-BCSD has led to overall more skillful streamflow predictions at all forecast lead times over the forecast horizon and across the 4 study watersheds, in contrast to the case where only BCSD is employed.
3. My analysis of the "elasticity" of the added skill between precipitation and streamflow suggests that the "memory" of the hydrologic system could play an important role in terms of the accuracy of streamflow forecasts, especially at relatively shorter forecast lead times.
4. Although future advancements in S2S streamflow forecasting probably require efforts in multi-aspects, my study in this chapter highlights the potential success of data-driven ML or Deep Learning techniques to further improve the quality of S2S precipitation

forecasts at a typical watershed scale for overall superior streamflow forecasting at S2S timescale.

Acknowledgments

The material is based upon work supported by the National Science Foundation (NSF) under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3, and the Oklahoma EPSCoR Research Seed Grand supported by NSF. This work is also partially supported by U.S. Department of Defense, Army Corps of Engineers (DOD-COR) Engineering With Nature (EWN) Program (Award No. W912HZ-21-2-0038); This research was also supported by a grant from the Research Council of the University of Oklahoma Norman Campus.

References

AghaKouchak, A., Pan, B., Mazdiyasni, O., Sadegh, M., Jiwa, S., Zhang, W., Love, C., Madadgar, S., Papalexiou, S. and Davis, S. 2022. Status and prospects for drought forecasting: opportunities in artificial intelligence and hybrid physical–statistical forecasting. Philosophical Transactions of the Royal Society A 380(2238), 20210288.

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J. and Peng, Q. 2018. Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. Journal of Geophysical Research: Atmospheres 123(22), 12,543-512,563.

Borovikov, A., Cullather, R., Kovach, R., Marshak, J., Vernieres, G., Vikhliaev, Y., Zhao, B. and Li, Z. 2019. GEOS-5 seasonal forecast system. Climate Dynamics 53, 7335-7361.

Cao, Q., Shukla, S., DeFlorio, M.J., Ralph, F.M. and Lettenmaier, D.P. 2021. Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in coastal Western US watersheds. Journal of Hydrometeorology 22(6), 1535-1552.

Caruana, R., Lawrence, S. and Giles, C. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in neural information processing systems 13.

Chiew, F., Zhou, S. and McMahon, T. 2003. Use of seasonal streamflow forecasts in water resources management. Journal of Hydrology 270(1-2), 135-144.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, A.A. 2018. Generative adversarial networks: An overview. IEEE signal processing magazine 35(1), 53-65.

Day, G.N. 1985. Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management 111(2), 157-170.

de Andrade, F.M., Young, M.P., MacLeod, D., Hirons, L.C., Woolnough, S.J. and Black, E. 2021. Subseasonal precipitation prediction for Africa: Forecast evaluation and sources of predictability. Weather and Forecasting 36(1), 265-284.

Delaney, C.J., Hartman, R.K., Mendoza, J., Dettinger, M., Delle Monache, L., Jasperse, J., Ralph, F.M., Talbot, C., Brown, J. and Reynolds, D. 2020. Forecast informed reservoir operations using ensemble streamflow predictions for a multipurpose reservoir in Northern California. Water Resources Research 56(9), e2019WR026604.

Domeisen, D.I., White, C.J., Afargan-Gerstman, H., Muñoz, Á.G., Janiga, M.A., Vitart, F., Wulff, C.O., Antoine, S., Ardilouze, C. and Batté, L. 2022. Advances in the subseasonal prediction of extreme events: Relevant case studies across the globe. Bulletin of the American Meteorological Society 103(6), E1473-E1501.

Grillakis, M.G., Koutroulis, A.G. and Tsanis, I.K. 2013. Multisegment statistical bias correction of daily GCM precipitation output. Journal of Geophysical Research: Atmospheres 118(8), 3150-3162.

Gupta, H.V., Kling, H., Yilmaz, K.K. and Martinez, G.F. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology 377(1-2), 80-91.

Harrigan, S., Prudhomme, C., Parry, S., Smith, K. and Tanguy, M. 2018. Benchmarking ensemble streamflow prediction skill in the UK. Hydrology and Earth System Sciences 22(3), 2023-2039.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J. and Aila, T. 2020. Training generative adversarial networks with limited data. Advances in neural information processing systems 33, 12104-12114.

Kim, T., Yang, T., Zhang, L. and Hong, Y. 2022. Near real-time hurricane rainfall forecasting using convolutional neural network models with Integrated Multi-satellitE Retrievals for GPM (IMERG) product. Atmospheric Research 270, 106037.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P. and Becker, E. 2014. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bulletin of the American Meteorological Society 95(4), 585-601.

Krishnamurthy, V. 2019. Predictability of weather and climate. Earth and Space Science 6(7), 1043-1056.

Mayer, K.J. and Barnes, E.A. 2022. Quantifying the Effect of Climate Change on Midlatitude Subseasonal Prediction Skill Provided by the Tropics. Geophysical Research Letters 49(14), e2022GL098663.

Naeini, M.R., Yang, T., Sadegh, M., AghaKouchak, A., Hsu, K.-l., Sorooshian, S., Duan, Q. and Lei, X. 2018. Shuffled complex-self adaptive hybrid evolution (SC-SAHEL) optimization framework. Environmental Modelling & Software 104, 215-235.

Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., Viger, R., Blodgett, D., Brekke, L. and Arnold, J. 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences 19(1), 209-223.

Pan, B., Anderson, G.J., Goncalves, A., Lucas, D.D., Bonfils, C.J., Lee, J., Tian, Y. and Ma, H.Y. 2021. Learning to correct climate projection biases. Journal of Advances in Modeling Earth Systems 13(10), e2021MS002509.

Pegion, K., Kirtman, B.P., Becker, E., Collins, D.C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D. and Zhu, Y. 2019. The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. Bulletin of the American Meteorological Society 100(10), 2043-2060.

Richter, J.H., Glanville, A.A., Edwards, J., Kauffman, B., Davis, N.A., Jaye, A., Kim, H., Pedatella, N.M., Sun, L. and Berner, J. 2022. Subseasonal Earth system prediction with CESM2. Weather and Forecasting 37(6), 797-815.

Schaake, J. and Larson, L. 1998 Ensemble streamflow prediction (ESP): Progress and research needs, pp. J19-J24.

Shukla, S., Sheffield, J., Wood, E.F. and Lettenmaier, D.P. 2013. On the sources of global land surface hydrologic predictability. Hydrology and earth system sciences 17(7), 2781-2796.

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(3), 1477-1490.

Troin, M., Arsenault, R., Wood, A.W., Brissette, F. and Martel, J.L. 2021 Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years, Wiley Online Library.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E. and Fuentes, M. 2017. The subseasonal to seasonal (S2S) prediction project database. Bulletin of the American Meteorological Society 98(1), 163-173.

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X. and Wang, F.-Y. 2017. Generative adversarial networks: introduction and outlook. IEEE/CAA Journal of Automatica Sinica 4(4), 588-598.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J. and Murray, V. 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorological applications 24(3), 315-325.

White, C.J., Domeisen, D.I., Acharya, N., Adefisan, E.A., Anderson, M.L., Aura, S., Balogun, A.A., Bertram, D., Bluhm, S. and Brayshaw, D.J. 2022. Advances in the application and utility of subseasonal-to-seasonal predictions. Bulletin of the American Meteorological Society 103(6), E1448-E1472.

Wood, A.W. and Lettenmaier, D.P. 2008. An ensemble approach for attribution of hydrologic prediction uncertainty. Geophysical Research Letters 35(14).

Wood, A.W., Leung, L.R., Sridhar, V. and Lettenmaier, D. 2004. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Climatic change 62(1), 189-216.

Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D. and Peng, Q. 2021. A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region. Journal of Hydrology 602, 126723.

Zhang, L., Kim, T., Yang, T., Hong, Y. and Zhu, Q. 2021. Evaluation of Subseasonal-to-Seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous US. Journal of Hydrology 603, 127058.

Zhang, L., Yang, T., Gao, S., Hong, Y., Fan, M., Lu, D., Xu, H. and Xiao, C. 2023a. An alternative Ensemble Streamflow Prediction (ESP) approach using improved Subseasonal-to-Seasonal (S2S) precipitation forecasts from the North America Multi-Model Ensemble Phase II (NMME-2) dataset. Journal of Hydrology (under review).

Zhang, L., Yang, T., Gao, S., Hong, Y., Zhang, Q., Wen, X. and Cheng, C. 2023b. Improving Subseasonal-to-Seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous US using machine learning models. Atmospheric Research 281, 106502.

Chapter 6. Conclusions and Recommendations for Future Work

6.1 Conclusions

In this dissertation, I have studied the applicability of available S2S precipitation from NMME-2 in hydrology for streamflow forecasting. That includes (1) a comprehensive evaluation of the available S2S precipitation forecasts from the NMME-2 over CONUS; (2) initial application of the S2S precipitation through BCSD forecast adaptation technique at 4 study watersheds in the NCEI South climate region; (3) a novel forecast adaptation technique that combines Machine Learning (i.e., RF) with BCSD for better predictions of the extreme precipitation events over CONUS; and (4) validation of the joint application of RF and BCSD from (3) at the same 4 watersheds that have been studied in (2) for streamflow hindcast experiments under the ESP framework. The Sac-SMA is chosen to be the experimental hydrologic model in this dissertation.

The first research question I posted in chapter I is: "***(1) What are the levels of accuracy of the S2S ensemble precipitation forecasts from NMME-2 at different forecast lead times, within different seasons, and over different spatial domains across CONUS?***". To address this question, **Chapter 2** comprehensively evaluated the S2S precipitation from NMME-2 over CONUS. The Percentage Bias (PBIAS) and Anomaly Correlation Coefficient (ACC) of the S2S precipitation from NMME-2 are computed against the reference dataset PRISM over the entire CONUS. Regional analysis of the resulting PBIAS and ACC is also conducted based on the climate region portioning per the National Centers for Environmental Information (NCEI).

Chapter 2 concludes that substantial forecast biases are commonly associated with the S2S precipitation from NMME-2 over CONUS. It also concludes that the forecast skill of the S2S precipitation from NMME-2 decreases rapidly from week 1 to week 2 and remains at a

marginal level at week 3 and week 4 across CONUS. The PBIAS is most sensitive to seasonality, while ACC is most sensitive to forecast lead times. The simple model averaging (SMA) of the NMME-2 members shows higher deterministic forecast skill than any single NMME-2 model under most comparison scenarios. Spatially, all NMME-2 models show higher forecast skills in the western coastal regions of CONUS, and the lowest predictive skill is observed in the central regions of CONUS.

