IMPROVING POST PROCESSING OF ENSEMBLE STREAMFLOW FORECAST FOR SHORT-TO-LONG RANGES: A MULTISCALE APPROACH

by

BABAK ALIZADEH

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at
The University of Texas at Arlington
August, 2019

Arlington, Texas

Supervising Committee:
Dr. Dong-Jun Seo, Supervising Professor
Dr. Nick Fang
Dr. Yu Zhang
Dr. Haksu Lee

Abstract

Improving Post Processing of Ensemble Streamflow Forecast for Short-to-long Ranges: A Multiscale Approach

Babak Alizadeh, Ph.D.
The University of Texas at Arlington, 2019
Supervising Professor: Dong-Jun Seo

A novel multi-scale post-processor for ensemble streamflow prediction, MS-EnsPost, and a multiscale probability matching (MS-PM) technique for bias correction in streamflow simulation are developed and evaluated. The MS-PM successively applies probability matching (PM) across multiple time scales of aggregation to reduce scale-dependent biases in streamflow simulation. For evaluation of MS-PM, 34 basins in four National Weather Service (NWS) River Forecast Centers (RFC) in the US were used. The results indicate that MS-PM improves over PM for streamflow prediction at a daily time step, and that averaging the empirical cumulative distribution functions to reduce sampling uncertainty marginally improves performance. The performance of MS-PM, however, quickly reaches a limit with the addition of larger temporal scales of aggregation due to the increasingly large sampling uncertainties. MS-EnsPost represents a departure from the PM-based approaches to avoid large sampling uncertainties associated with distribution modeling, and to utilize fully the predictive skill in model-simulated and observed streamflow that may be present over a range of temporal scales.

MS-EnsPost uses data-driven correction of magnitude-dependent bias in simulated flow, multiscale regression over a range of temporal aggregation scales, and ensemble generation using parsimonious error modeling. For evaluation of MS-EnsPost, 139 basins in eight RFCs

were used. Streamflow predictability in different hydroclimatological regions is assessed and characterized, and gains by MS-EnsPost over the existing streamflow ensemble post processor in the NWS Hydrologic Ensemble Forecast Service, EnsPost, are attributed. The ensemble mean prediction results show that MS-EnsPost reduces the root mean square error of Day-1 to -7 predictions of mean daily flow from EnsPost by 5 to 68 percent, and for most basins, the improvement is due to both bias correction and multiscale regression. The ensemble prediction results show that MS-EnsPost reduces the mean Continuous Ranked Probability Score of Day-1 to -7 predictions of mean daily flow from EnsPost by 2 to 62 percent, and that the improvement is due mostly to improved resolution than reliability.

Examination of the mean Continuous Ranked Probability Skill Scores (CRPSS) indicates that, for most basins, the improvement by MS-EnsPost is due to both magnitude-dependent bias correction and full utilization of hydrologic memory through multiscale regression. Comparison of the mean CRPSS results with hydroclimatic indices indicates that the skill of ensemble streamflow prediction with post processing is modulated largely by the fraction of precipitation as snow and, for non-snow-driven basins, mean annual precipitation.

The positive impact of MS-EnsPost is particularly significant for a number of basins impacted by flow regulations. Examination of the multiscale regression weights indicates that the multiscale regression procedure is able to capture and reflect the scale-dependent impact of flow regulations on predictive skills of observed and model-predicted flow. One of the motivations for MS-EnsPost is to reduce data requirement so that nonstationarity may be considered. Comparative evaluation of MS-EnsPost with EnsPost indicates that, under reduced data availability, MS-EnsPost generally outperforms EnsPost for those basins exhibiting significant changes in flow regime.

Acknowledgements

At the early steps of my Ph.D. research, it became obvious to me that a researcher cannot and does not work without the help and support from his/her peers. While the list of individuals that I wish to thank extends beyond the limits of this format, I would like to acknowledge the dedication and support of the following persons:

First, I would like to express my deep sense of gratitude to my advisor Dr. Dong-Jun Seo, who has the attitude and the substance of a true professional researcher. Without his guidance and continuous support, I would have not been even close to where I am now. I would also like to thank the members of my Ph.D. committee, Dr. Nick Fang, Dr. Yu Zhang and of course Dr. Haksu Lee who never gave up supporting my research and answering my numerous questions.

I warmly thank my current and former colleagues at the Hydrology and Water Resources Laboratory who helped me with research and coursework as well as difficulties in personal life, Dr. Sunghee Kim, Dr. Seongjin Noh, Dr. Behzad Nazari, Dr. Hamideh Habibi, Dr. Reza Ahmad Limon, Dr. Amir Norouzi, Mr. Mohammad Nabatian, Mr. Behzad Rouhanizadeh, Mr. Ali Jozaghi, Mr. Vaghef Ghazvinian, Mrs. Soona Habibi and Mr. Miah Mohammad Saifuddin.

The material presented in this document is based upon work supported in part by the NOAA Climate Program Office under Grant NA15OAR4310109, the NWS COMET Program under UCAR Subaward No. SUBAWD000020, and the NSF under Grant CyberSEES-1442735. These supports are gratefully acknowledged. I would like to thank John Lhotak of CBRFC, Brett Whitin and Ark Henkel of CNRFC, Seann Reed of MARFC, Lisa Holts of MBRFC, Andrea Holz of NCRFC, Erick Boehmler of NERFC, Brad Gillies of NWRFC, Andrew Philpott, Frank Bell and others at WGRFC for providing the data used and help during the course of this work.

Dedication

I would like to dedicate this dissertation wholeheartedly to the my beloved ones:

My mother, for her unconditional love, patience, sacrifices and faith in me. She encouraged me to spread the wings and fly when everything in her wanted to keep me safe in her nest.

My grandmother, who nurtured and protected me with the greatest care one can imagine. I will never forget all she has done for me since I was a toddler.

My closest relatives, Manoochehr, Amin and Fereydoon, for encouraging me to go on every adventure, especially this one.

My sisters, Tahmineh and Taraneh, whom I know will become much more successful persons than their older brother.

The memories of my father, grandfather and aunt who inspired me to start, continue and complete this journey.

And Almighty God for always showing me that, wherever I go, I am not alone.

Thank you all for helping to give me the life I love today.

Table of Contents

List of Illustrations

Chapter 1

Introduction

Accurate short- to long-range streamflow forecast is critical to effective water management. Due to multiple sources of uncertainty, however, streamflow forecast is subject to large errors. For quantifying and communicating uncertainty, ensemble forecasting has been fast gaining acceptance (Cloke and Pappenberger 2009). For risk-based decision making in water management, reliable and skillful ensemble streamflow forecasts are a prerequisite (Demargne et al. 2014). The major sources of uncertainty in hydrologic forecasting include forecasts of precipitation and temperature at weather and climate scales, hydrologic, hydraulic and reservoir modeling, unmodeled or unknown human control of movement and storage of water, and anthropogenic changes to hydrdoclimatology (McMillan et al. 2011; Marimo et al. 2015; Subbey et al. 2004; Groves et al. 2008) (see Fig 1).

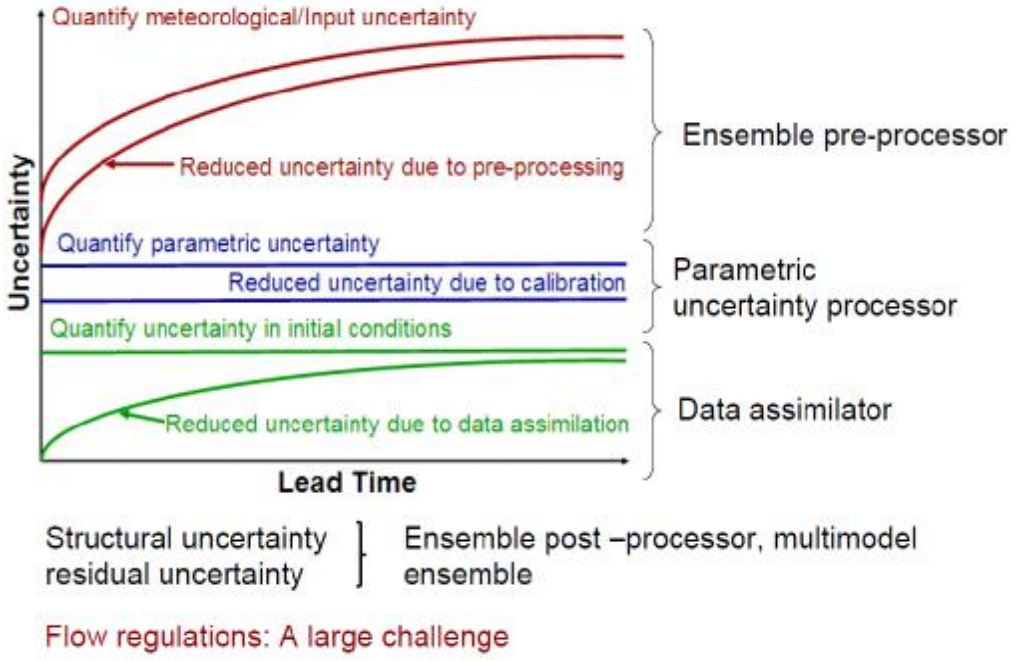


Figure 1: Major sources of hydrologic uncertainty.

This research focuses on advancing characterization, modeling, and reduction of hydrologic uncertainties in ensemble streamflow forecasting. The specific aim is to develop a post-processing methodology that:

- Reduces hydrologic uncertainty and improves streamflow prediction by fully utilizing skill in simulated and observed flow over a range of temporal scales of aggregation,
- Handles intermittency of flow in ephemeral streams in arid and semi-arid regions, and
- Reduces data requirement to allow nonstationarities arising from changing hydroclimatology.

Streamflow simulations from hydrologic models contain errors propagated from uncertain forcings, model initial conditions (IC), parameters and structures, and human control of storage and movement of water (Ajami et al. 2007; Doherty and Welter 2010; Gupta et al. 2012; Krzysztofowicz 1999; Montanari and Brath 2004; NRC 2006; Renard et al. 2010; Schaake et al. 2007; Seo et al. 2006; Wood and Schaake 2008). For risk-based management of water resources and water-related hazards, it is necessary to quantify the uncertainties from these sources (Borgomeo et al. 2014; Butts et al. 2004; Georgakakos et al. 2004; Hall and Borgomeo 2013; Hall et al. 2019). Ensemble forecasting has emerged in recent years as the methodology of choice for modeling and communicating forecast uncertainty (Cloke and Pappenberger 2009; Demargne et al. 2014; Demeritt et al. 2010; NRC 2006; Schaake et al. 2007). In the US, the National Weather Service (NWS) has recently implemented the Hydrologic Ensemble Forecast Service (HEFS; Demargne et al. 2014) at all River Forecast Centers (RFC) (Lee et al. 2018) following experimental operation at selected RFCs (Hartman et al. 2015; Kim et al. 2018; Wells 2017). To reduce and quantify hydrologic uncertainty in streamflow prediction, the HEFS employs the ensemble post-processor, EnsPost (NWS 2015; Seo et al. 2006) (see Fig 2).

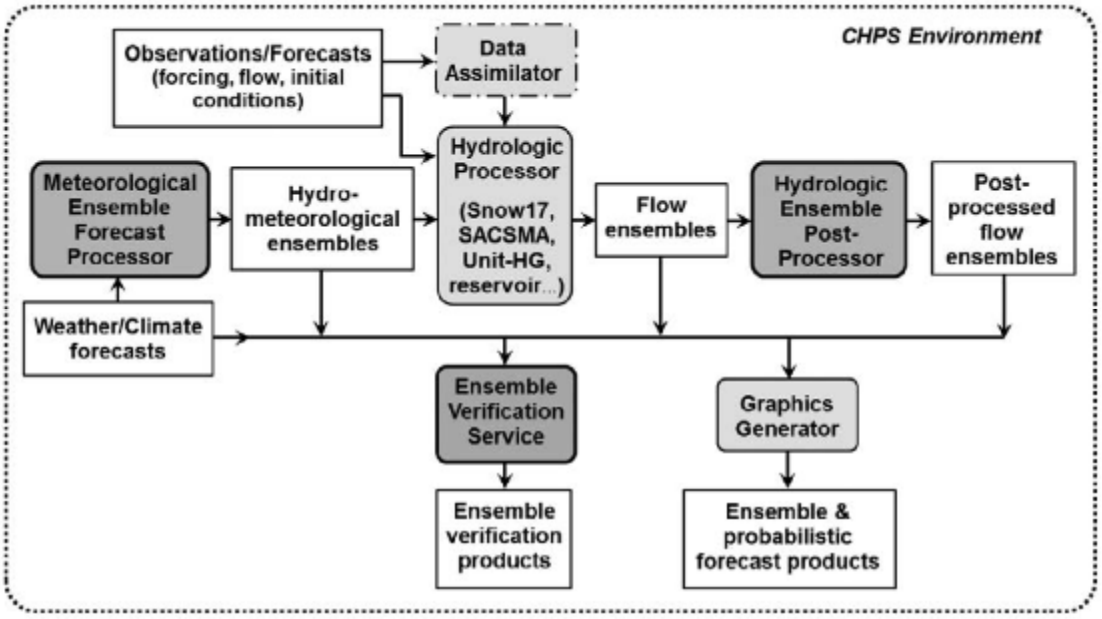


Figure 2: Schematic of HEFS's elements (Demargne et al. 2014).

Originally developed for short-range forecasting of natural flows in headwater basins, EnsPost models predictive hydrologic uncertainty using a combination of probability matching (PM; Hashino et al. 2002; Madadgar et al. 2014) and autoregressive (AR)-1 model with an exogenous variable, or ARX (1,1) (Bennett et al. 2014; Damon and Guillas 2002), in bivariate normal space (Krzysztofowicz 1999; Seo et al. 2006). EnsPost applies PM and ARX(1,1) at a daily scale only. In reality, however, the characteristic time scales of error in model-simulated flow may span a range of scales depending on the residence time of the hydrologic processes involved and the error characteristics of the forcings and the hydrologic models used (Blöschl and Sivapalan 1995). In addition, if the flow is strongly regulated, the errors may be reducible only over a certain range of temporal scales of aggregation due to the altered residence time and storage-outflow relationships.

This research develops and evaluates a multiscale probability matching (MS-PM) technique for improved bias correction in lieu of PM used in EnsPost, and a multiscale ensemble post processing methodology, MS-EnsPost, to improve skill in streamflow ensemble forecast from

short to long ranges. As part of the evaluation, the following research questions are also addressed.

- How do errors in operational model-simulated streamflow vary according to the time scale of aggregation?
- How do biases in model-simulated streamflow vary according to the magnitude of the simulated flow?
- How does the predictive skill in observed and model-predicted streamflow vary according to the time scale of aggregation?
- What is the relative importance between correcting biases and reducing uncertainty in the ICs among different basins in different hydroclimatological regions?
- How does the prediction skill of MS-EnsPost vary among different RFCs, and among different basins within an RFC? How does the skill compare with that of EnsPost?
- How does the above skill relate to hydroclimatology of the basin?
- How do flow regulations impact the above?
- How do the data availability and nonstationarity impact the above?

The fundamental contributions of this work are:

- Advances in understanding, statistical modeling, and assessment of errors and predictive skill in operational model-simulated flow at different time scales of aggregation and in different hydroclimatological regimes,
- Development and evaluation of a statistical post-processor that combines flow magnitude-dependent bias correction, multiscale regression utilizing hydrologic memory over a range of time scales, and parsimonious parametric modeling of the error,

- Advances in understanding of errors in model-simulated regulated flow, and in improving predictive skill via the multiscale approach, and
- Advances in understanding and assessment of data requirements for post-processing of ensemble streamflow forecast under nonstationarity.

This dissertation is organized as follows. Chapter 2 describes literature review for this research. Chapter 3 describes the developed methods in this work. In Chapter 4, the study basins and data used are described. Chapter 5 provides the evaluation measures. In Chapter 6, results for the developed methods are discussed. And finally, in Chapter 7, the conclusions of this work are decribed and future research recommendations are provided.

Chapter 2

Literature review

The positive impact of post-processing raw model simulations of streamflow in ensemble streamflow forecasting has been widely reported (Kim et al. 2018; Kim et al. 2016; Madadgar et al. 2014). It has also been shown recently that EnsPost significantly increases skill in ensemble forecasts of outflow from a water supply reservoir in North Texas during significant releases, in addition to that in ensemble inflow forecasts (Limon 2019). With increasing acceptance and adoption of ensemble streamflow forecasting by the operational community, developing more effective post-processing methods has been a very active area of research. To that end, a number of comparison studies have been carried out. For post-processing of meteorological forecast, Wilks (2006) compared direct model output (Wilks 2006), rank histogram recalibration (Hamill and Colucci 1998), single-integration Model Output Statistics (MOS; Erickson 1996), ensemble dressing (Roulston and Smith 2003), logistic regression (Hamill et al. 2004), non-homogeneous Gaussian regression (Gneiting et al. 2005), forecast assimilation (Stephenson et al. 2005), and Bayesian model averaging (Raftery et al. 2005). He concluded that logistic regression (Duan et al. 2007; Hamill et al. 2004), ensemble MOS (Gneiting et al. 2005), and ensemble dressing outperform the others. Boucher et al. (2015) compared the regression and dressing methods using synthetic data (Li et al. 2017). They concluded that the techniques have similar overall performance, and that the regression and dressing methods perform better in terms of resolution and reliability, respectively. Mendoza et al. (2016) used medium-range ensemble streamflow forecasts from the System for Hydrometeorological Applications, Research and Prediction, and compared quantile mapping (Mendoza et al. 2016; Hashino et al. 2006; Piani et al. 2010; Regonda and Seo 2008; Wood and Schaake 2008; Zhu and Luo 2015), logistic regression,

quantile regression (Bjørnar Bremnes 2004; Bogner et al. 2016; Coccia and Todini 2011; Koenker and Bassett 1978) and the general linear model post-processor (Zhao et al. 2011). They found that no single method performed best in all situations, and that the post processors' performance depended on factors such as soil type and land use, and hydroclimatic conditions of the basin.

Since the launch of the HEFS (NWS 2007, Demargne et al. 2014), the NWS has implemented the application at a number of RFCs (Fresch 2015). The experience thus far with EnsPost (Seo et al. 2006) indicates the following.

- Whereas the EnsPost ensembles are generally skillful for largely natural flows at short ranges (3 days or less), they provide little skill at longer ranges or for regulated flows.
- EnsPost requires long periods of record due to heavily parameterized stochastic modeling. Its performance is hence susceptible to data availability and nonstationarities of hydroclimatolological and other origins.

EnsPost is a statistical model of streamflow simulation error. It inputs raw ensemble streamflow forecast and corrects biases in the mean and higher-order moments, and outputs bias-corrected ensemble streamflow forecast. Fig 3 schematically depicts what EnsPost does. Successful correction would render the post-processed ensemble forecast reliable, i.e., unbiased in the probabilistic sense, and improve the skill of the ensemble streamflow forecast.

The resulting post-processed ensemble forecast reflects both the input uncertainty, i.e., uncertainty in the forcings, and the hydrologic uncertainty, i.e., uncertainty in converting forcings to streamflow using hydrologic, hydraulic and reservoir models. Generally speaking, the input uncertainty is larger than the hydrologic uncertainty. Because of the limits of predictability in weather and climate forecasting, reducing input uncertainty is a large challenge.