Given the compromised performance of the S2S precipitation from NMME-2, it is natural that one would wonder if available S2S precipitation will still lead to superior performance in hydrology for streamflow forecasting. This echoes the second research question I posted in **Chapter 1**: "***In contrast to the resampled precipitation forecasts, how informative are the bias-corrected and downscaled S2S precipitation forecasts in streamflow forecasts under the ESP framework at typical watershed scales?***". To address this second research question, **Chapter 3** carried out a series of streamflow hindcast experiments using the S2S precipitation from NMME-2 at 4 watersheds in the NCEI South region, where the S2S precipitation from NMME-2 performs consistently worse than that in other climate regions. The underlying logic is that if S2S precipitation from NMME-2 leads to better S2S streamflow forecasts at study watersheds in the NCEI South climate region, then it is very likely that S2S precipitation would lead to superior S2S streamflow forecasts in other climate regions as well. In Chapter 3, the raw S2S precipitation from NMME-2 is adapted to the 4 study watersheds with the Bias Correction and Spatial Disaggregation (BCSD) approach. The performance of BCSD-adapted S2S precipitation in streamflow forecasting is compared with the baseline resampled precipitation under the Ensemble Streamflow Prediction (ESP) framework.

Chapter 3 concludes that the predictive skills of the S2S precipitation from NMME-2 and its associated streamflow predictions are generally higher than that of the resampled precipitation at the study watersheds. However, the advantage of the streamflow predictions associated with the S2S precipitation appears insignificant, especially at longer forecast lead times exceeding ~10 days. Nevertheless, the specific evaluation of high-volume streamflow values suggests that S2S precipitation leads to better categorical predictions of extreme events, which highlights the potential benefits of S2S precipitation in flood forecasting in the future. Finally, visual inspections of a few selected high-volume streamflow events at the study watersheds confirmed the advantages of S2S precipitation in streamflow forecasting in a more direct manner. From the visual inspections, I found that the streamflow hindcasts generated with S2S precipitation present either closer volumetric predictions or a smaller temporal offset of the measured peak streamflow volumes.

My results from **Chapter 3** confirmed that the S2S precipitation from NMME-2 does lead to overall better streamflow predictions under the ESP framework. However, given the marginal advantage brought by the application of S2S precipitation in streamflow forecasting at longer forecast lead times, a novel forecast adaptation approach is still needed to further improve the predictive skill of the raw S2S precipitation from a post-processing perspective. Therefore, in **Chapter 4**, a novel post-processing approach is developed to address the third research question I posted in Chapter I: "***Can Machine Learning (ML) algorithms provide a more reliable post-processing method with respect to improving the quality of precipitation forecasts over different geographical domains and lead times?***". In **Chapter 4**, a popular Machine Learning (ML) technique termed the Random Forest Classifier (RF) is trained to improve the capability of the raw S2S precipitation NASA Goddard Earth Observation System

Model version 5 (i.e., GEOS5, a contributing model to the NMME-2 data archive) in predicting the occurrence of weekly extreme precipitation events above 99% over the entire CONUS. Four S2S forecast variables are used, including precipitation, surface air temperature, and geopotential height at 500 hPa and 850 hPa, during the training of RF.

Chapter 4 further develops a proof-of-concept forecast adaptation framework using the RF to post-process the raw S2S forecasts over CONUS. The results show that the application of RF can significantly improve S2S forecasts in terms of predicting the occurrence of weekly extreme precipitation events at week 1. However, if additional S2S forecast variables are included as the RF inputs, the forecast performance is further improved at longer lead times (i.e., weeks 2-4). Additional sensitivity analysis on the hyperparameters of RF indicates that tunning of the RF would result in overall better forecasts at all lead times over CONUS.

With the novel RF-based forecast adaptation framework tested over CONUS on weekly events, I further examined the applicability of the RF at typical watershed scale for daily S2S streamflow predictions in **Chapter 5** and intends to address the last and fourth research question that I posted in **Chapter 1**: "***How much of the added predictive skill obtained from the ML-improved S2S precipitation forecasts propagate into the streamflow predictions under the ESP framework at typical watershed scales?***".

In **Chapter 5**, RF (developed in **Chapter 4**) is jointly applied with BCSD to improve the predictive skill of the raw S2S precipitation from GEOS5 at the same 4 study watersheds in **Chapter 2** for daily S2S streamflow predictions under the ESP framework. The performance of the joint application of RF and BCSD (RF-BCSD) in S2S precipitation forecast adaptation are compared with the baseline scenario where only BCSD is applied. The propagation of the added predictive skills from precipitation to streamflow is quantified with the predefined "elasticity"

of KGE (i.e., unit change of KGE in streamflow prediction per unit change of KGE in S2S precipitation forecasts).

Chapter 5 concludes that the application of RF-BCSD can not only adapt the raw S2S precipitation forecasts to typical watersheds but also improve the predictive skill of S2S precipitation, in contrast to the case where only BCSD is applied. But more importantly, the added skill brought by the application of RF-BCSD can propagate into streamflow predictions through the ESP framework across different study watersheds. My analysis of the "elasticity" of the added skill between precipitation and streamflow suggests that the "memory" of the hydrologic system could play an important role in terms of the accuracy of streamflow forecasts, especially at relatively shorter forecast lead times.

6.2 Recommendations for Future Work

This dissertation demonstrates the capability of available S2S precipitation in hydrology for streamflow forecasting. According to the results presented in previous sections, available S2S precipitation should bring overall more accurate and reliable streamflow forecasts, both deterministically and probabilistically. However, further advancements in streamflow forecasting at S2S timescale require joint efforts in not only meteorology/climate forecasting but also in hydrology, as the quality of streamflow forecasting is bounded by both.

Based on the results from this dissertation, I reckon that the accuracy of available S2S hydrometeorological forecast products might be the current limiting factor for more skillful streamflow forecasts at S2S timescale. For this dissertation, the 4 selected study watersheds are dry with their streamflow dominated by precipitation. Therefore, I believe continuously improving the quality of S2S precipitation forecasts should still be the focus of future works.

However, achieving accurate and reliable S2S precipitation has long been recognized as extremely challenging. To improve the quality of S2S precipitation forecasts, advancements need to be made to identify additional predictability sources of precipitation at S2S timescales (de Andrade et al., 2021; Huang et al., 2021; Moron and Robertson, 2020), and/or to enhance current forecasting GCM models (Robertson et al., 2015; Xue et al., 2021).

An alternative for hydrologists or other forecast end users to improve available S2S precipitation could be various post-processing techniques. For instance, this dissertation (**Chapters 4** and **5**) developed a proof-of-concept ML-based forecast adaptation framework and successfully improved the predictive skill of the raw S2S forecasts. The successful application of this ML-based framework highlights the potential success of other data-driven ML or Deep Learning (DL) techniques in S2S forecast adaptations. Due to a rather limited training sample size, the RF is trained to give categorical predictions in this dissertation, leaving the corresponding volumetrical information to be further corrected by the BCSD. Most recently, Pan et al. (2021) have applied a deep learning technique, termed generative adviser network (GAN), in correcting the bias of GCM-simulated precipitation over CONUS. The successful application of GAN by Pan et al. (2021) shed some light on potential future works at typical watershed scales, as GAN is known for its resistance to over-fitting with limited training data (Creswell et al., 2018; Wang et al., 2017). Therefore, I encourage future studies to further explore the effectiveness of more advanced data-driven approaches in adapting S2S precipitation at a typical watershed scale for more accurate and reliable S2S streamflow forecasting.

On the other hand, deepening our understanding of hydrology could also benefit streamflow forecasting at S2S timescale. In this chapter, I used a rather standard lumped hydrologic model

(i.e., Sac-SMA) that has been long recognized with an advanced parameter estimation algorithm. However, the performance of the Sac-SMA is far from perfect even after model calibration, indicating that the hydrologic characteristics of the study watersheds need be better represented with novel development in hydrology in the future. Gao et al. (2021) suggested through the application of distributed hydrologic models, the common existing non-perennial streamflow in the NCEI South climate region might be better considered. Similar studies of such could lead to better hydrologic simulations, which eventually will result in more advanced S2S streamflow forecasting.

The initial hydrologic conditions (i.e., IHCs, initial soil, and/or snow conditions) also play important roles in streamflow forecasting. For study watersheds in this dissertation, the IHCs tend to contribute to the predictive skill of streamflow at very short forecast lead times within ~5 days. But for watersheds with different characteristics, the effect coming from IHCs might have longer effects on streamflow at an extended forecast lead time. For instance, it is known that snow ablations play a pivotal role and accounts for around 70% of the streamflow at watersheds located in the Upper Colorado region (Li et al., 2017; Xiao and Lettenmaier, 2021). Therefore, future work should also extend and transfer the current hydrologic forecasting experiments to watersheds in other regions over CONUS for a more comprehensive evaluation.

Further, in this dissertation, the IHCs are considered under the ESP framework through a technical process called the “spin-up”. In which, a chosen hydrologic model is forced with measured hydrometeorological time series up to the “current” time step for the best estimation of the IHCs of a watershed. Although such a process has proven to be effective and has been in operational use for decades, novel alternative techniques and measurements have become available nowadays for an alternative. Due to the rapid advancements in remote sensing, many

informative large-scale datasets have become available, including the Soil Moisture Active Passive (SMAP) mission (Entekhabi et al., 2010), groundwater condition estimations from the twin Gravity Recovery and Climate Experiment (GRACE) mission (Frappart and Ramillien, 2018; Tapley et al., 2004), and/or surface water condition estimation through the recently launched Surface Water and Ocean Topography (SWOT) mission (Morrow et al., 2019) among some others. For streamflow forecasting at S2S timescale, an alternative to the "spin-up" in considering the IHCs could be merging these available large-scale hydrologic remote sensing datasets through advanced data assimilation techniques.

In summary, future works could explore from either a meteorological/climate forecasting or a hydrologic perspective. The quality of available S2S forecast products essentially determines the accuracy of streamflow predictions at S2S timescale, especially at longer forecast lead times. Advancements in weather/climate predictions at S2S timescale could be made by enhancing dynamical forecast models, identifying additional predictability sources, and/or through post-processing techniques of various kinds. However, future advancements in hydrology could also contribute to the topic, as proper considerations of the IHCs and hydrologic models are essentially the backbone of hydrologic forecasting at S2S timescale. Further deepening our understanding of hydrology for more accurate hydrologic simulations at watersheds with different characteristics, and/or integrating available hydrologic remote sensing datasets when performing hydrologic forecasts could lead our way to further success in S2S hydrologic forecasting.