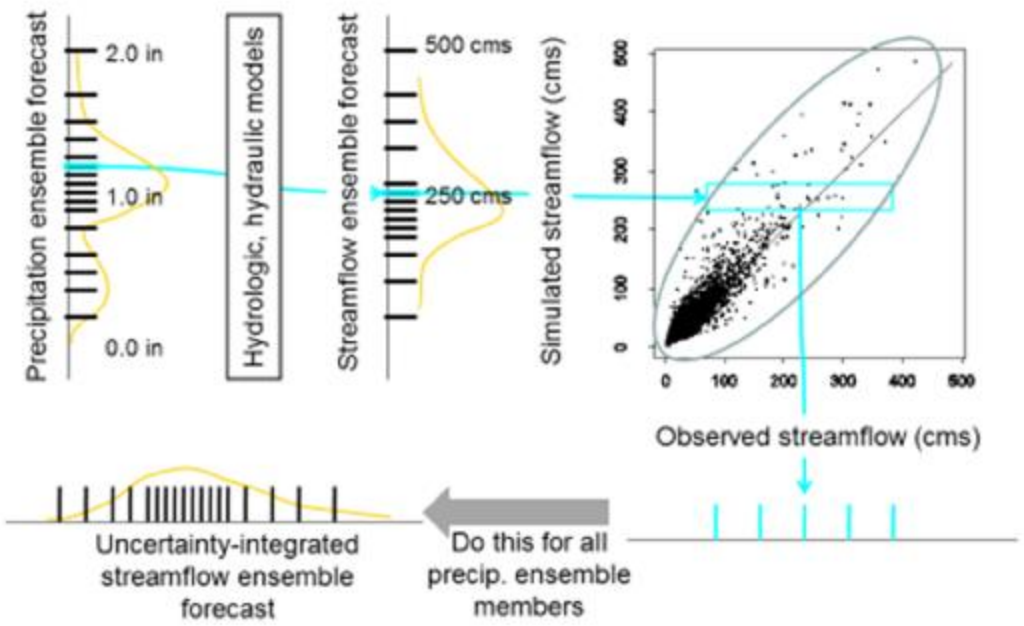


Figure 3: Illustration of integration of input and hydrologic uncertainties in hydrologic ensemble forecasting (NWS 2015).

Hydrologic uncertainty, on the other hand, may be reduced comparatively easily via statistical means if the data required are available.

PM, or cumulative distribution function (CDF) matching, was first introduced in ensemble streamflow post processing in NWS for Extended Streamflow Prediction (ESP, Day 1985, Perica et al. 1998, 1999) in support of water resources forecasting. Later, a regression model was combined with PM to form EnsPost in support of short-range ensemble forecasting (Seo et al. 2006). Though the parameter estimation procedure for EnsPost was subsequently modified to improve longer-range performance (but at the expense of compromising performance at very short ranges), the algorithm is limited by the single-scale nature of bias correction in that PM is performed only at a daily scale. EnsPost is limited also by the autoregressive (with an exogenous variable) nature of statistical modeling in that a single ARX(1,1) (Seo et al. 2006) model is applied recursively at a daily scale over the entire forecast horizon. In reality, biases exist in model-simulated flow over a range of scales due to various sources of error. For example, in

semi-arid regions where larger uncertainties exist in antecedent soil moisture, spatiotemporal variability of rainfall, soil moisture dynamics, and surface water dynamics, the dominant biases may have time scales of a rainfall event, or a few days. In such cases, bias correction at a multi-daily scale is likely to be more effective than that at a daily scale. Similarly, due to large uncertainties in flow regulations, model-simulated flow may have skill only at multi-daily, or even lager, time scales of aggregation. In such cases, one may not expect PM, which operates only at a daily scale, to be very effective. Furthermore, regulated flows are not, in general, autoregressive. The $ARX(\cdot,\cdot)$ class of models may hence be of limited effectiveness. For the above reasons, EnsPost is not likely to capture all available skill that may be present in the model-predicted flow and in the real-time streamflow observations over a wide range of time scales of aggregation.

To illustrate the importance of the time scale of bias correction used in PM, in Fig 4 the reduction in percent root mean square error (RMSE) of model-simulated mean daily flow (SQME) is shown. The reduction in RMSE over raw SQME by bias-corrected flow using PM at times scales of a single day and multiple days are shown in blue and green, respectively. The six basins used are located in the Upper Trinity River Basin in North Central Texas (Kim et al. 2018). The hydrologic models used are the Sacramento (SAC) and unit hydrograph (UH). Note that, for three of the basins, PM at daily scale increases RMSE over raw model simulations, whereas PM at multi-daily scale reduces RMSE for all six basins.

Because PM requires accurate modeling particularly of the upper tails of observed and simulated flows, EnsPost requires a long period of record. Due to urbanization and climate change, streamflow responses have changed or are changing significantly in many parts of the US. In addition, with the transition from gauge-only precipitation analysis (MAP) to multisensor

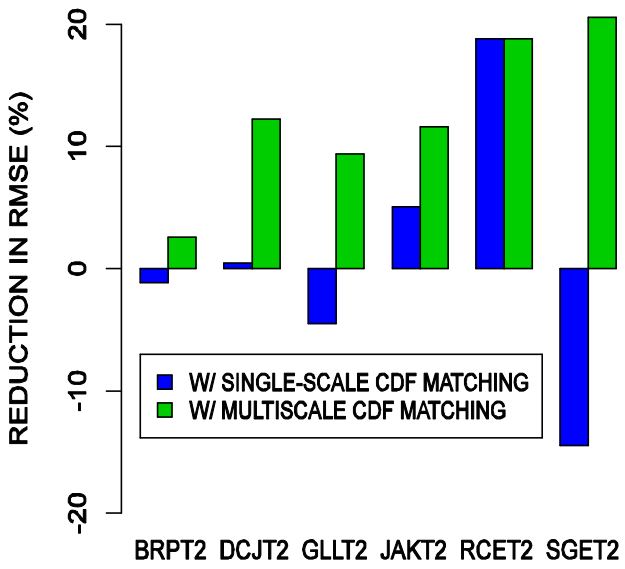


Figure 4: Reduction in percent RMSE of SQME as bias-corrected using PM at a daily time scale (in blue) and at a time scale of multiple days (in green) over the raw model-simulated QME for 6 basins in the Upper Trinity River Basin.

analysis (MAPX, Breidenbach 2001, 2002), there are significant changes in the statistical properties of model-simulated flow over different periods for many basins, even if there may be no changes in the actual hydroclimatology. Under changing conditions, the data requirement for EnsPost poses a rather difficult tradeoff between capturing possible nonstationarities vs. keeping sampling uncertainties smaller. To capture nonstationarity, one may divide the period of record or model trends. Such an operation, however, would significantly increase sampling uncertainties in statistical modeling. One may maximize sample size by keeping the entire period of record, but at the expense of introducing biases due to nonstationarities. To illustrate the above point, Fig 5 shows a 37-yr time series of the multiplicative bias for raw SQME to achieve unbiasedness against the observed for HUNP1 in the Juniata Basin of the MARFC's service area (Seo et al. 2006). A bias greater or less than unity indicates under- and oversimulation, respectively.

The figure indicates that the simulated flow tends to be biased high (i.e., oversimulate) in the first half of the time series whereas it is biased slightly low (i.e., undersimulate) in the second

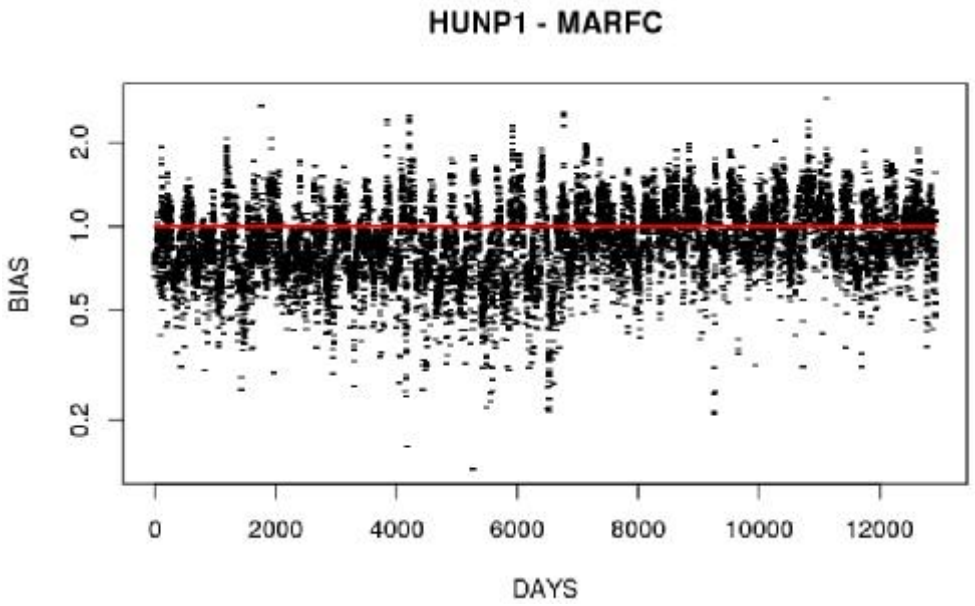


Figure 5: A 37-yr time series of the multiplicative bias for simulated QME to achieve unbiasedness against the observed for HUNP1 in the Juniata Basin, PA.

half. Because such nonstationarities in model bias distort distribution modeling, PM is not likely to be very effective without nonstationarity modeling, a tall order given PM's already large data requirements. In this work, MS-EnsPost is developed and evaluated which avoids data-intensive empirical variable transformation and thus minimizes distribution modeling. Owing to the parsimony, one may expect MS-EnsPost to reduce data requirements significantly while fully utilizing all available skill in model-simulated flow and real-time streamflow observations over a range of scales.

Chapter 3

Methods developed

This section describes MS-PM and MS-EnsPost. MS-PM was intended for EnsPost as an improved bias correction technique. MS-EnsPost is a new non-distributional approach for increased parsimony and reduced data requirement. MS-EnsPost is motivated by the findings from MS-PM that, whereas the multiscale approach improves bias correction, the gain is quickly lost as the temporal scale of aggregation increases by sampling uncertainty. MS-EnsPost hence departs from the distribution modeling-based approach of EnsPost and MS-PM in favor of parsimony and minimal variable transformation.

3.1 Multiscale probability matching (MS-PM)

MS-PM was first explored by NWS as a possible enhancement to EnsPost (Regonda and Seo 2008). In MS-PM, multiscale bias correction is applied to reduce biases in streamflow simulation. One may expect such correction to be effective when the characteristic time scales of the dominant model errors vary significantly from basin to basin or when there are multiple time scales at which the model errors may be dominant such as when the flow is regulated. MS-PM rests on the fact that reproducing marginal CDFs of the predictand at all scales encompassing the forecast horizon is a necessary condition for reproducing the multivariate CDF at the smallest scale over the entire forecast horizon under the stationarity assumption. The premise of MS-PM is that matching CDFs over a range of time scales of aggregation may render the resulting simulated flow at the smallest time scale statistically very similar in the multivariate sense to the observed flow at the smallest scale. If successful, the simulated flow at any time scale of aggregation would possess the same statistical properties of the observed flow at that scale,

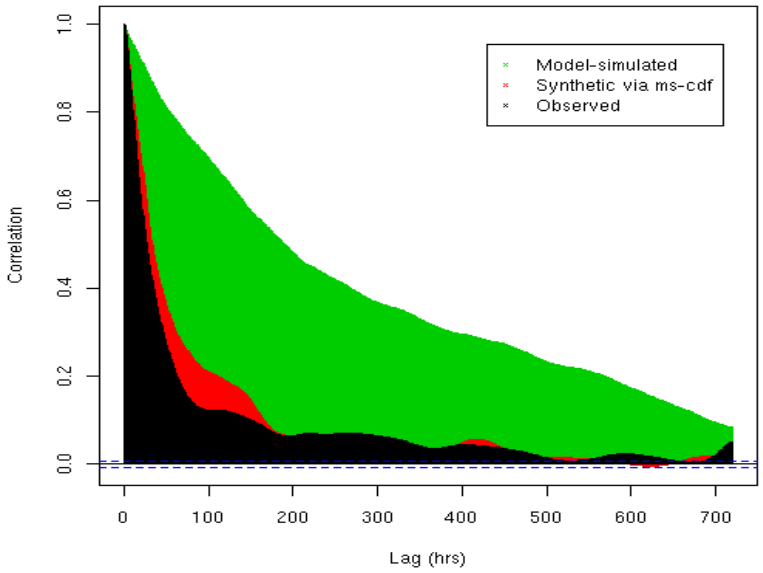


Figure 6: Serial correlation of simulated flow (green), observed flow (black) and multiscale post-processed flow (red).

thereby rendering streamflow ensembles reliable, i.e., probabilistically unbiased, regardless of the time scale of the user's interest. To illustrate, Fig 6 shows the very positive impact of MS-PM in which the serial correlation of model-simulated streamflow of rather poor quality (in green) is rendered very close (in red) to that of the observed streamflow (in black) via MS-PM.

The experience thus far with MS-PM is that, perhaps not surprisingly, it suffers from large sampling errors at larger time scales due to the increasingly smaller sample size. In addition, whereas single-scale PM at a multi-daily scale often corrects biases in daily flow more effectively than PM at a daily scale, MS-PM over a range temporal scales of aggregation does not necessarily improve over PM at a single scale, due presumably to the sampling uncertainties that accrue over multiple PM operations. On the other hand, larger aggregation scales may still be necessary, even with larger sampling uncertainties, for ephemeral basins in semi-arid regions. Accordingly, systematic assessment is necessary using a large number of basins in diverse hydroclimatology to ascertain the range of scales over which MS-PM is consistently effective

and to assess the associated data requirements. In this rersearch, the tradeoffs between the number of temporal scales of aggregation and the effectiveness of bias correction is assessed.

3.2 Multiscale ensemble post-processor (MS-EnsPost)

In this subsection, MS-EnsPost for ensemble streamflow prediction for short-to-long ranges is described. By short and long ranges, it is meant up to several days and at least 1 month ahead, respectively. MS-EnsPost is designed to reduce magnitude-dependent biases in raw model-simulated flow, and utilize all available skill that may exist over a range of temporal scales of aggregation in simulated and observed flows. MS-EnsPost consists of three elements: bias correction, multiscale regression, and ensemble generation. Fig 7 provides a schematic of the data flow and the associated processes.

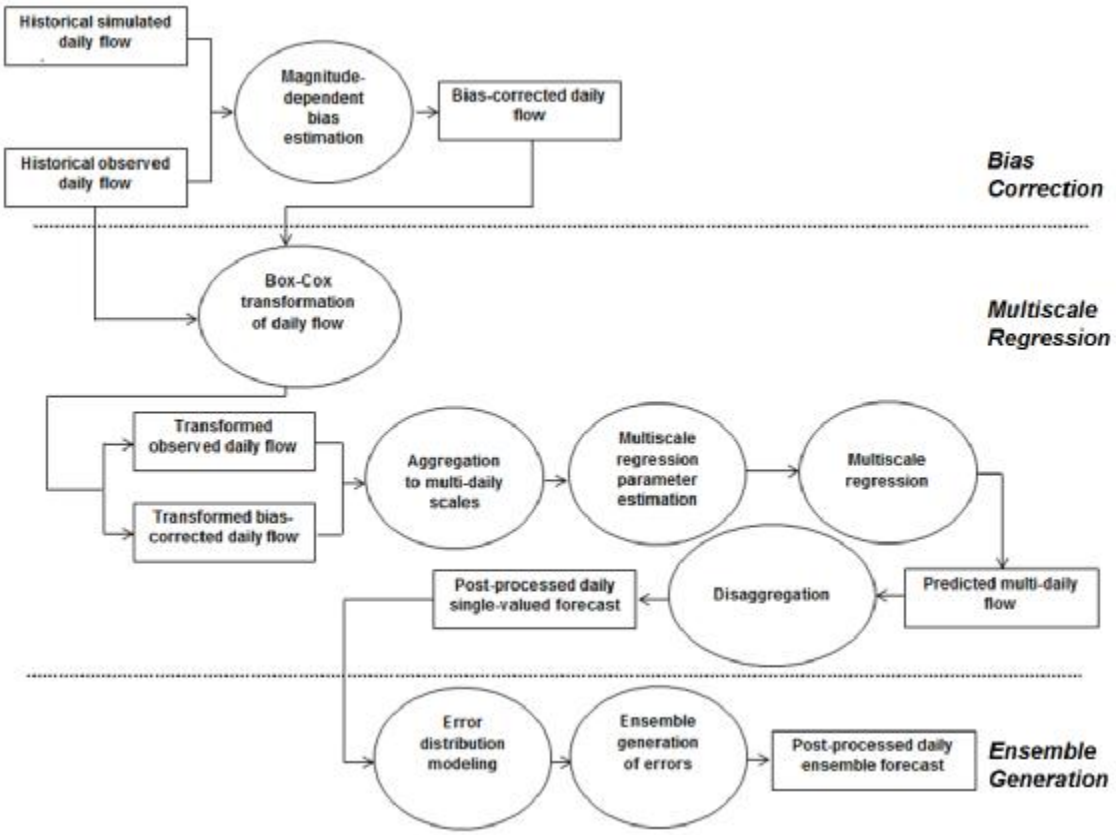


Figure 7: Schematic of MS-EnsPost elements and dataflow.

3.2.1 Magnitude-dependent bias correction

It is defined within some time period of interest the multiplicative bias, β_i, in the simulated flow valid at the i-th day, q_i^s, as:

$$\beta_i = \frac{q_i^o}{q_i^s} \tag{1}$$

where q_i^o denotes the observed flow valid at the i-th day. Model-simulated high and low flows generally have smaller and larger multiplicative biases, respectively (see Appendix A for examples). Hydrologic models tend to simulate the physical processes that govern high flows relatively more accurately (Dunne and Black 1970; Engman and Rogowski 1974; Freeze 1972; Horton 1933; Loague and VanderKwaak 2004). In addition, most hydrologic models, Continuous API being a prime example, are calibrated to perform better for high flows than for low flows (Fowler et al. 2018; Freer et al. 1996; Gan et al. 1997; Kim et al. 2007; Krause et al. 2005; Legates and McCabe Jr 1999; Nash and Sutcliffe 1970; Smith et al. 2014). The bias correction procedure in MS-EnsPost is designed to address this dependence. Sample estimates of β_i at a daily scale are very noisy due to very large variabilities in q_i^o and q_i^s. To obtain stable estimates of the magnitude-dependent bias, this procedure first pairs q_i^o and q_i^s, sorts them in the ascending order of q_i^s, and aggregates the resulting daily flows over different time scales. The temporal aggregation and the attendant noise cancellation greatly reduce the sampling uncertainty in the estimated bias, compared to that without aggregation. The time-aggregated flows are expressed as:

$$a_{k,(j)}^o = \sum_{i=(j_k-1)L_k+1}^{j_k L_k} q_{(i)}^o \tag{2}$$

$$a_{k,(j)}^s = \sum_{i=(j_k-1)L_k+1}^{j_k L_k} q_{(i)}^s \tag{3}$$

In the above, the symbol, (i), signifies that the variable subscripted is sorted in the ascending order of q_i^s, L_k denotes the k-th time scale of aggregation, and j_k denotes the j-th aggregation window of the k-th scale within the period of record, and $a_{k,(j)}^o$ and $a_{k,(j)}^s$ denote the sorted observed and simulated flows aggregated over the j-th time window of the k-th time scale, respectively, where the symbol, (j), signifies that the aggregation is based on the sorted daily flow. Eqs.(4) and (5) pool the simulated and observed flows such that, when averaged over the respective aggregation periods, the aggregated simulated flows are similar in magnitude to the conditioning flow, q_i^s, in Eq.(1). The bias for q_i^s at the k-the aggregation scale, $\beta_{k,i}$, is given by:

$$\beta_{k,i} = \frac{a_{k,(j)}^o}{a_{k,(j)}^s}, i \in [(j_k - 1)L_k + 1, j_k L_k] \quad (4)$$

In the above, the range for the index i, which is associated with the simulated flow to be bias-corrected, q_i^s, identifies the time scale of aggregation associated with the bias being estimated. In this work, we used $L_k = 2^k$ (days), $k = 1, \ldots, 14$, for the aggregation scales, but other choices are possible. In the above, the largest aggregation scale is almost 45 years long with which one would be applying a single multiplicative bias for all simulated daily flows regardless of their magnitude. Among the total of K different temporal scales of aggregation, the best-performing scale is identified via leave-one-year-out (or similar) cross validation using a period of record of N years as described below. First, the magnitude-dependent biases are estimated at the K different scales of aggregation using an (N-1)-year period of observed flow and matching model simulation. The resulting biases are applied to the simulated daily flow valid on each Julian day of the withheld year. The procedure then identifies the aggregation scale that produces the smallest RMSE in the bias-corrected simulated flow by comparing with the verifying observed flow. Once completed for all N years, the leave-one-year-out cross validation produces a total of N different sets of magnitude-dependent biases for simulated daily flow. For a given q_i^s, the

procedure arithmetically averages the N different biases associated with the respective time windows that enclose q_i^s in the sorted series. The procedure repeats the above steps for all possible values of q_i^s, from which the final single relationship between the simulated flow and the magnitude-dependent bias results.

3.2.2 Multiscale regression

Post-processing generally seeks predictions at the highest possible temporal resolution. High dimensional stochastic modeling necessary for such predictions, however, is a large challenge due to the complexity involved and large data requirements. In the multiscale regression approach used in this work, instead a large number of very low-dimensional statistical modeling problems are solved. Fig 8 illustrates the basic idea behind the approach in the context of predicting L_M day-ahead observed daily flow using the model-simulated daily flow valid over the L_M day-long prediction horizon, and the observed daily flow $L_M - 1$ days into the past. In this approach, rather than predicting $q_i^o, i = 1, \dots, L_M$, using $q_i^s, i = 1, \dots, L_M$, and $q_i^o, i = 1, \dots, L_M - 1$, we predict $a_{k,1}^o = \Sigma_{i=1}^{L_k} q_i^o$ using $a_{k,1}^b = \Sigma_{i=1}^{L_k} q_i^b$ and $a_{k,0}^o = \Sigma_{i=-(L_k-1)}^{0} q_i^o$ for all time scales of aggregation, $k = 1, \dots, M$, where q_i^b denotes the bias-corrected model-simulated daily flow, $\beta_{k,i} q_i^s$ (see Eqs.(3) and (6)).

The predicted multi-daily flow is then dissagregated to daily flow using the granular patterns of daily flow in the bias-corrected model-simulated daily flow. The above approach is motivated by the fact that, the larger the temporal scale of aggregation is, the more skillful $a_{k,1}^b$ is likely to be (Kim et al. 2018; Limon 2019). Similar approaches have also been used in post-processing forecasts of precipitation (Kim et al. 2018; Schaake et al. 2007) and streamflow (Regonda and Seo 2008). In this work, the prediction horizon, L_M, used is 32 days, and the aggregation scales

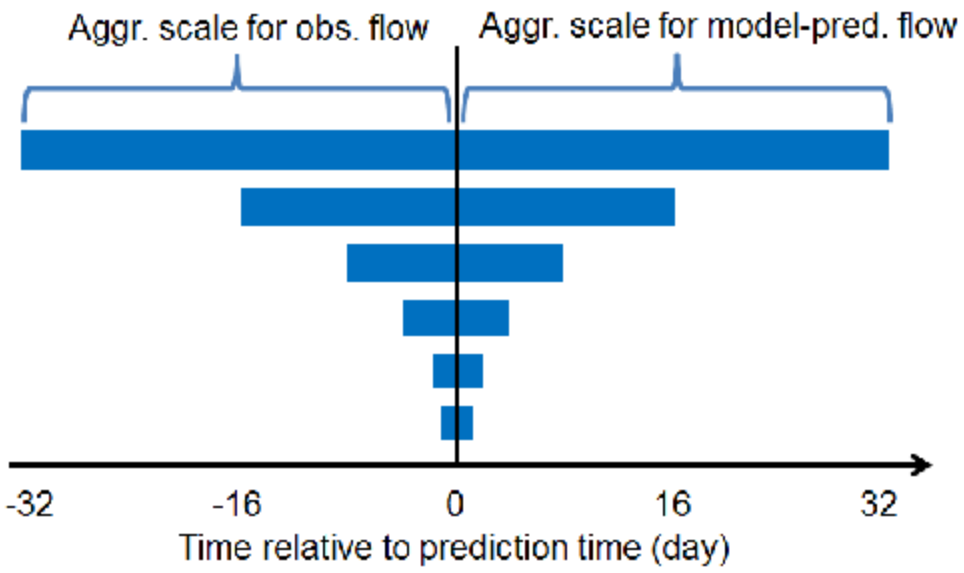


Figure 8: Schematic of multiscale regression.

used are 1, 2, 4, 8, 16, and 32 days (see Fig 8). Depending on the application and the pattern of time scale-dependent predictability, however, one may choose a different set of scales. To predict the observed flow at the k-th scale, the following linear model is used:

$$a^{p}_{k,1} = \varpi_k a^{o}_{k,0} + (1-\varpi_k) a^{b}_{k,1}, k = 1, \ldots, M \quad (5)$$

In the above, $a^{p}_{k,1}$ denotes the predicted, time-aggregated observed flow at the k-th time scale, where the subscript "1" signifies that the prediction is for a single time step ahead at the k-th time scale, and ϖ_k denotes the optimal weight for the time-aggregated observed flow at the k-th scale, $a^{o}_{k,0}$, where the subscript "0" signifies that $a^{o}_{k,0}$ is valid at the current time step at the k-th time scale. The optimal weight, ϖ_k, in Eq.(5) may be obtained via optimal linear (i.e., maximum likelihood) estimation (Schweppe 1973) as:

$$[\varpi_k \ \ (1-\varpi_k)] = [U^T \mathrm{R}^{-1} U]^{-1} U^T \mathrm{R}^{-1} \quad (6)$$

In the above, U denotes the (2×1) unit vector, and R denotes the error covariance matrix:

$$R = \begin{bmatrix} Var\left[a_{k,0}^{o} - a_{k,1}^{o}\right] & Cov[a_{k,0}^{o} - a_{k,1}^{o}, a_{k,1}^{b} - a_{k,1}^{o}] \\ Cov[a_{k,0}^{o} - a_{k,1}^{o}, a_{k,1}^{b} - a_{k,1}^{o}] & Var\left[a_{k,1}^{b} - a_{k,1}^{o}\right] \end{bmatrix} \quad (7)$$

The predicted daily flow for the i-th day from the multiscale regression at the k-the time scale, $q_{k,i}^{p}$, may be obtained by disaggregating $a_{k,1}^{p}, k = 1, \ldots, M$, according to:

$$q_{k,i}^{p} = \frac{q_i^b}{a_{k,1}^{b}} a_{k,1}^{p} = \frac{a_{k,1}^{p}}{a_{k,1}^{b}} q_i^b \quad (8)$$

Eq.(8) amounts to adjusting the bias-corrected model-simulated daily flow, q_i^b, based on how much larger or smaller the predicted time-aggregated flow is relative to the time-aggregated bias-corrected flow, i.e., $a_{k,1}^{p}/a_{k,1}^{b}$. Once the disaggregation process is complete for all time scales of aggregation, the final prediction of observed daily flow, $q_i^p, i = 1, \ldots, L_M$, is constructed from $q_{k,i}^{p}, k = 1, \ldots, M$, by choosing for each day, i, in the prediction horizon, $i = 1, \ldots, L_k$, the predicted daily flow $q_{k,i}^{p}$ associated with the smallest k, i.e., the smallest time scale of aggregation. In this way, if there are multiple predictions with overlapping prediction horizons, the procedure selects the one associated with the shortest lead time. Fig 8 shows the resulting time scales of aggregation over the prediction horizon of 32 days. Albeit heuristic, the above selection rule is based on the extremely reasonable assumption that, the shorter the lead time is, the more skillful $q_{k,i}^{p}$ is. If the period of record is too short relative to the largest time scale of aggregation, the estimation of the error covariance terms in Eq.(7) may not be possible due to small sample size. In such a case, the largest time scale may have to be reduced or dropped. The last element of MS-EnsPost models the error in the above prediction and its temporal structure for ensemble generation, which is described below.

3.2.3 Error modeling and ensemble generation

This element of MS-EnsPost models the time-correlated errors in $q_{p,i}, i = 1, \ldots, L_M$, from multiscale regression. the error, ε_i, is defined in the predicted daily flow, $q_{p,i}$, valid for the i-th day in the prediction horizon as:

$$\varepsilon_i = q_i^o - q_i^p,\ \varepsilon_i \geq -q_i^p \tag{9}$$

where q_i^o denotes the verifying observed daily flow. In EnsPost, q_i^o and q_i^p are normal quantile-transformed (NQT) empirically (Krzysztofowicz and Kelly 2000; Krzysztofowicz and Herr 2001) which renders ε_i normal in the transformed space (Seo et al. 2006). In MS-EnsPost, q_i^o and q_i^p are Box-Cox-transformed (Box and Cox 1964) to avoid data-intensive empirical distribution modeling. We then have for the error in the transformed space, ε_i^t:

$$\varepsilon_i^t = q_i^{to} - q_i^{tp} = \frac{(q_i^o)^\lambda - 1}{\lambda} - \frac{(q_i^p)^\lambda - 1}{\lambda} = \frac{(q_i^o)^\lambda - (q_i^p)^\lambda}{\lambda},\ \varepsilon_i^t \geq -(q_i^p)^\lambda/\lambda \tag{10}$$

In the above, λ denotes the Box-Cox parameter, and q_i^{to} and q_i^{tp} denote the transformed observed and predicted daily flows, respectively. The parameter λ is chosen such that ε_i^t may be approximated with a truncated normal distribution (Robert 1995):

$$f_1\left(\varepsilon_i^t \middle| \varepsilon_i^t \geq q_i^{tp}\right) = N\left(m_{\varepsilon_i^t}, \sigma_{\varepsilon_i^t}^2; q_i^{tp}\right) \tag{11}$$

where $m_{\varepsilon_i^t}$ and $\sigma_{\varepsilon_i^t}^2$ denote the mean and variance, respectively, of ε_i^t conditional on $\{\varepsilon_i^t \geq q_i^{tp}\}$. The mean and variance of ε_i^t in Eq.(11) may be equated with the sample mean and variance estimated from all available observed flow and the corresponding predicted flow as follows:

$$\mu_{\varepsilon_i^t} = \int_0^\infty \int_{-(q_i^p)^\lambda/\lambda}^\infty \varepsilon_i^t f_1\left(\varepsilon_i^t \middle| q_i^p\right) f_2(q_i^p) dq_i^p \tag{12}$$

$$s_{\varepsilon_i^t}^2 = \int_0^\infty \int_{-(q_i^p)^\lambda/\lambda}^\infty (\varepsilon_i^t)^2 f_1\left(\varepsilon_i^t \middle| q_i^p\right) f_2\left(q_i^p\right) dq_i^p - \mu_{\varepsilon_i^t}^2 \tag{13}$$

where $\mu_{\varepsilon_i^t}$ and $s^2_{\varepsilon_i^t}$ denote the sample mean and variance of ε_i^t, respectively, $f_1(\,|\,)$ denotes the conditional probability density function (PDF) of ε_i^t, and $f_2(\,)$ denotes the PDF of q_i^p. Using Eqs.(14) and (15) and the empirical distribution of q_i^p, one may solve for $m_{\varepsilon_i^t}$ and $\sigma^2_{\varepsilon_i^t}$ given the sample estimates of $\mu_{\varepsilon_i^t}$ and $s^2_{\varepsilon_i^t}$. Once $m_{\varepsilon_i^t}$ and $\sigma^2_{\varepsilon_i^t}$ are prescribed, an ensemble realization of ε_i^t, or $\varepsilon_i^t(\omega)$, may be generated from:

$$\varepsilon_i^t(\omega) = \mu_{\varepsilon_i^t} + s_{\varepsilon_i^t}\Phi^{-1}[\Phi(z_i^l) + U(\omega)\{1 - \Phi(z_i^l)\}] \tag{14}$$

In the above, $U(\omega)$ denotes the realization of the [0,1] uniform random variable, and $\Phi(\,)$ denotes the standard normal cumulative distribution function (CDF). The normalized lower bound of $\varepsilon_{t,i}$ in Eq.(14) is given by:

$$z_i^l = \frac{-(q_i^p)^\lambda/\lambda - \mu_{\varepsilon_i^t}}{s_{\varepsilon_i^t}} \tag{15}$$

An ensemble trace of the post-processed daily flow, $q_i^o(\omega)$, may then be obtained from:

$$q_i^o(\omega) = q_i^p + \varepsilon_i(\omega) = \{(q_i^p)^\lambda + \lambda\varepsilon_i^t(\omega)\}^{1/\lambda}, \varepsilon_i^t(\omega) \geq -(q_i^p)^\lambda/\lambda \tag{16}$$

where $q_i^o(\omega)$ and $\varepsilon_i(\omega)$ denote the ensemble realizations of q_i^o and ε_i, respectively.

The error modeling as described above requires estimation of λ, $\mu_{\varepsilon_i^t}$, and $s^2_{\varepsilon_i^t}$ that render the distribution of ε_i^t, approximately truncated normal given q_i^{tp} (see Eq.(10)). In addition, Eq.(11) assumes that ε_i^t is approximately homoscedastic with respect to q_i^{tp} except near the origin where the lower bound strongly suppresses variability. In reality, the above assumptions may not be met for all basins. In addition, there may not be enough data points over the tail ends of q_i^{tp} to test the conditional truncated normality or homoscedasticity. In this work, the reasonableness of the above assumptions is checked by examining for each basin the sample moments, normal quantile plots, histograms, and scatter plots of ε_i^t vs. q_i^{tp}. For those basin showing significant

departures from truncated normality or homoscedasticity, λ and/or $s_{\varepsilon_i^t}^2$ are adjusted until the results passed the visual test. Figs 9 and 10 show examples of how λ is adjusted based on the mean CRPS results. Note that, for HUNP1, it was necessary to adjust λ significantly from the nominal optimum associated with zero skewness, but that, for MPLP1, no adjustment was necessary for λ as the nominal optimum produced the minimum mean CRPS.

Admittedly, the above error modeling procedure is less than fully objective for improvement of which additional research is necessary. In this study, only truncated normal is considered for $f_1\left(\varepsilon_i^t \middle| q_i^p\right)$ for simplilcity. Other distributions, such as truncated gamma (Chapman 1956), are also possible.

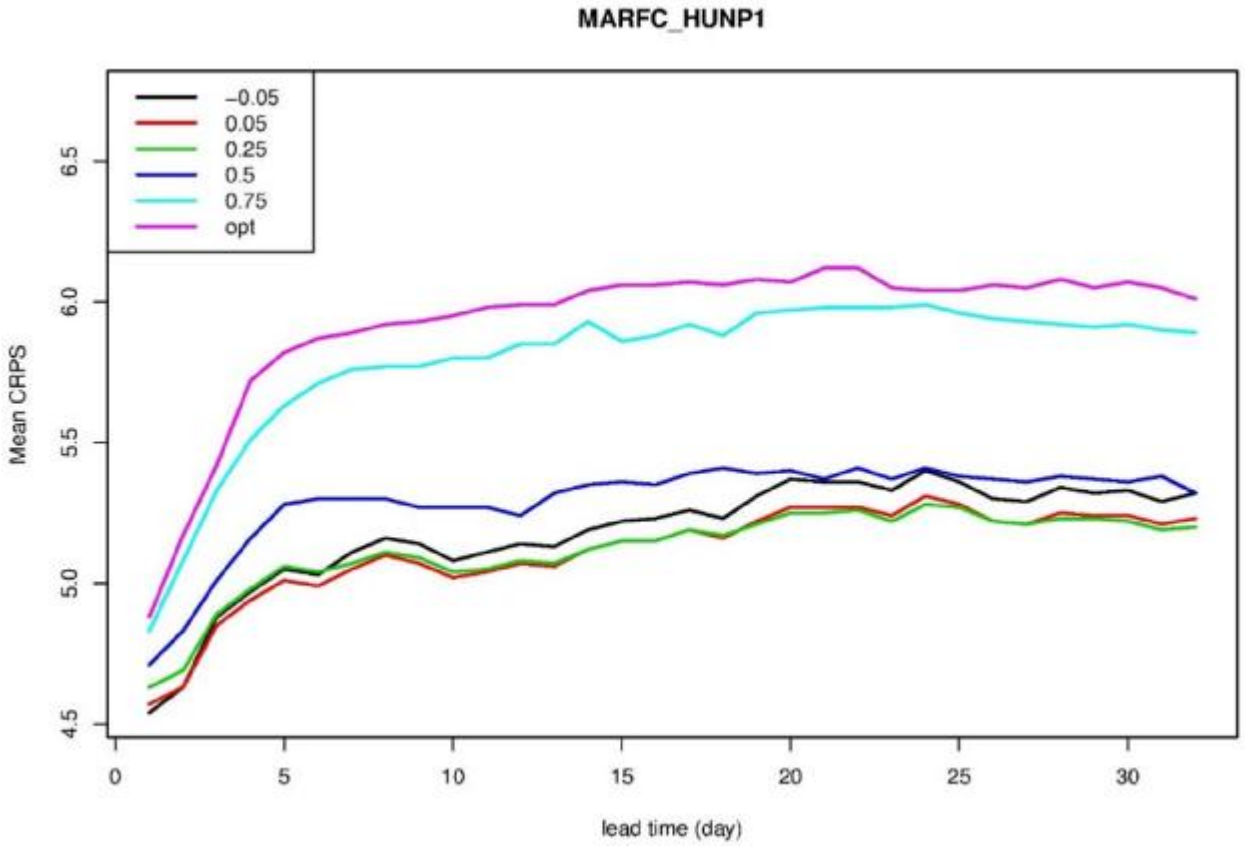


Figure 9: Adjustement of λ based on mean CRPS for HUNP1 basin in Juniata River Basin, PA.