References

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, A.A. 2018. Generative adversarial networks: An overview. IEEE signal processing magazine 35(1), 53-65.

Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J. and Johnson, J. 2010. The soil moisture active passive (SMAP) mission. Proceedings of the IEEE 98(5), 704-716.

Frappart, F. and Ramillien, G. 2018. Monitoring groundwater storage changes using the Gravity Recovery and Climate Experiment (GRACE) satellite mission: A review. Remote Sensing 10(6), 829.

Gao, S., Chen, M., Li, Z., Cook, S., Allen, D., Neeson, T., Yang, T., Yami, T. and Hong, Y. 2021. Mapping dynamic non-perennial stream networks using high-resolution distributed hydrologic simulation: A case study in the upper blue river basin. Journal of Hydrology 600, 126522.

Li, D., Wrzesien, M.L., Durand, M., Adam, J. and Lettenmaier, D.P. 2017. How much runoff originates as snow in the western United States, and how will that change in the future? Geophysical Research Letters 44(12), 6163-6172.

Morrow, R., Fu, L.-L., Ardhuin, F., Benkiran, M., Chapron, B., Cosme, E., d'Ovidio, F., Farrar, J.T., Gille, S.T. and Lapeyre, G. 2019. Global observations of fine-scale ocean surface topography with the Surface Water and Ocean Topography (SWOT) mission. Frontiers in Marine Science 6, 232.

Pan, B., Anderson, G.J., Goncalves, A., Lucas, D.D., Bonfils, C.J., Lee, J., Tian, Y. and Ma, H.Y. 2021. Learning to correct climate projection biases. Journal of Advances in Modeling Earth Systems 13(10), e2021MS002509.

Robertson, A.W., Kumar, A., Peña, M. and Vitart, F. 2015. Improving and promoting subseasonal to seasonal prediction. Bulletin of the American Meteorological Society 96(3), ES49-ES53.

Tapley, B.D., Bettadpur, S., Watkins, M. and Reigber, C. 2004. The gravity recovery and climate experiment: Mission overview and early results. Geophysical research letters 31(9).

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X. and Wang, F.-Y. 2017. Generative adversarial networks: introduction and outlook. IEEE/CAA Journal of Automatica Sinica 4(4), 588-598.

Xiao, M. and Lettenmaier, D.P. 2021. Atmospheric rivers and snow accumulation in the Upper Colorado River basin. Geophysical Research Letters 48(16), e2021GL094265.

Xue, Y., Yao, T., Boone, A.A., Diallo, I., Liu, Y., Zeng, X., Lau, W.K., Sugimoto, S., Tang, Q. and Pan, X. 2021. Impact of initialized land surface temperature and snowpack on subseasonal to seasonal prediction project, phase i (ls4p-i): organization and experimental design. Geoscientific Model Development 14(7), 4465-4494.

Appendix A

The Sac-SMA model is a classical lumped hydrologic model first introduced by Burnash and Ferral (1973). Figure A.1 presents the conceptual structure of the Sacramento Soil Moisture Accounting (Sac-SMA) model. The Sac-SMA incorporates a two-story soil layer to account for both fast and slow runoff. Both upper and lower soil layers are further divided into primary and supplemental water buckets. The input precipitation will enter the first layer of soil and be partitioned into supplemental and primary water buckets for fast and slow flows. Based on the available water storage of tension and free water in the upper zone, input precipitation will be partitioned into surface runoff and infiltration into the lower soil zone. The amount of precipitation exceeding the upper zone tension water capacity (UZTWM) will then try to fulfill the upper zone free water capacity (UZFWM). The overall remaining amount of excess precipitation from the upper soil zone will become surface runoff. The amount of available water in the upper soil zone slowly contributes to the lower zone tension water capacity (LZTWZ), the lower zone free water capacity (LZFWC), and the lower zone supplementary water capacity (LZSWC) through percolations. Percolations between the upper and lower zones are described with two parameters: the maximum rate of percolation (ZPERC) and an exponent value (REXP). The total amount of percolated water is further split with a fractional value (PFREE) into different lower-zone water buckets. All three free water buckets from the upper and lower zones have different depletion coefficients (UZK, LZSK, LZPK), which describe the linear relationships between the amount of runoff and the corresponding water storage.

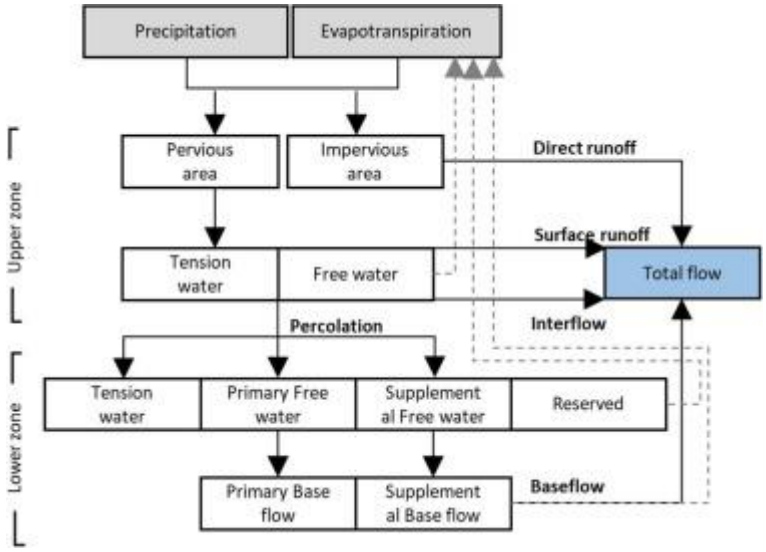

Figure A.1 Conceptual structure of the Sac-SMA model modified from Kim et al. (2021).