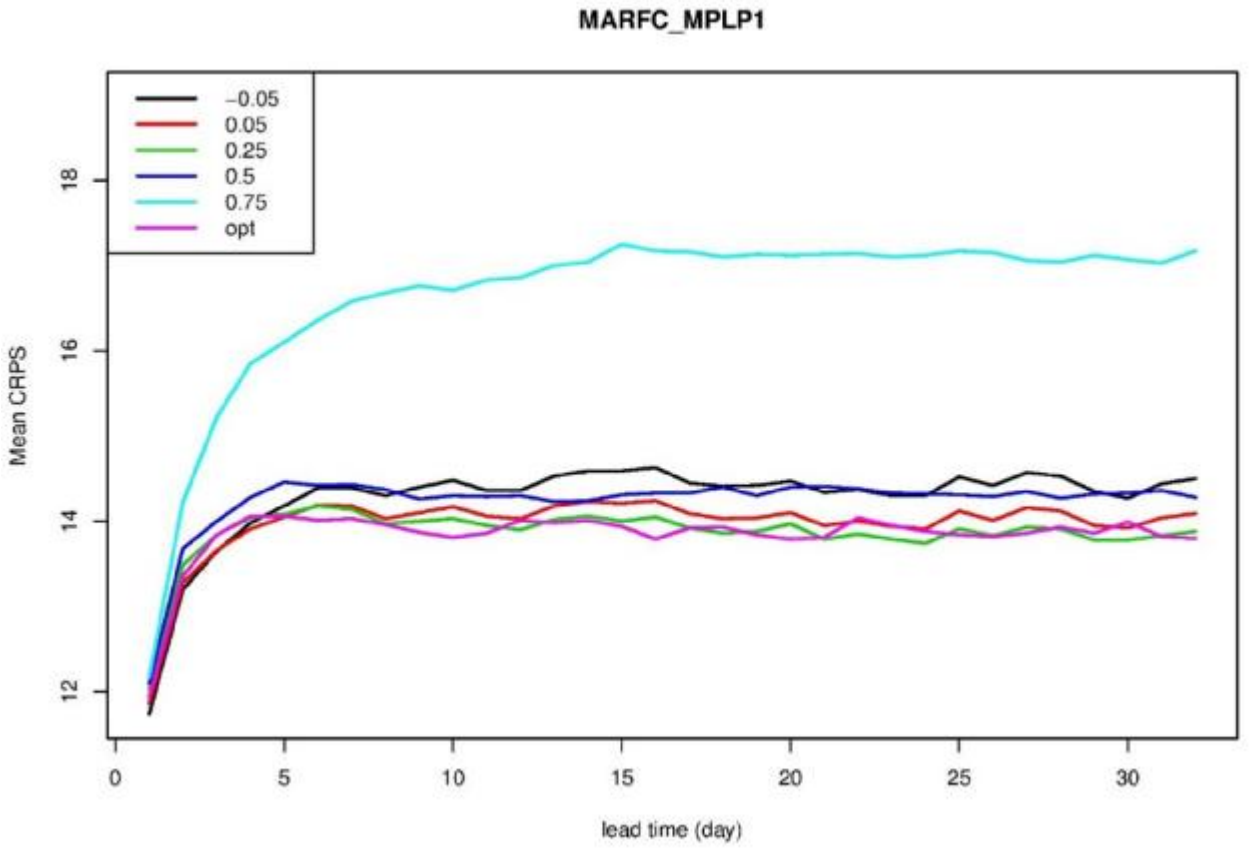


Figure 10: Adjustement of λ based on mean CRPS for MPLP1 basin in Juniata River Basin, PA.

To capture the distributional characteristics of multi-daily flow, it is necessary to model the temporal dependence of ε_i^t. Due to the large number, basin-specific modeling of error time series (Box and Jenkins 1976) was outside of the scope of this study. Instead, the error, $\varepsilon_i^t = q_i^{to} - q_i^{tb}$, is modelled with AR(1) as a first-order approximation for all basins. The use of q_i^{tb} rather than q_i^{tp} in the above is motivated by the fact that q_i^{tb} is not lead time-dependent, and hence greatly simplifies the modeling. The impact of this simplification is relatively small compared to the goodness of the time series modeling of $\varepsilon_i^t = q_i^{to} - q_i^{tb}$. To assess the adequacy of AR(1), structure identification were carried out for a small number of basins in the MA-, NC-, NE-, and NWRFCs' service areas. The results indicate that the error structures are generally more complex

than AR(1), and contain both autoregressive and moving-average components of higher order. Appendix B shows examples of the time series modeling results. The above is not very surprising given the very widely varying hydroclimatology of the basins and goodness of the hydrologic modeling. The simplifying choice of AR(1) in this work is additionally motivated by its use in EnsPost which facilitates direct comparison between MS-EnsPost and EnsPost. Though limited in sample size, the above findings suggest that additional improvement in ensemble prediction of multi-daily flow may be possible with improved modeling of temporal dependence of prediction error.

Chapter 4

Study basins and data used

For MS-PM, a total of 34 basins, comprising 13, 8, 7 and 6 in the service areas of CB-, CN-, MA- and WGRFC, respectively, are used. They were also used in previous studies of EnsPost (Regonda and Seo 2008; Seo et al. 2006). For MS-EnsPost, a total of 139 basins, comprising 11, 13, 7, 19, 13, 28, 42, and 6 in the service areas of CB-, CN-, MA-, MB-, NC-, NE-, NW-, and WGRFC, respectively, are used (see Fig 11). The basin areas range from 91 to 30,700 km^2 with 48, 41 and 50 basins under 500, between 500 and 1,000, and over 1,000 km^2, respectively. The basins cover a wide range of hydroclimatology as may be seen in mean annual precipitation (Fig 11), the aridity index (Fig 12) and the fraction of precipitation as snow (Fig 13).

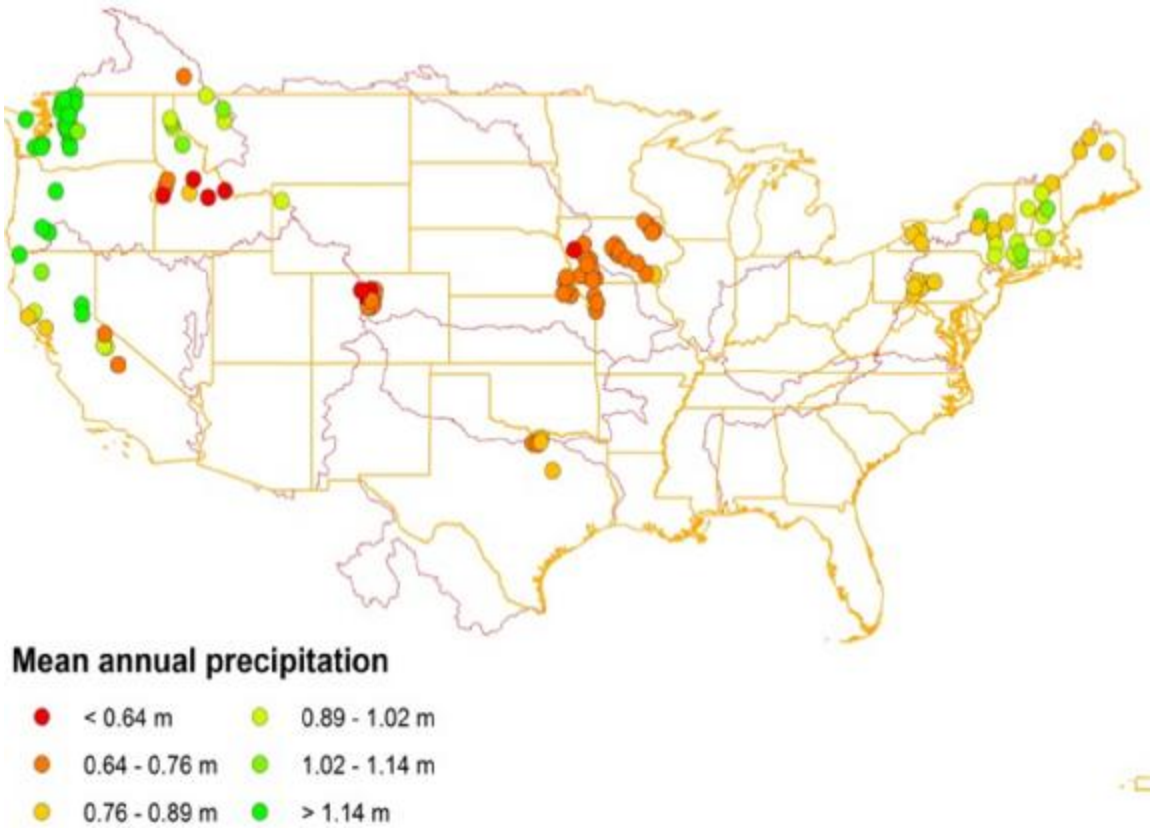


Figure 11: Map of mean annual precipitation.

The aridity index, φ, is defined as (Budyko et al. 1974):

$$\varphi = \frac{\bar{E}}{\bar{P}} \tag{17}$$

where $\bar{E}$ and $\bar{P}$ denote the mean potential evaporation and precipitation (in mm/day), respectively. Fig 12 shows the location and the corresponding aridity index for each basin in the study area. Most of the basins in the northeast and the Pacific northwest are humid and hence have small values of aridity index whereas basins in midsection of the continent and those in California are generally semi-arid to arid.

The fraction of precipitation as snow, f_s, is defined as:

$$f_s = \frac{\bar{P}[T<0°C]}{\bar{P}} \tag{18}$$

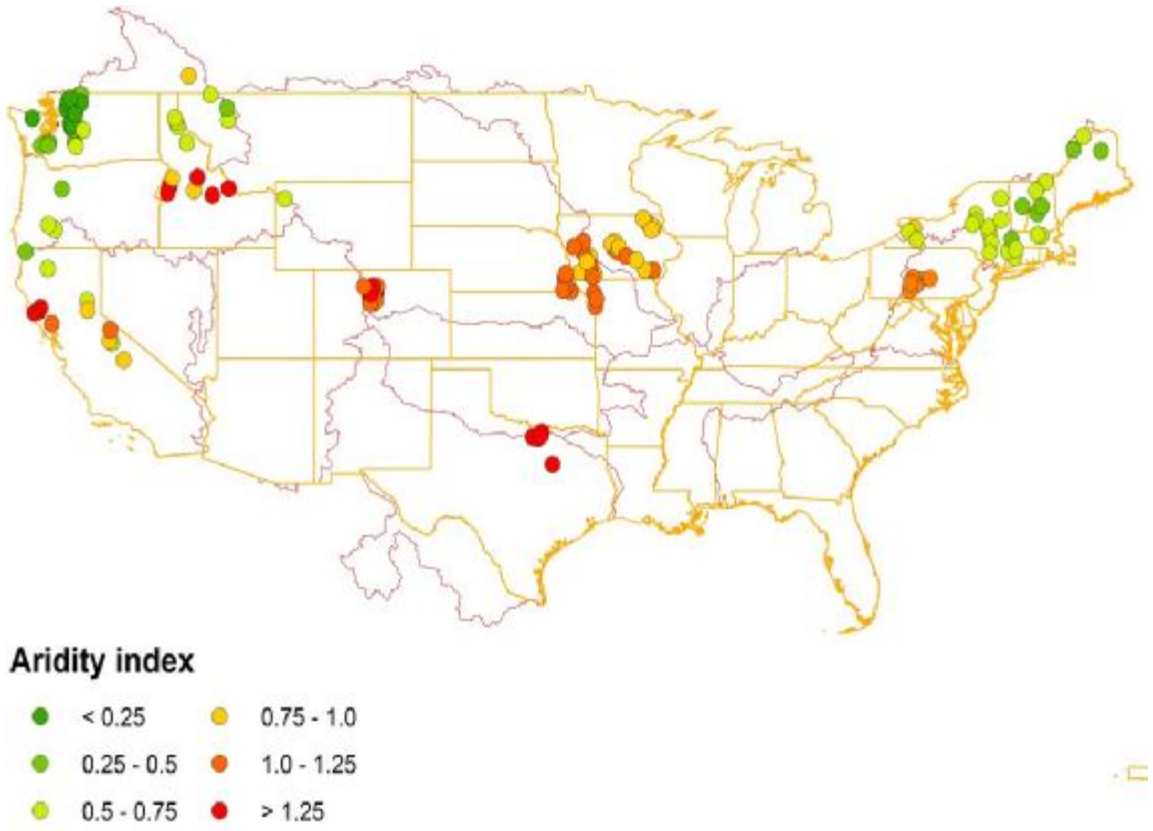


Figure 12: Map of aridity index.

where T denotes the surface air temperature in ℃ and $\bar{P}[\ ]$ denotes the mean of precipitation for which the event bracketed holds true. The fraction of precipitation as snow contributes to the accumulation of snow in winter, a delay in contribution to soil moisture and recharge to groundwater and melting during spring, contributing to a longer lag in streamflow generation (Berghuijs et al. 2014). Fig 13 shows the location and the fs value for the basins used in this research. Snow-driven basins along the Rocky Mountains in Colorado, Idaho and Montana have fs exceeding 50 percent. The majority of the basins in the Middle Atlantic and Northeast have fs of 0.1 to 0.3, and only a small fraction of precipitation occurs as snow for those in the Midwest and Texas.

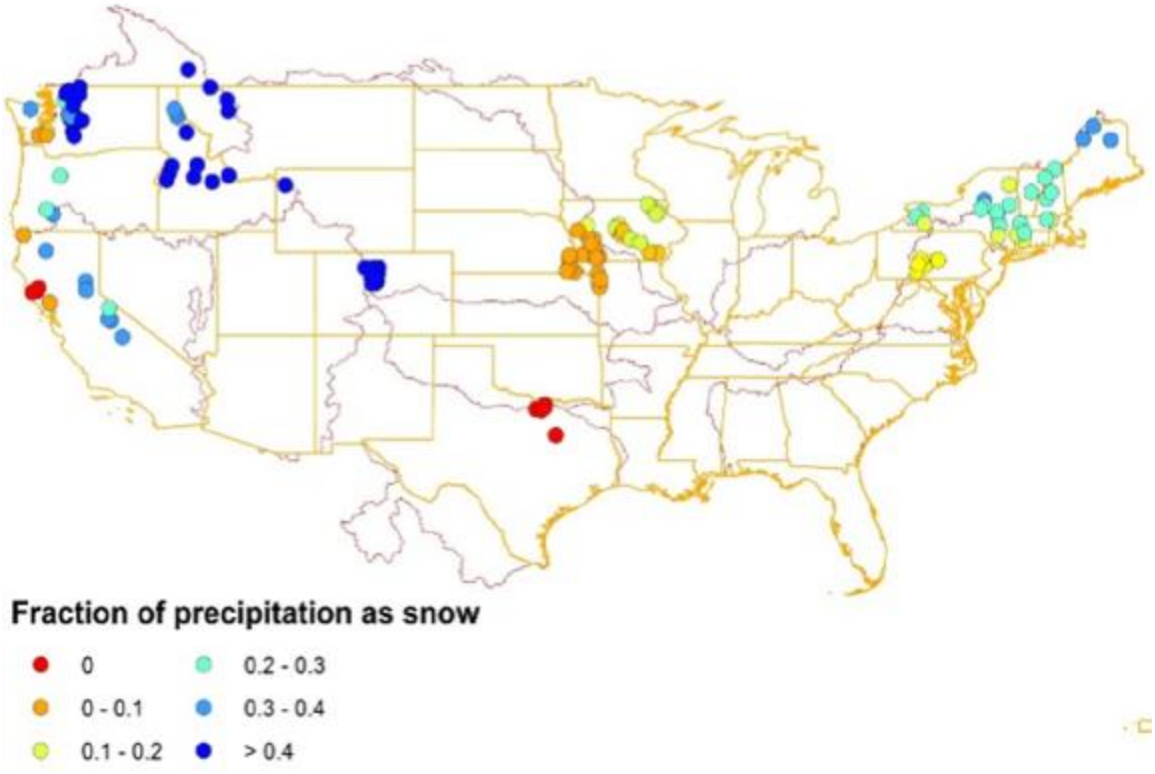


Figure 13: Map of fraction of precipitation as snow.

In general, an arid basin has smaller predictability of streamflow than a humid basin, and a snow-driven basin has larger predictability than a rainfall-driven basin. Appendix C shows mean monthly streamflow for all basins used in this research.

The data used for this study are mean daily observed and simulated streamflow. The historical observed mean daily streamflow, referred to as QME in the NWS, is obtained from the US Geological Survey. The focus of this work is on reducing and quantifying hydrologic uncertainty. As such, the model output of interest is the simulated streamflow, which reflects hydrologic uncertainty only, rather than the streamflow forecast, which reflects both input and hydrologic uncertainties (Krzysztofowicz 1999; Seo et al. 2006). The simulated mean daily flow, or SQME, is derived from the simulated instantaneous flow, or SQIN, generated at a 6-hr interval using the operational hydrologic models, and the observed forcings of mean areal precipitation, temperature, and potential evapotranspiration. For the remainder of this dissertation, by daily flow, it is meant mean daily flow. The hydrologic models used are the SAC (Burnash et al. 1973) for soil moisture accounting, UH (Chow et al. 1988) for surface runoff routing, and SNOW-17 (Anderson 1973) for snow ablation. The MARFC uses the continuous Antecedent Precipitation Index model (API-CONT; Fedora and Beschta 1989; Sittner et al. 1969) instead of SAC. The SQIN time series were produced by the respective RFCs using the Community Hydrologic Prediction System (CHPS; Gijsbers et al. 2009) based on the RFCs' historical forcings and calibrated model parameters. The CHPS is currently the main operational forecasting system at the RFCs, and uses the single (RES-SNGL) and joint (RES-J) reservoir regulation models, and the SSARR reservoir regulation (SSARRESV) model for simulation of reservoir operations (Adams III 2016; NWS 2008a, 2008b). RES-SNGL can model sophisticated operating rules, but only for reservoirs operating individually, whereas RES-J can model

multiple reservoirs with limited operating rules, but only from up- to downstream (Limon 2019). The SSARRESV is based on the Streamflow Synthesis and Reservoir Regulation System developed by the US Army Corps of Engineers and NWRFC (NWS 2015) with which flows may be routed as a function of multi-variable relationships involving backwater effects from a downstream reservoir. Reservoir models were included in the hydrologic modeling of 20 of the 139 locations used for MS-EnsPost evaluation (Chapter 6). There are about 14 additional locations impacted by reservoir regulations which are not modeled. Limon (2019) has shown that, for a water supply reservoir in North Texas, the magnitude of reservoir modeling uncertainty may be comparable to that of all other hydrologic uncertainties combined, and may even approach that of the input uncertainty. As such, flow regulations present a large additional challenge to streamflow post-processing. Experience thus far indicates that at least 20 years' worth of data is necessary for estimation of the EnsPost parameters (NWS 2015). The period of record used in this work common to both QME and SQME time series ranged from 30 to 54 years for MS-PM evaluation (Subsection 5.1) and 12 to 66 years (which exceeded 30 years for over 90 percent of the basins) for MS-EnsPost evaluation (Subsection 5.2)

Chapter 5

Evaluation

This section describes how MS-PM and MS-EnsPost are evaluated. For MS-PM, the evaluation is limited to single-valued predictions only whereas for MS-EnsPost it is for both single-valued and ensemble predictions.

5.1 MS-PM

We assess MS-PM for 34 basins in the service areas of CB-, CN-, MA- and WGRFCs. The following three approaches are evaluated:

A1) PM at a single multi-daily scale

A2) A1 but with generation and averaging of multiple CDFs

A3) A2 but at multiple multi-daily scales

In A1, for a chosen time window of length K (days), PM is performed at that scale, and the CDF of the simulated flow is matched with that of observed flow in that scale. One may repeat the above for time windows of different lengths (Regonda et al. 2008). Figure 14 illustrates the above operation at a K-daily scale. In A2, we apply A1 for different starting days of 1 through K within the first window and the resulting K adjusted daily streamflows are averaged. In A3, we apply A2 for the largest window and replace the original simulation with the averaged daily simulation from A2. Then, we use the CDF of averaged time series of simulated flow for matching with that of observed flow in the second largest window. This process is repeated until the smallest time window is reached. Fig 15 shows examples of A1 (left panels) and A2 (right panels) for a range of multi-daily aggregation scales. The RMSE results for A1 (upper-left panel) indicate that multi-daily correction improves over daily correction for all multi-daily scales of

aggregation, but that the margin of improves varies significantly from one multi-daily scale to another. The RMSE results for A2 (upper-right panel) indicate that using the average CDFs reduce scale-to-scale variations from reduced sampling uncertainty, that, for this particular example, 5- and 13-daily scales of aggregation provide the largest improvement, and that all multi-daily scales significantly improve over daily scale.

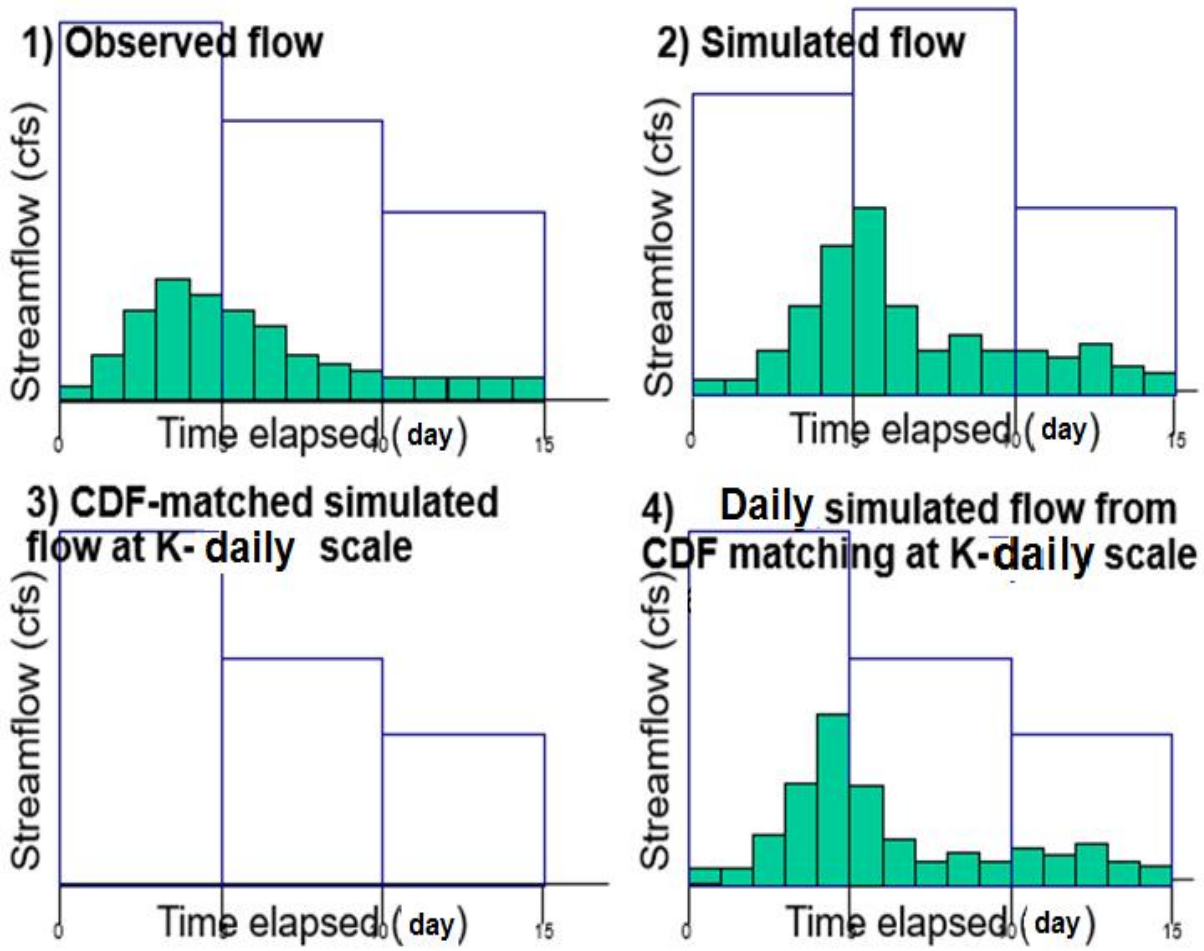


Figure 14: Schematic of multi daily CDF-matching (from Regonda and Seo 2008).

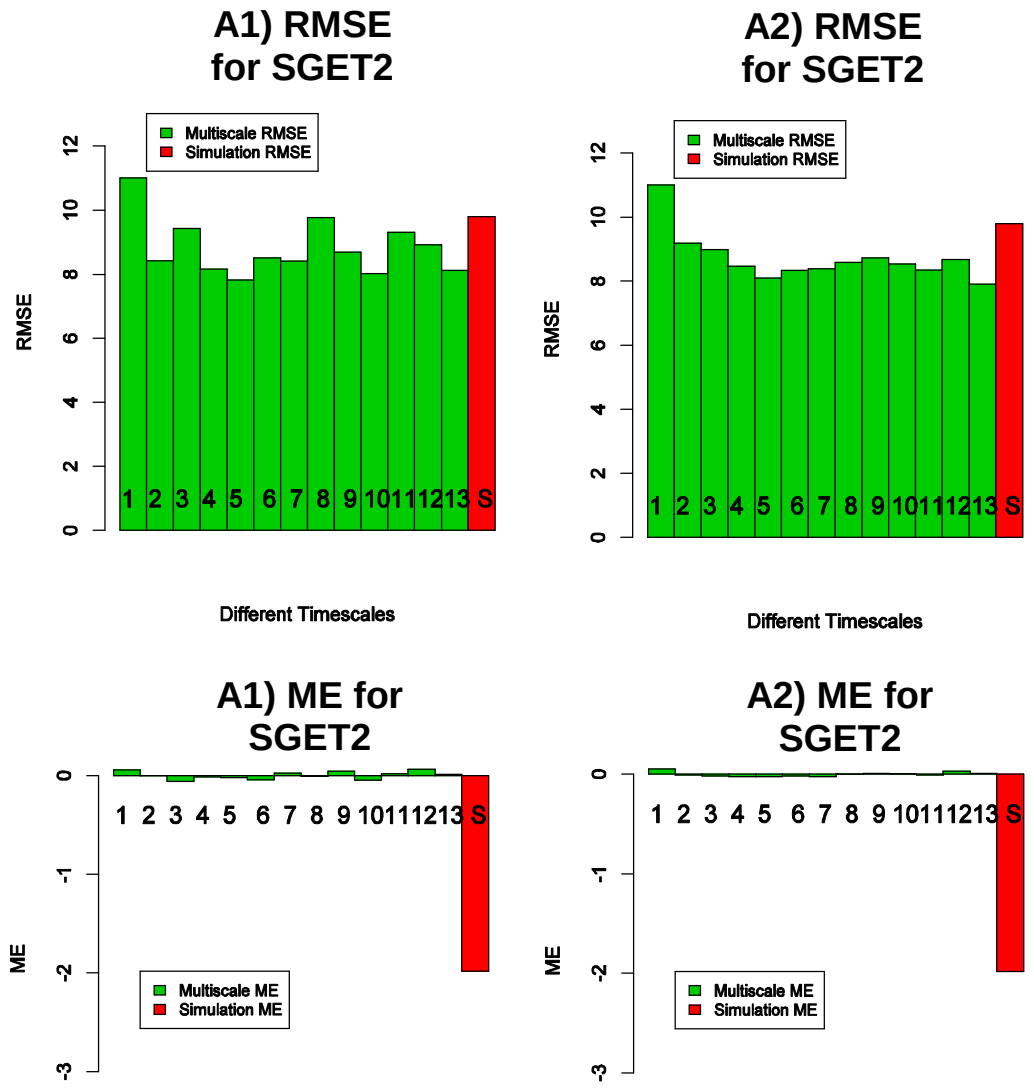


Figure 15: An example error statistics of PM at a single multi-daily scale without (left panel) and with (right panel) generation and averaging of multiple CDFs for SGET2 in Upper Trinity River Basin.

5.2 MS-EnsPost

For comparative evaluation of MS-EnsPost, both single-valued and ensemble verification of MS-EnsPost via leave-two-years-out cross validation were carried out. The leave-one-year-out cross validation results are similar. In single-valued verification, the raw and bias-corrected predictions, and ensemble mean predictions from post-processing with EnsPost and MS-EnsPost

are evaluated. In ensemble verification, the ensemble predictions from EnsPost and MS-EnsPost are evaluated, and their skill in reference to sample climatology of historical observed flow is assessed. In both, predictions of daily flow with lead times of 1 to 32 days, and of monthly flow with a lead time of one month are considered. For single-valued predictions, RMSE is used as the primary measure of performance:

$$\text{RMSE(k)} = \sqrt{\frac{1}{n(k)}\sum_{i=1}^{n(k)}[q_i^p(\text{k}) - q_i^o(\text{k})]^2} \quad (19)$$

where $q_i^p(k)$ denotes the i-th Day-k prediction of daily flow; $q_i^o(\text{k})$ denotes the verifying observed daily flow; and $n(k)$ denotes the total number of Day-k daily flow predictions. For ensemble predictions, the mean Continuous Ranked Probability Score (CRPS), its decomposition, and mean Continuous Ranked Probability Skill Score (CRPSS) are used as primary measures (Brown and Seo 2010; Kim et al. 2018). The CRPS represents the integral squared difference between the CDF of the predicted variable, $F_Y(q)$, and that of the verifying observed variable, $F_X(q)$ (i.e., a step function):

$$CRPS = \int\{(F_Y(q) - F_X(q)\}^2\, dq \quad (20)$$

The mean CRPS is the average of the CRPS values from the individual pairs of ensemble forecasts and observations and reflects the overall quality of an ensemble forecasting system (the smaller, the better). The CRPS is decomposed into reliability (REL), resolution (RES), and uncertainty (UNC), or into REL and potential CRPS ($CRPS_{POT}$) (Hersbach 2000):

$$CRPS = REL - RES + UNC = REL + CRPS_{POT} \quad (21)$$

Smaller REL indicates more reliable ensembles (desirable) and larger RES means better resolution (desirable). The RES component (=UNC – $CRPS_{POT}$) is positive if the ensemble forecast is better than the climatological ensemble forecast (Hersbach 2000). The UNC component reflects climatological uncertainties in the observations and does not relate to forecast

attributes. The $CRPS_{POT}$ (=CRPS – REL) represents the CRPS for a perfectly reliable forecast (Hersbach 2000). Similarly to the CRPS, smaller $CRPS_{POT}$ indicates smaller error or greater forecast quality. The mean CRPSS measures this skill relative to climatology, i.e., historical traces of observed daily flow valid at the same time of the year as the subject forecast:

$$\overline{CRPSS} = \frac{\overline{CRPS}_{clim} - \overline{CRPS}}{\overline{CRPS}_{clim}} \tag{22}$$

Perfect and skill-less ensemble forecasts have mean CRPSS of unity and zero, respectively.

Chapter 6

Results and discussion

This section summarizes the evaluation results for MS-PM and MS-EnsPost. For MS-PM, the evaluation is limited to single-valued results only whereas for MS-EnsPost it is for both single-valued and ensemble results.

6.1 MS-PM

MS-PM was applied for 34 basins in four different RFCs. Time windows of 1 to 5 days were used for the CN- and WGRFC basins, and 1 to 128 days for the MA- and CNRFC basins. The choice for the latter is to capture the temporal correlation structure at larger scales of aggregation. For evaluation of single-valued predictions, RMSE is used as the primary measure of performance:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[q_i^b - q_i^o]^2} \tag{23}$$

where q_i^b denotes the i-th bias-corrected simulation of daily flow; q_i^o denotes the verifying observed daily flow; and n denotes the total number of daily flow predictions. Fig 16 shows the CBRFC basin results. They are based on leave-one-year-out cross validation. In the figure, the red bars indicate the reduction in RMSE by the CDF-matched simulation at a daily scale over the raw simulations. The blue, cyan and green bars indicate the RMSE reductions from A1, A2 and A3, respectively. Note that PM at daily scale is not able to reduce RMSE over raw simulation. The single multi-daily scale (i.e., A1) results show positive reduction in RMSE, an indication that the multiscale approach provides improvement. Single and multiple multi-daily scales with CDF averaging (i.e., A2 and A3) reduce RMSE by 5% or more for the majority of the basins and outperform A1. Fig 17 shows the CNRFC results which indicate negative or little reduction

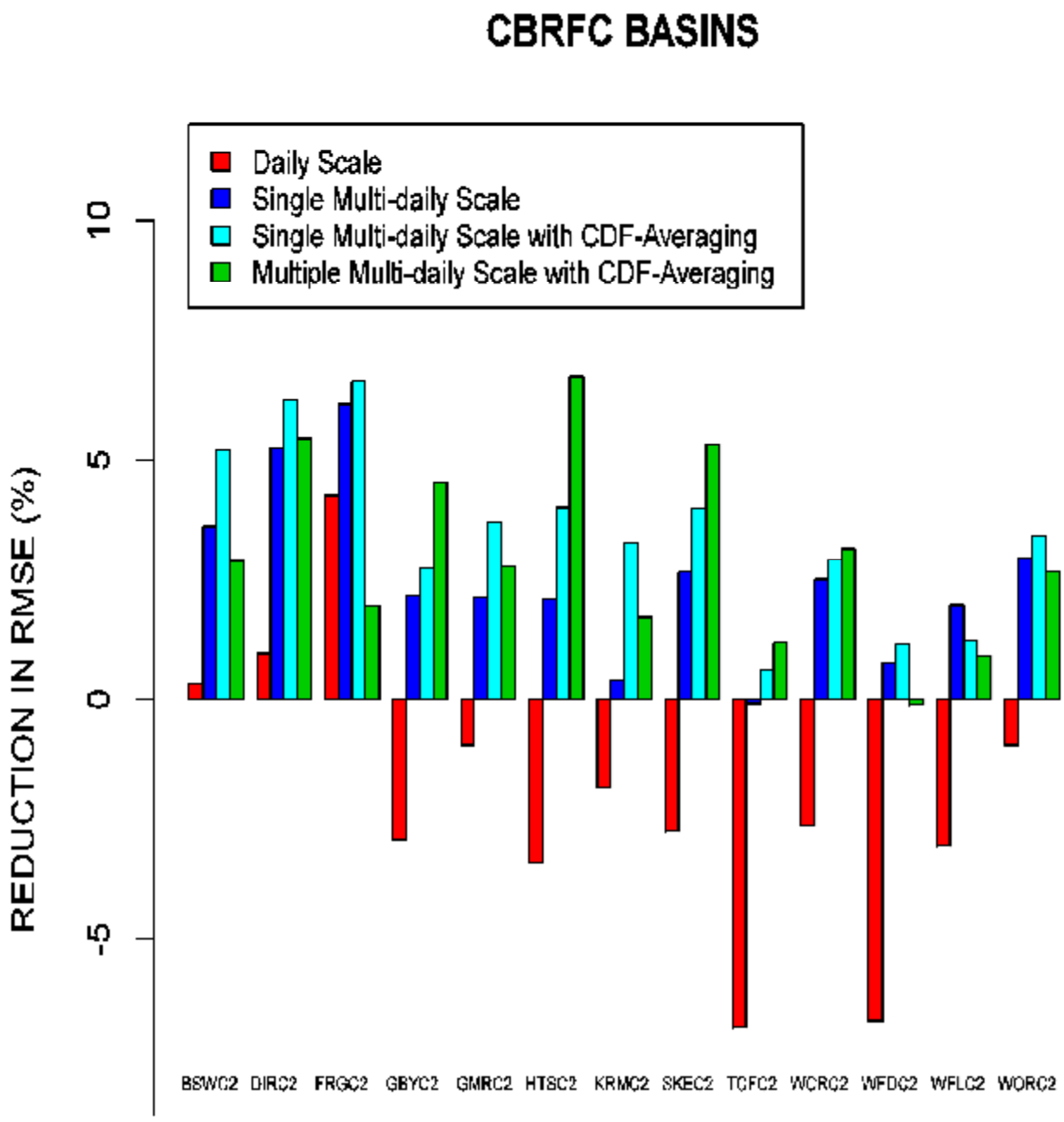


Figure 16: Root mean square error results for CBRFC basins.

in RMSE with PM at daily scale. HOPC1 shows very large reduction which is due to large biases present in model-simulated streamflow. All three MS-PM techniques are able to reduce RMSE for five basins. Fig 18 shows results for the MARFC basin all of which are located in the Juniata River Basin. It is seen that PM at daily scale reduces biases effectively for three out of seven

basins. However, more improvement can be gained by PM at multi-daily scales for all seven basins. Fig 19 shows the results for the six headwater basins in the Upper Trinity River Basin in north Texas. For five of the basins, the three techniques are able to reduce biases by over 10%.

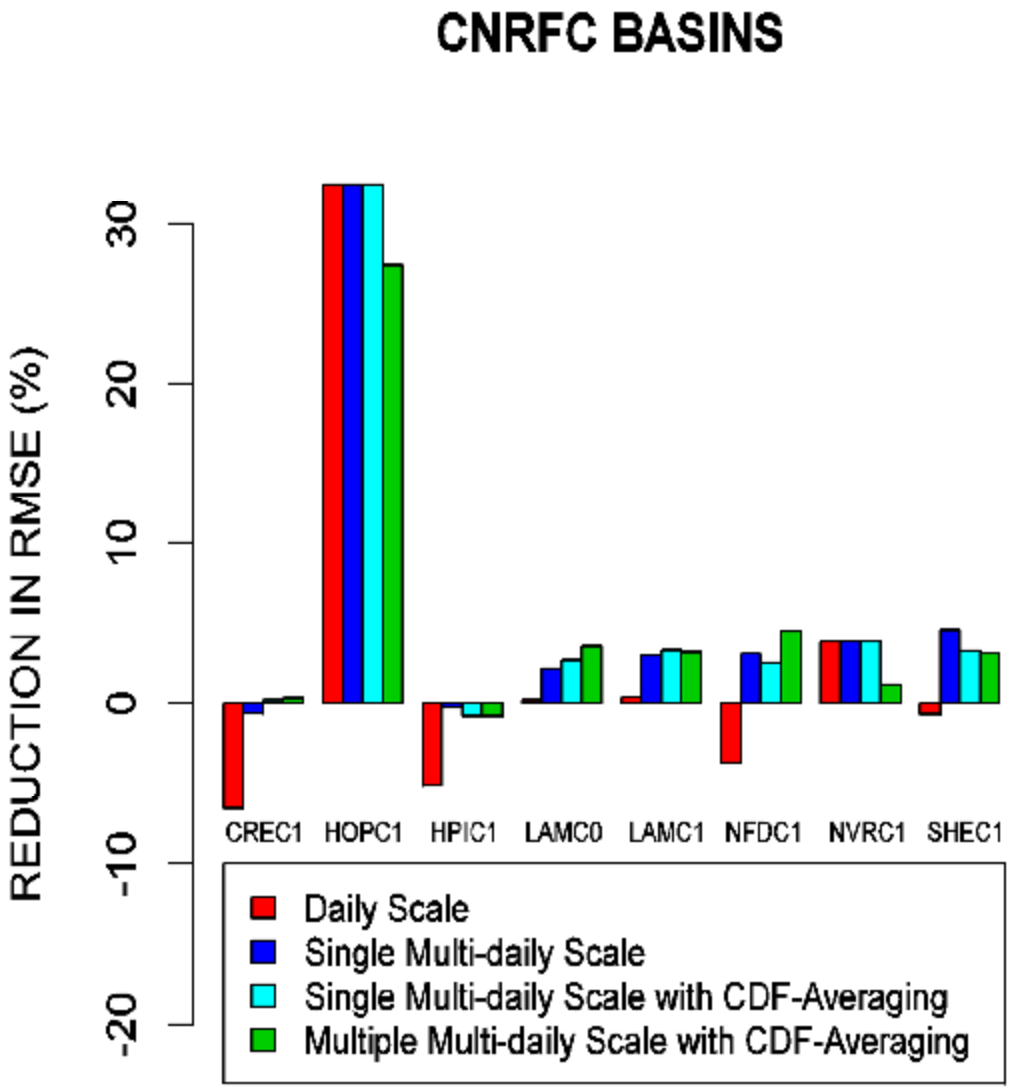


Figure 17: Root mean square error results for CNRFC basins.