The Sac-SMA model has a total of 16 parameters. Among which, 13 of them are tunable (Brazil, 1989). The 13 tunable parameters of the Sac-SMA are listed in Table A.1.

Table A.1 13 tunable parameters of the Sac-SMA.

No.	Parameter	Description	Range
1	UZTWM	Upper zone tension water maximum storage (mm)	[10, 300]
2	UZFWM	Upper zone free water maximum storage (mm)	[5, 150]
3	UZK	Upper zone free water lateral drainage rate (day^{-1})	[0.1, 0.75]
4	PCTIM	Impervious fraction of the watershed area (decimal fraction)	[0, 0.1]
5	ADIMP	Additional impervious area (decimal fraction)	[0, 0.2]
6	ZPERC	Maximum percolation rate (dimensionless)	[5, 350]
7	REXP	Exponent of the percolation equation (dimensionless)	[1, 5]
8	LZTWM	Lower zone tension water maximum storage (mm)	[10, 500]
9	LZFSM	Lower zone supplemental free water maximum storage (mm)	[5, 400]
10	LZFPM	Lower zone primary free water maximum storage (mm)	[10, 1000]
11	LZSK	Lower zone supplemental free water lateral drainage rate (day^{-1})	[0.01, 0.35]
12	LZPK	Lower zone primary free water lateral drainage rate (day^{-1})	[0.001, 0.05]
13	PFREE	Fraction of water percolating from upper zone directly to lower zone free water (decimal fraction)	[0.0, 0.9]

References

Burnash, R.J. and Ferral, R.L. (1973) A generalized streamflow simulation system: Conceptual modeling for digital computers, US Department of Commerce, National Weather Service, and State of California ….

Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., ... & Hong, Y. (2021). Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study of four watersheds with different hydro-climatic regions across the CONUS. Journal of Hydrology, 598, 126423.

Appendix B

Figure B.1 to B.9 below presents Receiver Operating Curves (ROC) under experiment scenario 1 (E1), experiment scenario 6 to 10 (E6-E10) from **Chapter 4**, at 9 NCEI climate regions and at 4 different lead times (week 1-4). Different colored lines indicate ROC curves resulting from different experiment scenarios. The blue-dotted diagonal line represents a classifier constantly making random guesses.

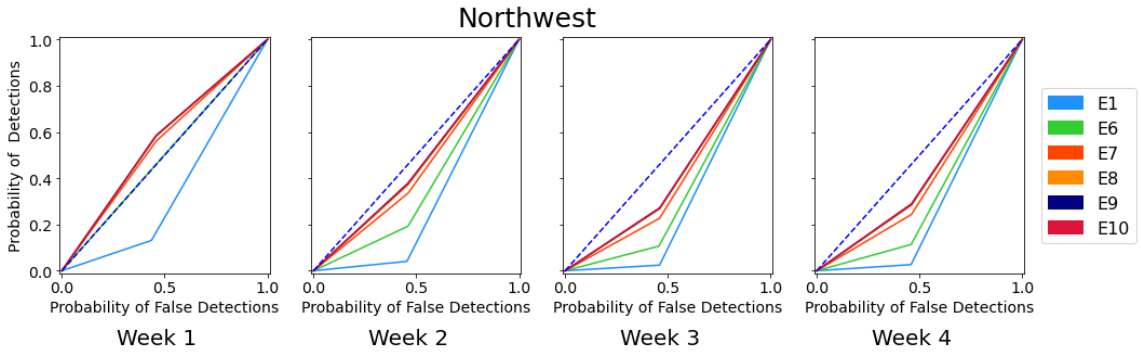

Figure B.1 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Northwest NCEI climate region, at different lead times (weeks 1 to 4).

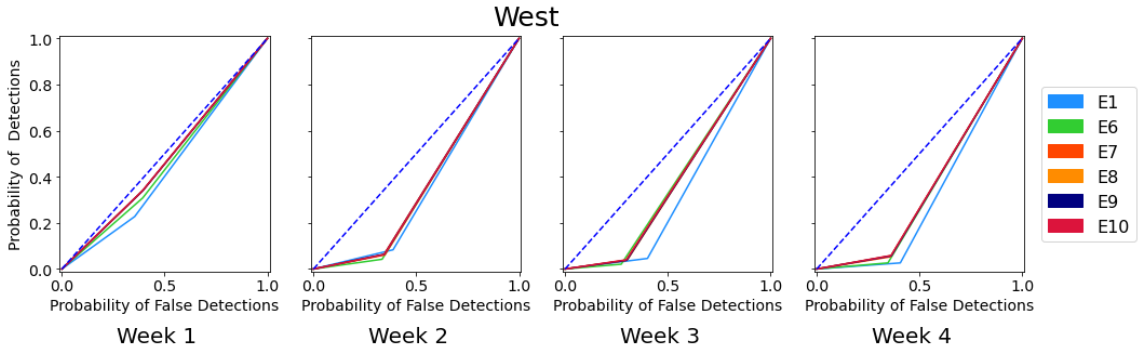

Figure B.2 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the West NCEI climate region, at different lead times (weeks 1 to 4).