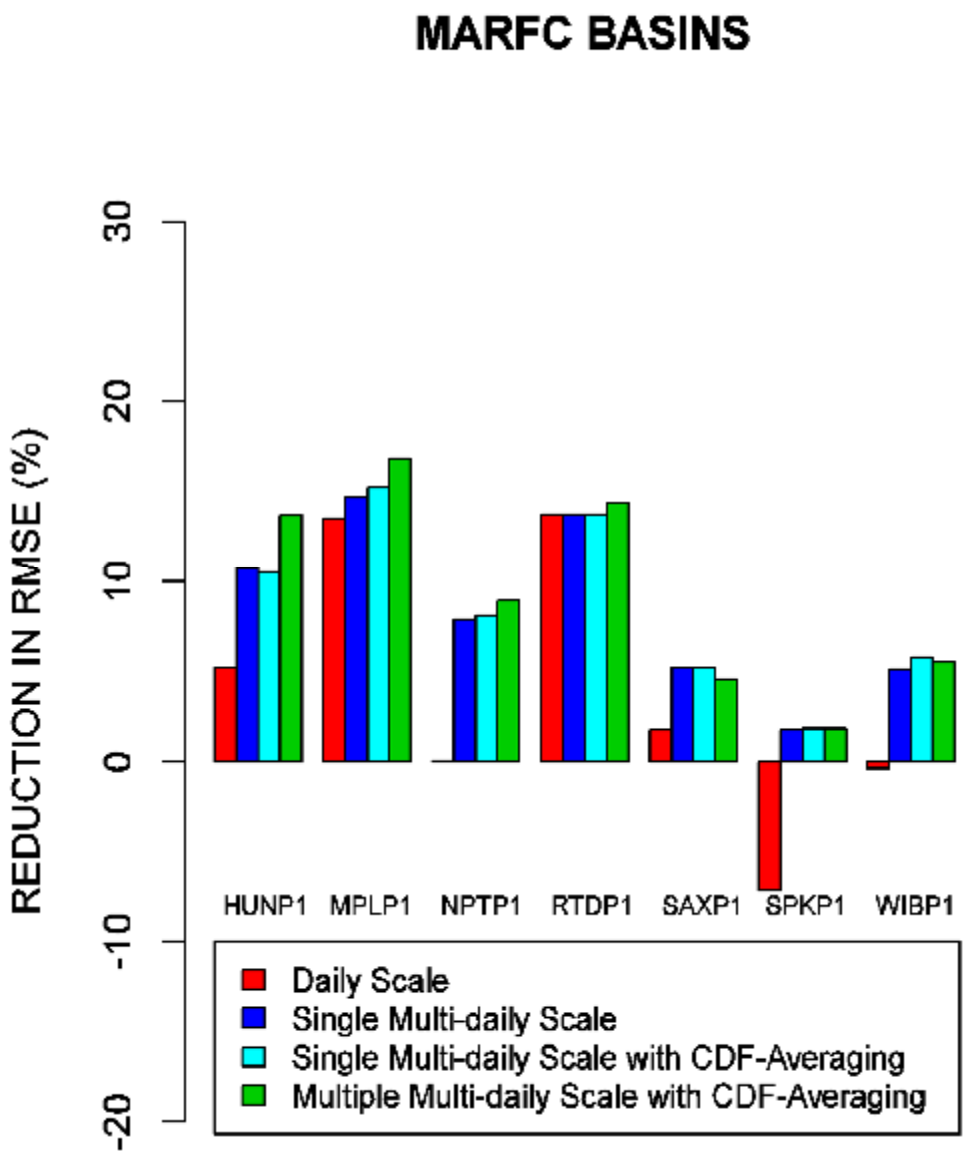


Figure 18: Root mean square error results for MARFC basins.

For four of the basins, PM at daily scale iscreases RMSE over raw model simulations whereas the three MS-PM techniques reduce RMSE for all five basins. BRPT2, however, shows slight increase in RMSE using MS-PM at multiple scales which is presumably due to large sampling uncertainties as well as additional degrees of freedom introduced at large scales.

The MS-PM results above show that, perhaps not surprisingly, PM suffers from sampling uncertainties at larger time scales due to increasingly smaller sample size.

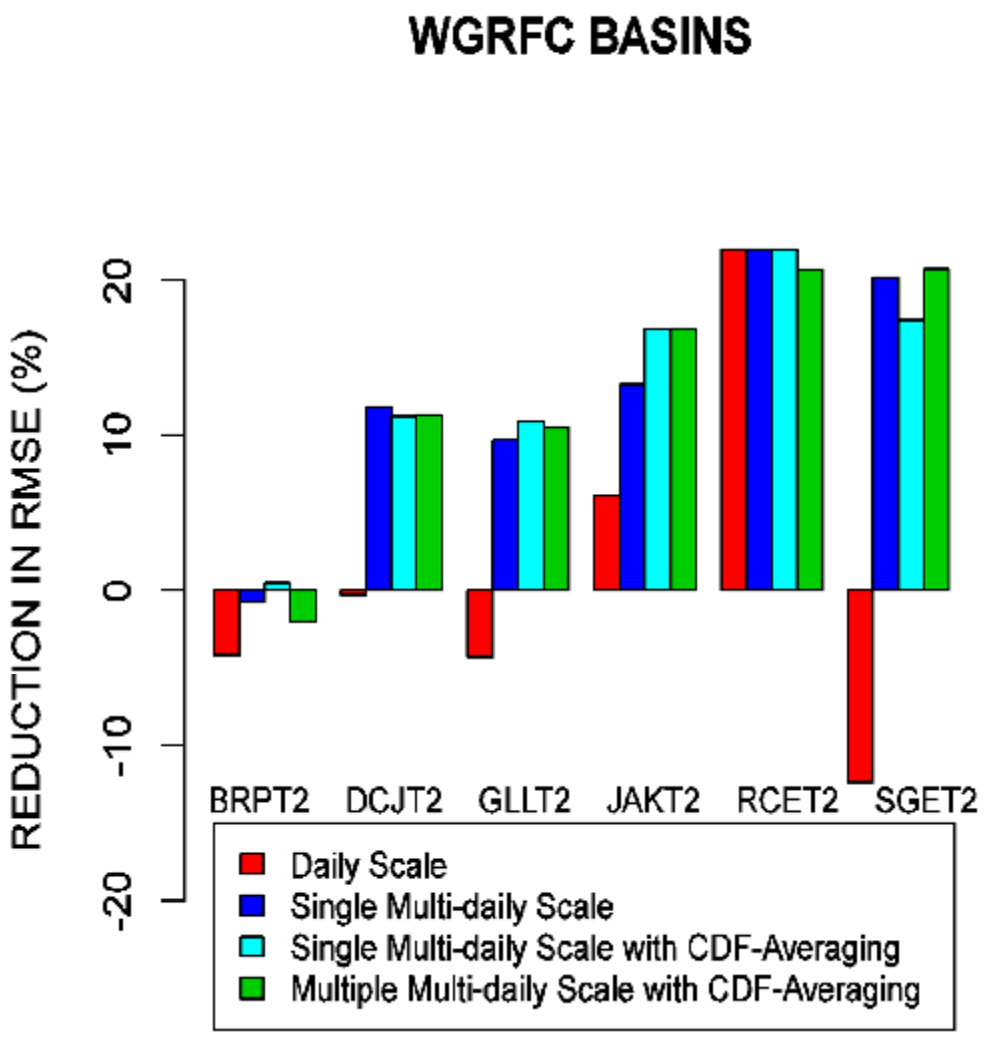


Figure 19: Root mean square error results for WGRFC basins.

In addition, whereas single-scale PM at a multi-daily scale often corrects biases in daily flow more effectively than PM at a daily scale, MS-PM over a range of temporal aggregation scales does not necessarily improve over PM at a single scale due presumably to the sampling uncertainties that accrue over multiple PM operations. On the other hand, larger aggregation scales may be necessary, even with larger uncertainties for, e.g., ephemeral basins in semi-arid to arid regions.

6.2 MS-EnsPost

This section presents the comparative evaluation results for single-valued and ensemble predictions, and assesses the predictability of streamflow as measured from the ensemble prediction results for different hydroclimatological regimes.

6.2.1 Single-valued streamflow prediction

Fig 20 through Fig 27 show the RMSE of the raw, bias-corrected, MS-EnsPost ensemble mean, and EnsPost ensemble mean streamflow predictions for lead times of 1 to 7 days, and 1 month for the basins in eight River Forecast Centers' service areas. In these figures, the RMSE values for each basin are connected to help assess the relative performance among the four different predictions for each basin. Reduction in RMSE by the bias-corrected prediction over the raw is an indication that significant magnitude-dependent biases exist in the raw model-simulated flow due to parametric or structural errors in the hydrologic models, biases in the forcings, or flow regulations. A reduction in RMSE by the MS-EnsPost ensemble mean prediction over the bias-corrected is due to multiscale regression, and indicates that significant uncertainties exist in the initial conditions of the hydrologic models, or significant hydrologic memory exists in the surface and soil water storages of the basin. The monthly results (right-most columns in Fig 20 through Fig 27) reflect bias correction, which impacts over the entire forecast horizon, more than multiscale regression, which impacts only over the range of hydrologic memory. Due to the temporal aggregation, the monthly results amplify the relative performance of the bias correction components of MS-EnsPost and EnsPost. The results for all 139 basins indicate that, MS-EnsPost reduces the RMSE of the raw model predictions of daily flow by 5 to 74 percent and when compared to the EnsPost predictions, by 5 to 68 percent, and

that MS-EnsPost is superior to EnsPost for 1 month-ahead streamflow prediction for all basins examined in this work. Below the main RMSE results for each of the 8 RFCs are summarized.

For CBRFC (Fig 20), the MS-EnsPost ensemble mean prediction improves over the raw and EnsPost ensemble mean predictions for all basins. The yellow circles in the figure indicate that the basin has flow regulations that are modeled with CHPS. In general, both bias correction and multiscale regression contributes to the improvement by MS-EnsPost. For a number of basins, the reduction in RMSE due to multiscale regression persists to Day 4 and beyond, a reflection of the longer hydrologic memory present in the Upper Colorado River Basin owing to the snowpack.

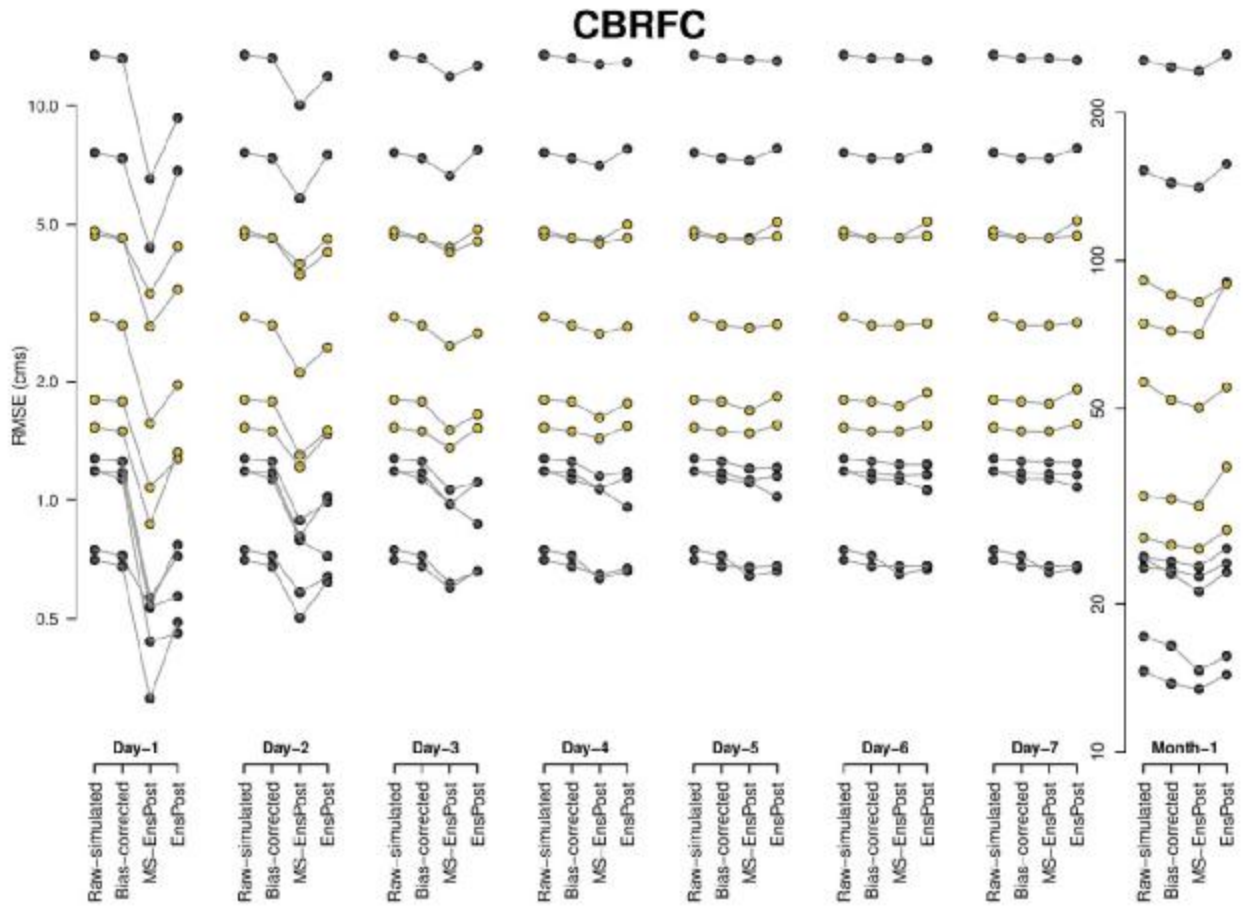


Figure 20: RMSE of the raw, bias-corrected, MS-EnsPost ensemble mean, and EnsPost ensemble mean predictions for lead times of 1 to 7 days, and 1 month for the basins in the CBRFCs' service area (yellow dots indicate basins with reservoir model included).

For CNRFC (Fig 21), MS-EnsPost improves over the raw model and EnsPost predictions for almost all basins. In Fig 21, the empty circles indicate that the basin has unmodeled regulated flow. Also, dots with green and blue outline indicate the basins in coastal range and Sierra Nevada mountain range, respectively. In general, both bias correction and multiscale regression contribute to the reduction in RMSE. The magnitude of reduction, however, is not as large as that for CBRFC. Also, the effect of multiscale regression is shorter-lived than that for CBRFC. For the coastal basins, the impact of multiscale regression is smaller due to the weaker hydrologic memory.

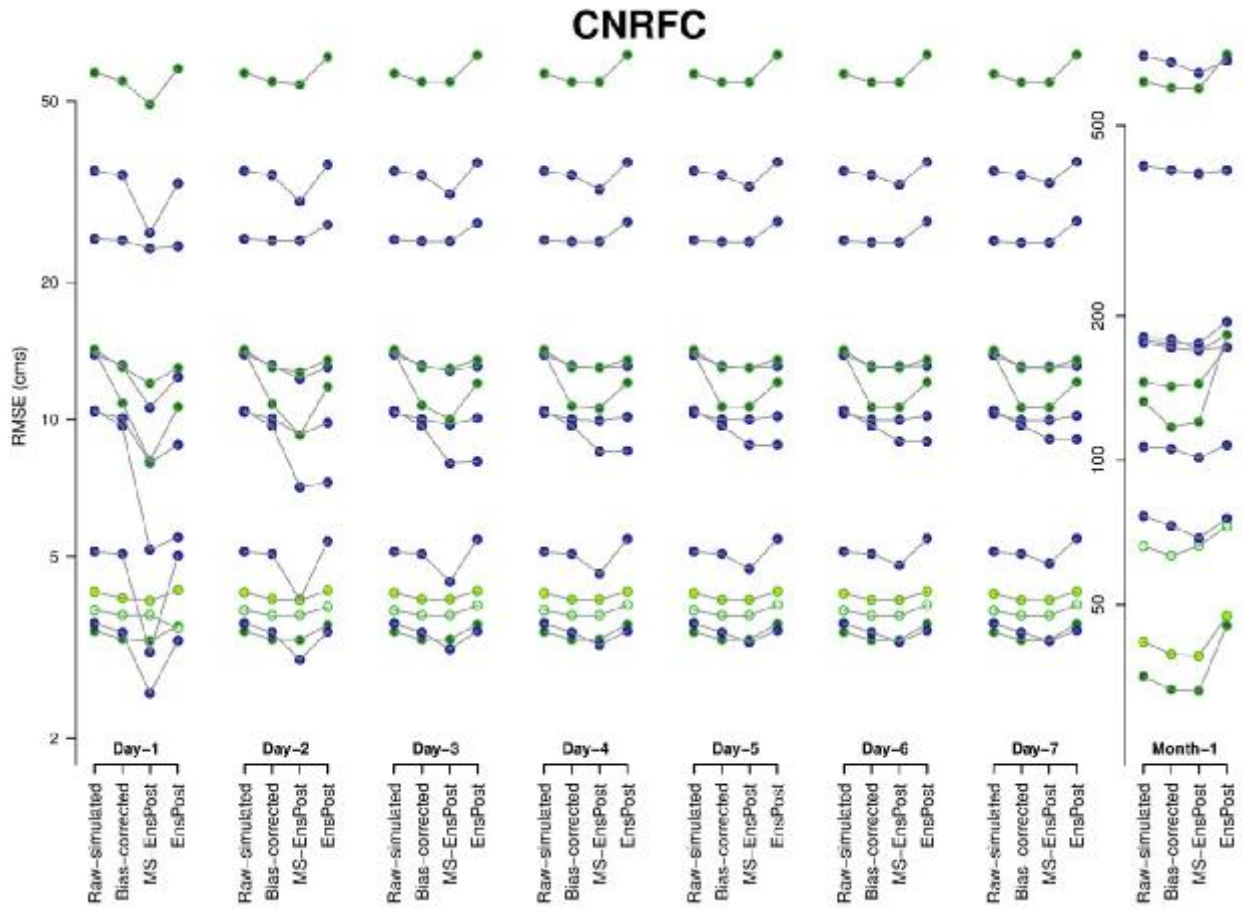


Figure 21: Same as Fig 20 but for the CNRFC basins (empty circles indicate basins with unmodeled regulated flow, green and blue outline indicate basins in coastal and Sierra Nevada mountain range, respectively).

For MARFC (Fig 22), bias correction generally contributes more to the RMSE reduction by MS-EnsPost than multiscale regression. The largest improvement by MS-EnsPost over EnsPost is in the Day-1 prediction for RTDP1 (third from the top) which is downstream of Raystown Dam on the Raystown Branch of the Juniata River. Overall, the impact of multiscale regression is rather modest and wears off within the first two days of lead time, an indication that the hydrologic memory in the Juniata River Basin in PA is relatively short.