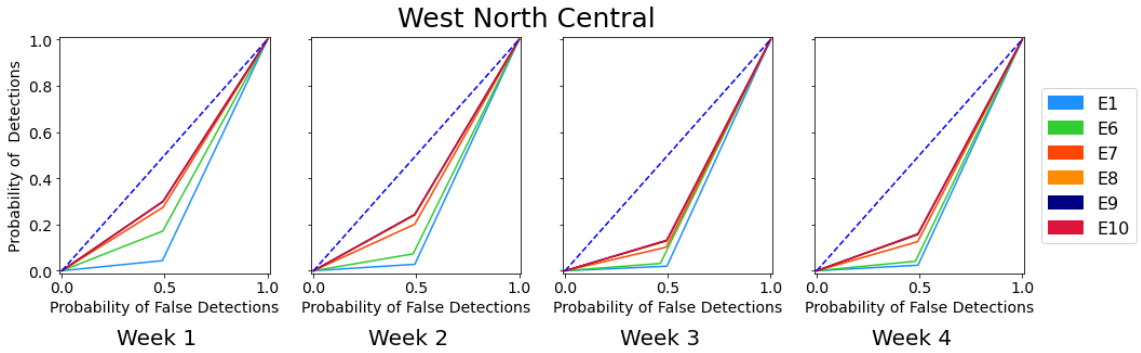

Figure B.3 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the West North Central NCEI climate region, at different lead times (weeks 1 to 4).

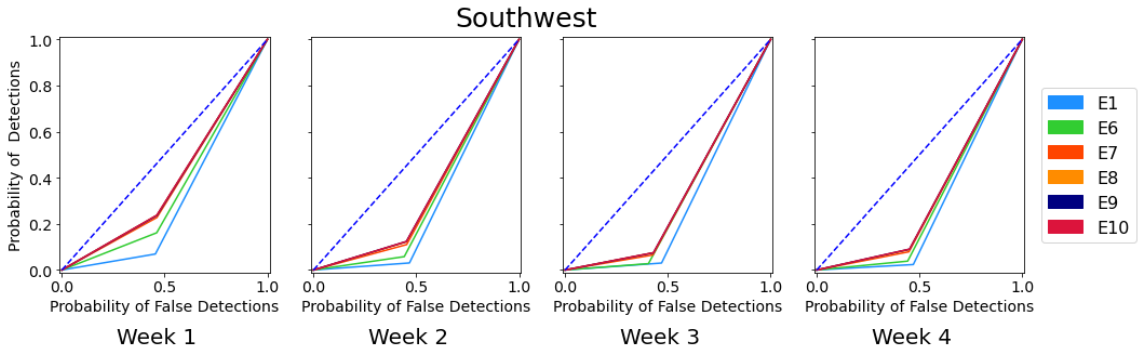

Figure B.4 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Southwest NCEI climate regions, at different lead times (weeks 1 to 4).

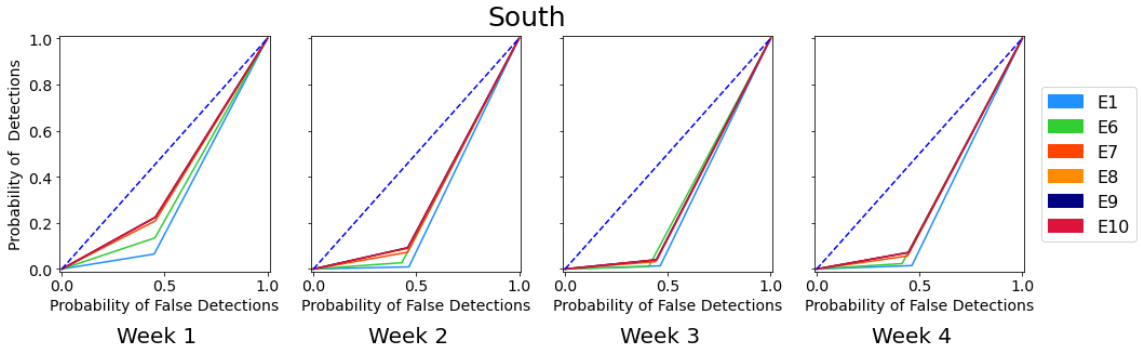

Figure B.5 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the South NCEI climate region, at different lead times (weeks 1 to 4).