For MBRFC (Fig 23), bias correction has significantly larger impact than multiscale regression for most basins. Visual examination of the scatter plots of the raw model predictions vs. verifying observations confirm that relatively large errors exist in the raw model predictions. Additional research is needed for its attribution.

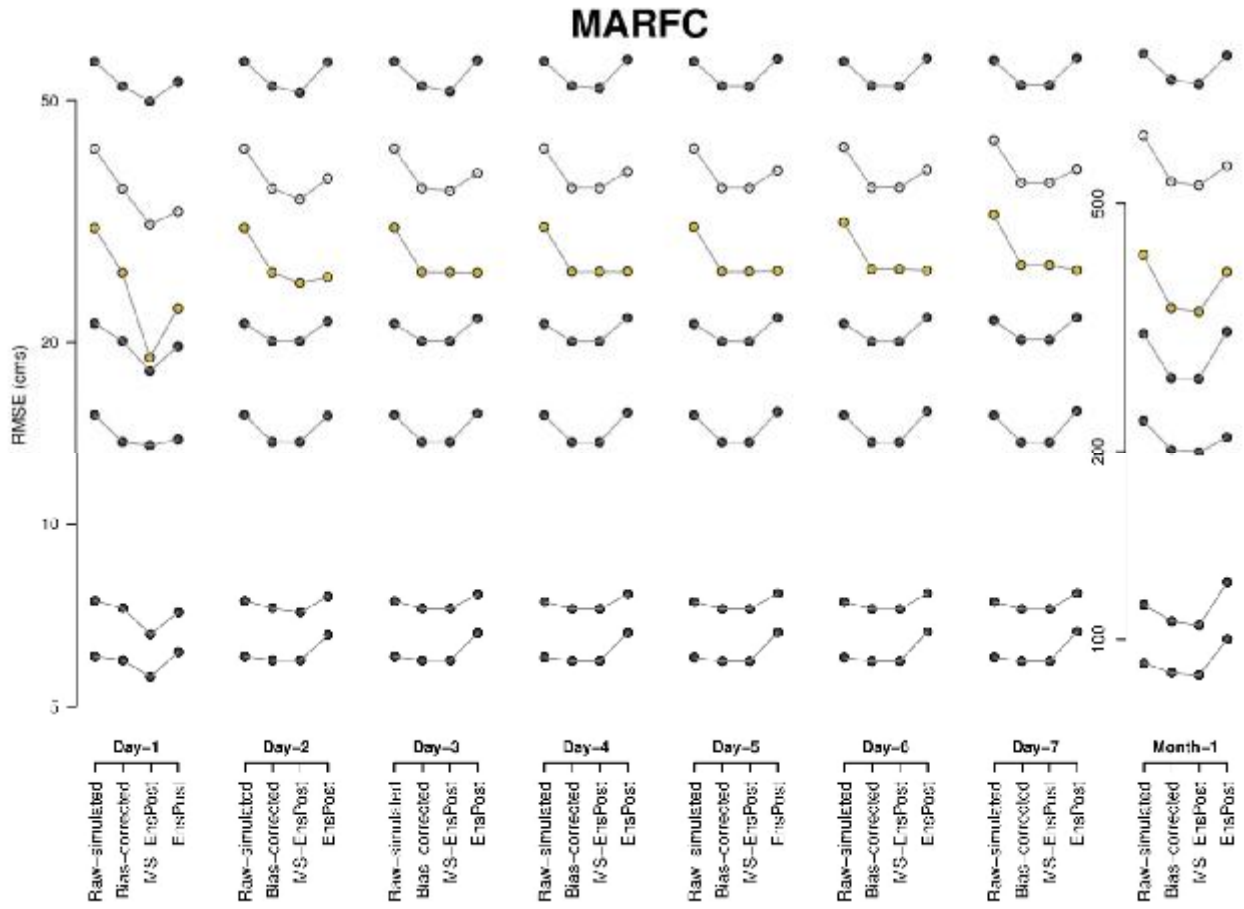


Figure 22: Same as Fig 20 but for the MARFC basins.

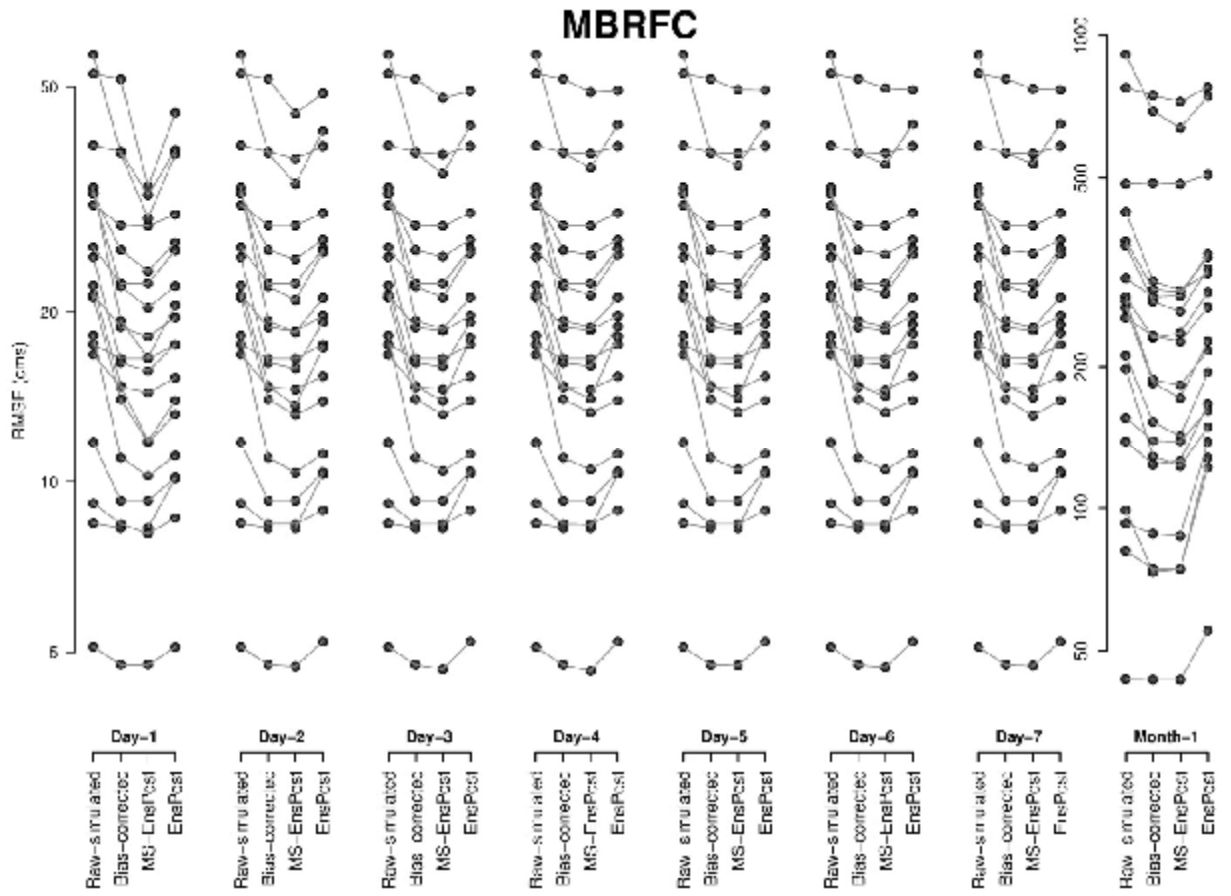


Figure 23: Same as Fig 20 but for the MBRFC basins.

For NCRFC (Fig 24), both bias correction and multiscale regression in MS-EnsPost significantly reduce the RMSE of raw model predictions for most basins. For a number of basins, the impact of multiscale regression is very large at short lead times due probably to the increased hydrologic uncertainty from ice jams, back water effects, frozen ground, agricultural diversions, and breakout flows that are common in this region.

For NERFC (Fig 25), both bias correction and multiscale regression in MS-EnsPost contribute to RMSE reduction. The impact of multiscale regression, however, is relatively short-lived. For a number of the NERFC basins, the impact of bias correction is relatively large. The scatter plots of raw model prediction vs. verifying observation indicate that significant

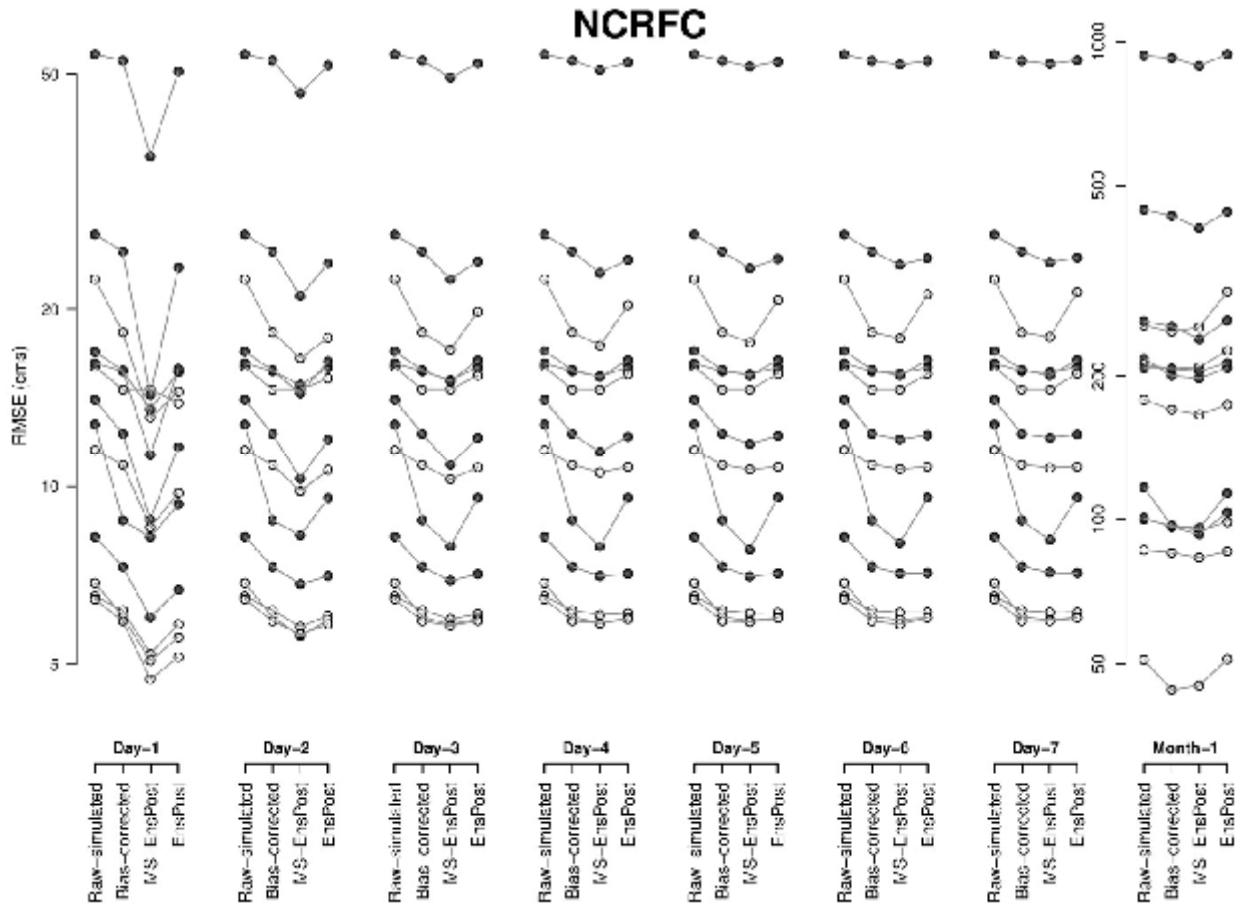


Figure 24: Same as Fig 20 but for the NCRFC basins.

magnitude-dependent biases exists for the above basins which MS-EnsPost is able to address effectively. The longer-lead time results for the NERFC basins indicate that the magnitude-dependent bias correction of MS-EnsPost outperforms the PM-based bias correction of EnsPost for most basins.

For NWRFC (Fig 26), bias correction is particularly effective for a number of flow-regulated basins that exhibit strong magnitude-dependent biases. In Fig 26, dots with green and blue and red outline indicate the basins in coastal and Cascade mountain and intermountain range, respectively. As with the CNRFC basins, the margin of improvement by MS-EnsPost is smaller for coastal basins. The longer-lead time results indicate that, as with the NERFC basins,

the magnitude-dependent bias correction of MS-EnsPost outperforms the PM-based bias correction of EnsPost for most of the NWRFC basins.

For the WGRFC basins (Fig 27), the improvement by MS-EnsPost over EnsPost is particularly large. These basins are located in the semi-arid western part of the Upper Trinity River Basin (Kim et al. 2018). As such, they have short memory in surface and soil water storages, and their streams are ephemeral despite relatively large basin size (441~1,764 km^2). Because EnsPost does not model intermittency of streamflow, its results are particularly poor for the WGRFC basins. MS-EnsPost, on the other hand, is able to address intermittency to a significant extent by aggregating flow which reduces or removes zero flows at sufficiently large

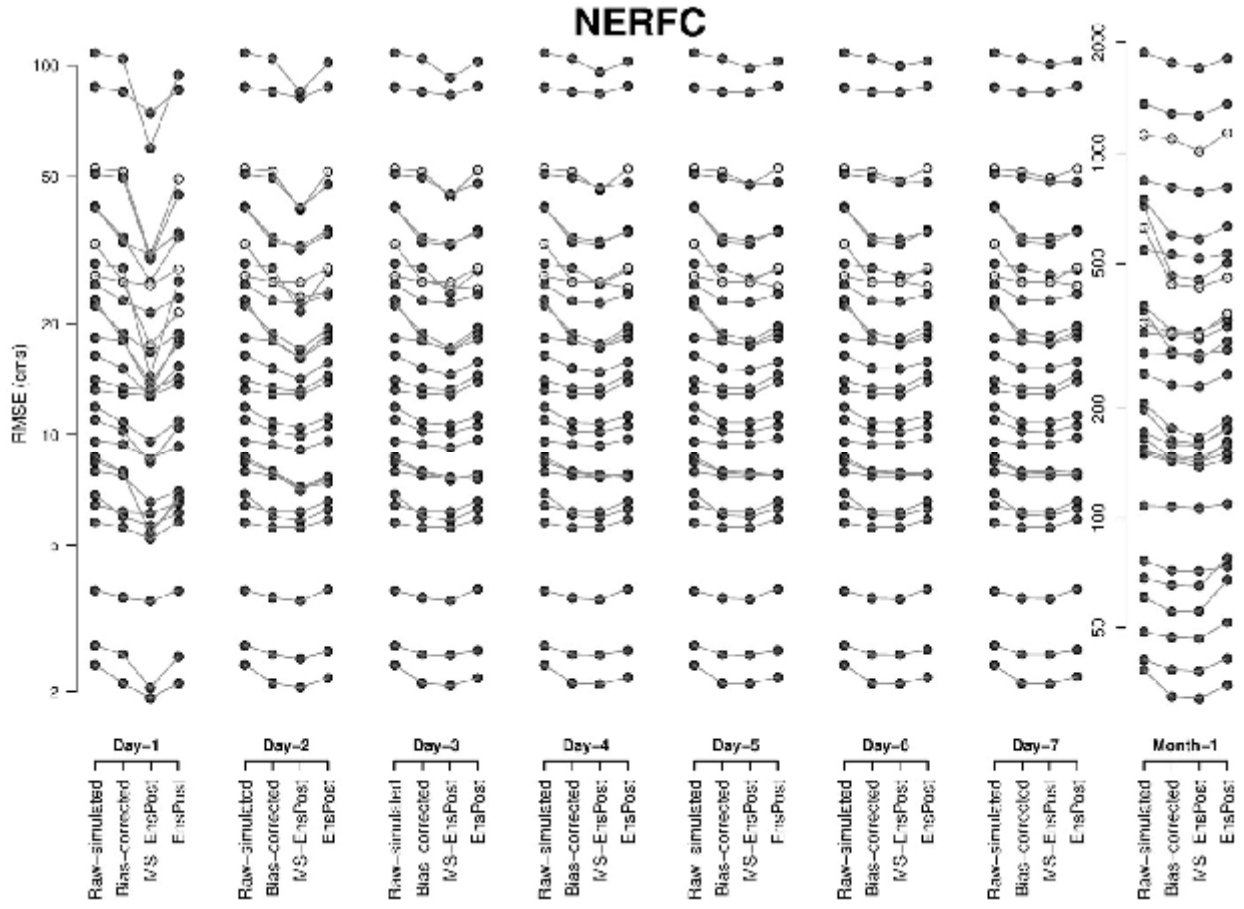


Figure 25: Same as Fig 20 but for the NERFC basins.

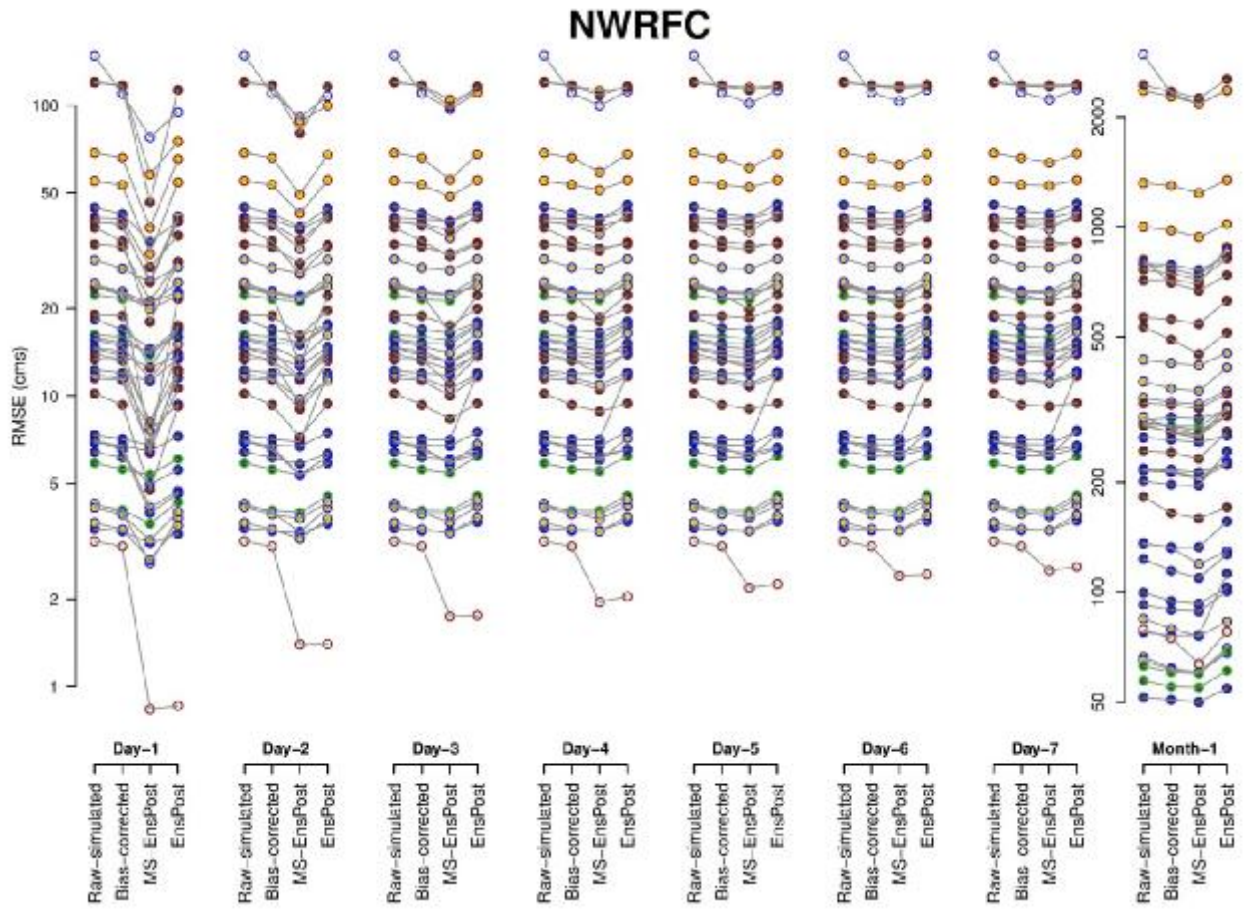


Figure 26: Same as Fig 20 but for the NWRFC basins.

temporal scales. Overall, the reduction in RMSE due to multiscale regression is rather short-lived. The monthly results (right-most panel in Fig 27) for JAKT2 and DCJT2 (2nd and 4th from the top, respectively) are unexpected in that multiscale regression in MS-EnsPost slightly increased RMSE over magnitude-dependent bias correction alone. The above observation indicates that statistical assimilation of observed streamflow up to a month in aggregation scale does not add skill due to the weak hydrologic memory in streamflow in these basins.