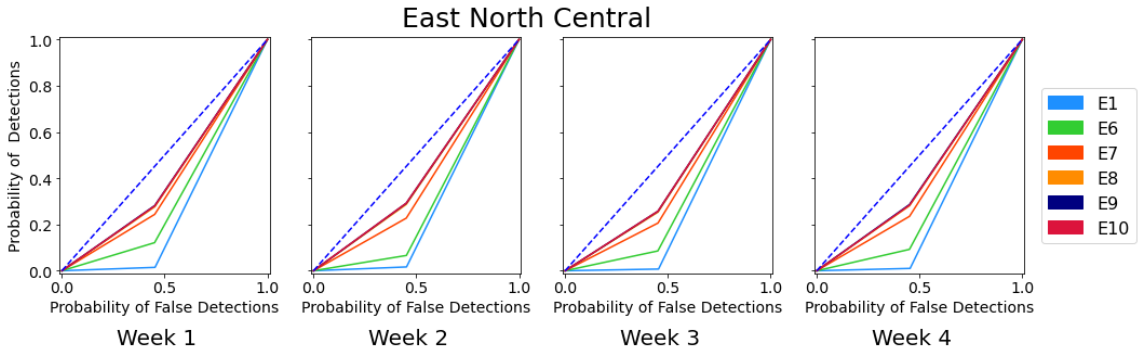

Figure B.6 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the East North Central NCEI climate region, at different lead times (weeks 1 to 4).

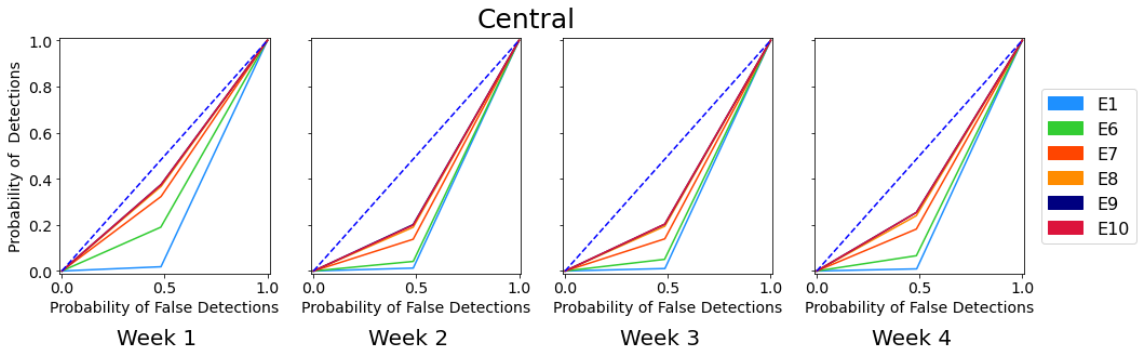

Figure B.7 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Central NCEI climate region, at different lead times (weeks 1 to 4).

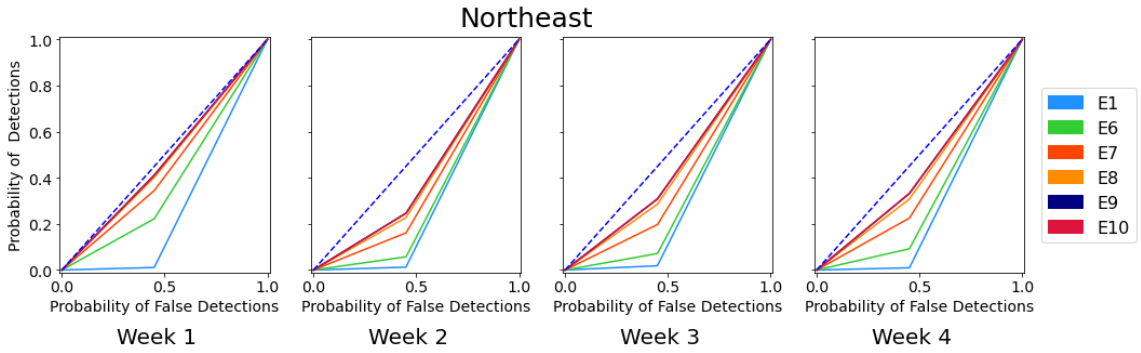

Figure B.8 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Northeast Central NCEI climate region, at different lead times (weeks 1 to 4).

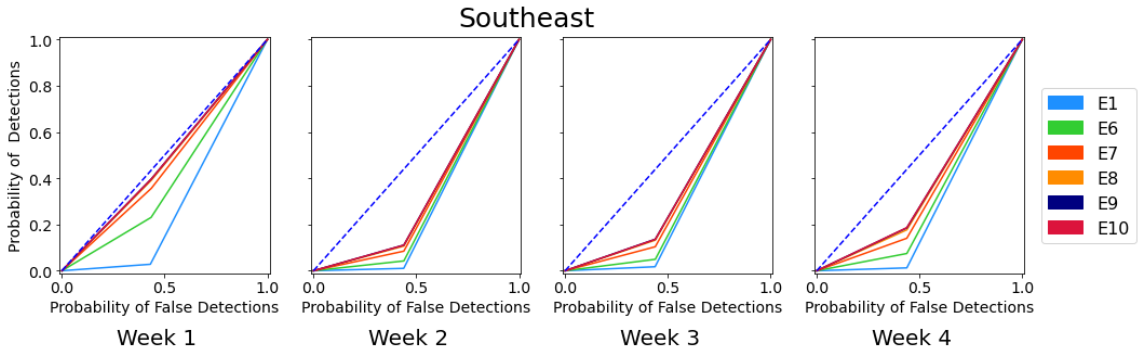

Figure B.9 ROC curves from experiment scenarios 1, and 6 to 10 from Chapter 4, in predicting extreme precipitation events at the Southeast NCEI climate region, at different lead times (weeks 1 to 4).