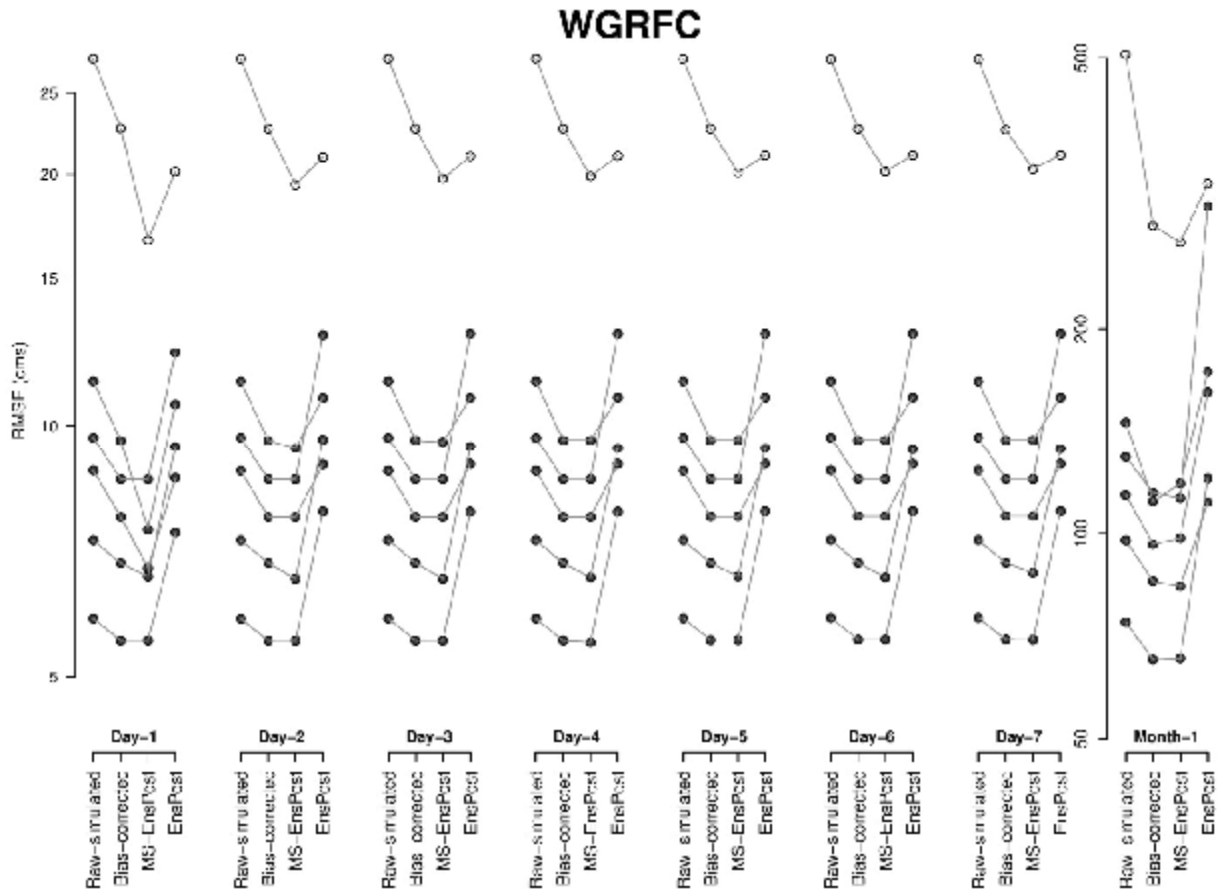


Figure 27: Same as Fig 20 but for the WGRFC basins.

6.2.2 Ensemble streamflow prediction

In this subsection, the MS-EnsPost ensemble streamflow predictions with the EnsPost are comparatively evaluated. To facilitate comparison for a large number of basins, I use “worm” plots in which the mean CRPS of the MS-EnsPost predictions (y-axis) vs. the EnsPost predictions (x-axis) are dot-plotted and connected for lead times of 1 to 7 days to form a “worm” for each basin. Fig 28 shows the worm plots in log-log scale for all study basins for each RFC. The lower and upper ends of each worm are associated with Day-1 and -7 predictions for that basin, respectively. If MS-EnsPost improves over EnsPost for 7 day-ahead prediction, the worms would stretch downward from the diagonal. The longer the downward stretch, the larger the

improvement by MS-EnsPost over EnsPost. If MS-EnsPost does not improve over EnsPost, the worms would lie along the diagonal. Fig 29 shows the mean CRPS scatter plots of 1 month-ahead MS-EnsPost predictions of monthly flow vs. the EnsPost.

Fig 28 shows that, for most basins, MS-EnsPost significantly improves over EnsPost. For most MARFC basins, however, little improvement is seen. For some MBRFC basin, MS-EnsPost performed worse than EnsPost for Day-1 and -2 predictions. For RTDP1 of MARFC (the 3rd worm from the top), MS-EnsPost clearly improves over EnsPost. Recall in the single-valued prediction results that MS-EnsPost generally showed significant improvement over EnsPost for regulated flows. The MBRFC basins results were unexpected in that MS-EnsPost was clearly superior to EnsPost in ensemble mean prediction. A closer examination indicates that, for the four MBRFC basins in question, the EnsPost ensemble predictions are superior to the MS-EnsPost only for the first one or two days of lead time, and that, for longer lead times, the MS-EnsPost predictions are superior. The effect of the above comparative performance may be seen in Fig 29 where the 1 month-ahead MS-EnsPost predictions of monthly flow is clearly superior to the EnsPost for all MBRFC basins. Note also in Fig 29 that the improvement by MS-EnsPost over EnsPost is larger for the smaller MBRFC basins. This is due to the fact that bias correction, rather than multiscale regression, is largely responsible for the improvement by MS-EnsPost which produces a large positive cumulative impact for prediction of monthly flow.

Overall, the CB-, NC- and NWRFC basins show particularly large improvement by MS-EnsPost over EnsPost. It was summarized in the single-valued prediction results that, for CB- and NWRFC basins, the reduction in RMSE was due more to multiscale regression than bias correction. For many basins in these two RFCs, streamflow is fed by snowmelt which increases hydrologic memory. The CB- and NWRFC results indicate that multiscale regression

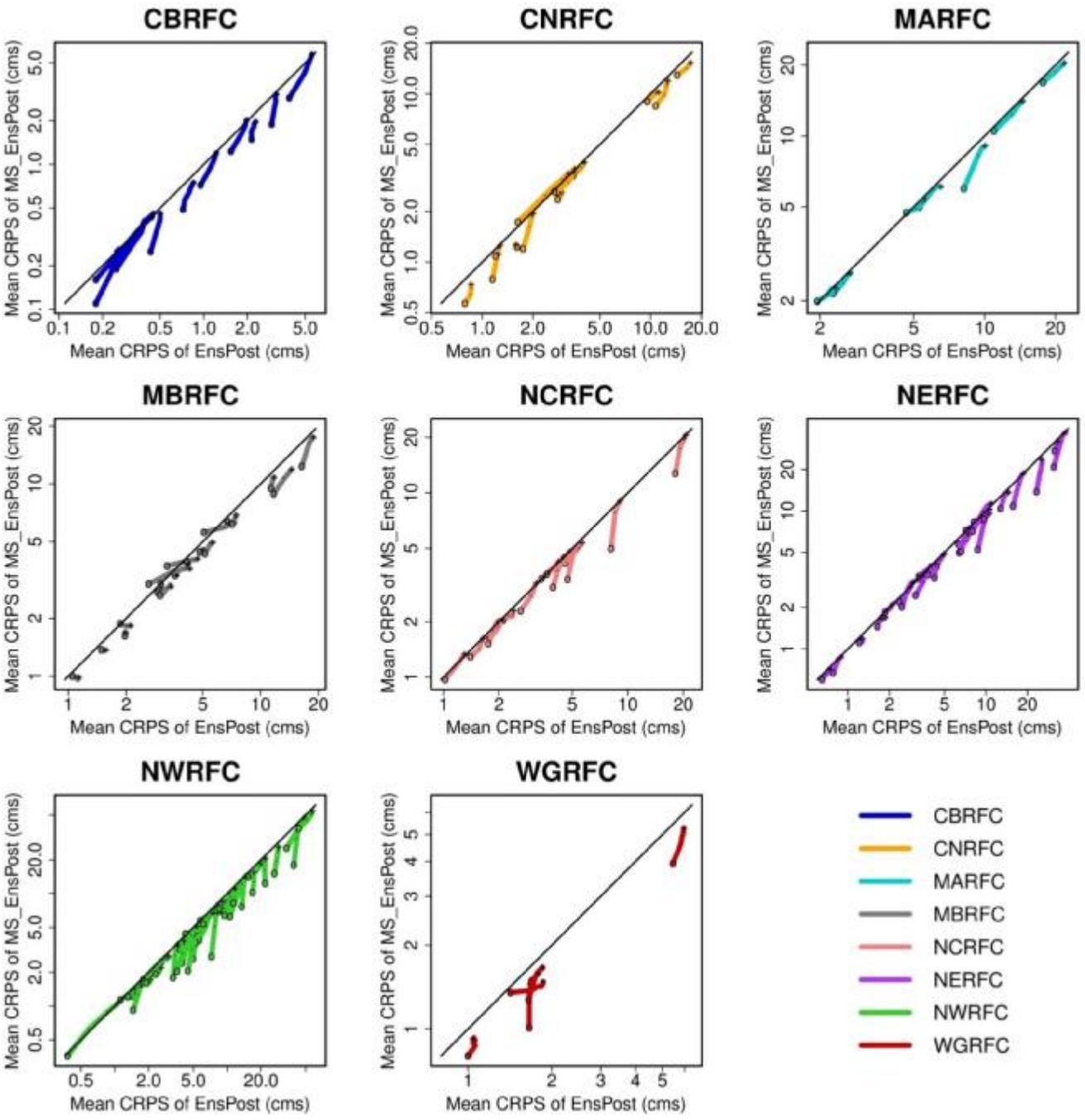


Figure 28: Worm plots (see text for explanation) of mean CRPS of ensemble predictions of daily flow from MS-EnsPost and EnsPost for lead times of 1 to 7 days.

in MS-EnsPost is able to utilize effectively the predictability present in the model-simulated and observed flows over a range of temporal scales of aggregation. For the NCRFC basins, on the other hand, the significant improvement by MS-EnsPost is due more to bias correction than

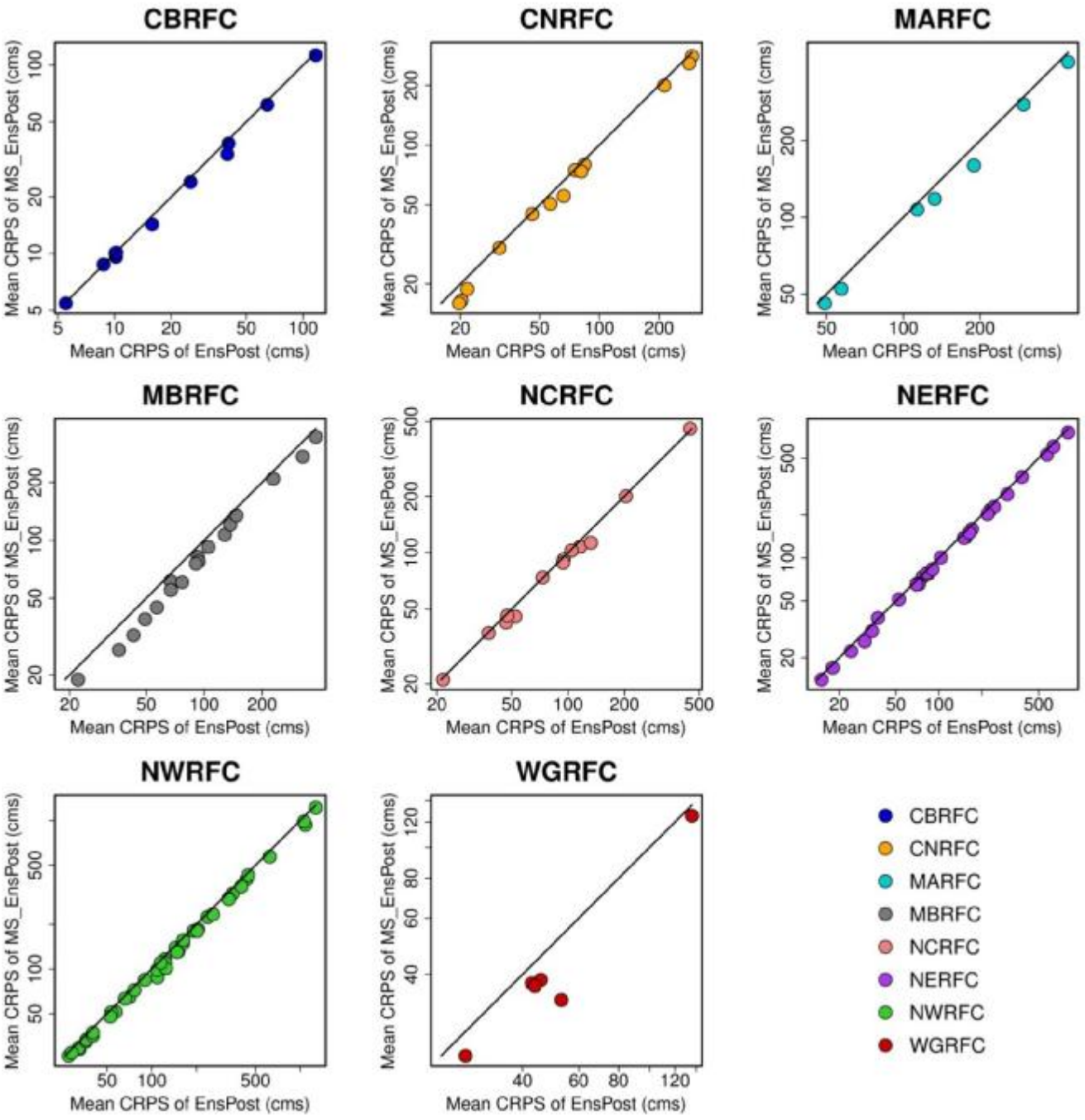


Figure 29: Same as Fig 28 but for 1 month-ahead predictions of monthly flow.

multiscale regression as explained in the single-valued prediction results. For the NERFC basins, MS-EnsPost shows significantly larger improvement over EnsPost for larger basins. For the CNRFC basins, MS-EnsPost significantly improves over EnsPost for some basins, while for the

others, MS-EnsPost and EnsPost perform similarly. As seen in the single-valued prediction results, the improvement is generally smaller for the coastal basins. For the WGRFC basins, MS-EnsPost significantly improves over EnsPost. It indicates that bias correction and multiscale regression are effective in addressing the flow magnitude-dependent biases in raw model-predicted flow and intermittency of streamflow in the semi-arid region.

Decomposition of the mean CRPS (see Eq.(21)) indicates that, for most basins, the reduction in mean CRPS by MS-EnsPost over EnsPost is due mostly to improved resolution, rather than improved reliability (See Appendix D for examples). This is not very surprising because the EnsPost uses empirical probability-matching based on NQT whereas MS-EnsPost relies on approximate distribution modeling via the Box-Cox transformation. If the historical record is long enough to model the tails of the distributions with accuracy, one may expect the ensemble traces sampled from the empirically-modeled distributions to be more reliable. To scrutinize reliability of MS-EnsPost ensemble predictions, also the reliability diagrams (Brown and Seo 2010; Jolliffe and Stephenson 2012; Wilks 2006) and Brier scores (1950) are examined for a wide range of thresholds. Figs 30 through 33 show comparative examples of Brier scores and reliability diagrams for EnsPost's and MS-EnsPost's daily predictions.

They indicate that the MS-EnsPost ensembles are generally as reliable as the EnsPost ensembles for the 90^{th} percentile or larger thresholds, but significantly less so for the 50^{th} percentiles or smaller thresholds. For flood and water supply forecasting, performance for larger flows is much more important than that for smaller flows. As such, deterioration in reliability at lower thresholds is not a large concern in most applications. The mean CRPS results for MS-EnsPost above indicate that, overall, the gain in resolution outweighs some loss in reliability in the low flow regime.

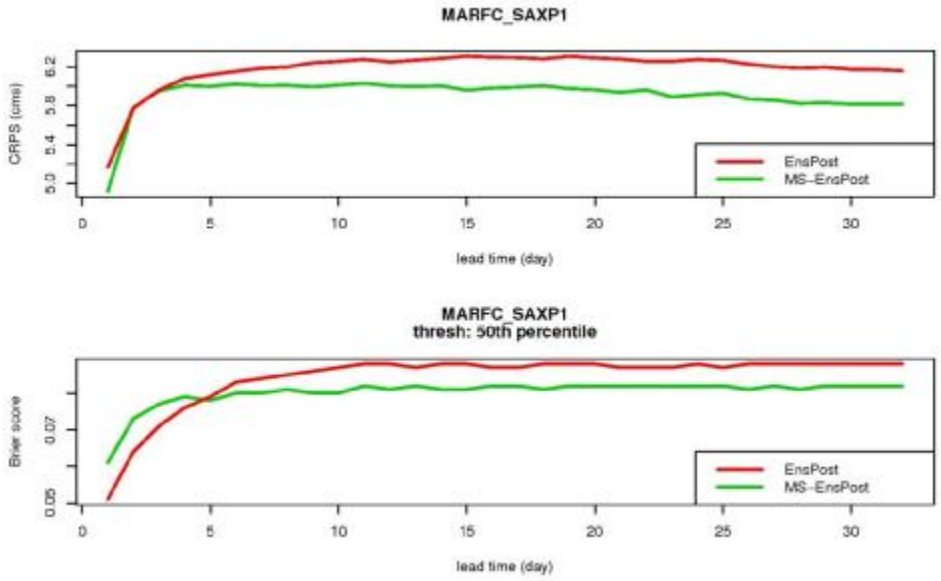


Figure 30: Comparison of CRPS vs. lead time (top) and Brier score (higher 50% of observed flow) vs. lead time (bottom) for SXTP1 in MARFC.

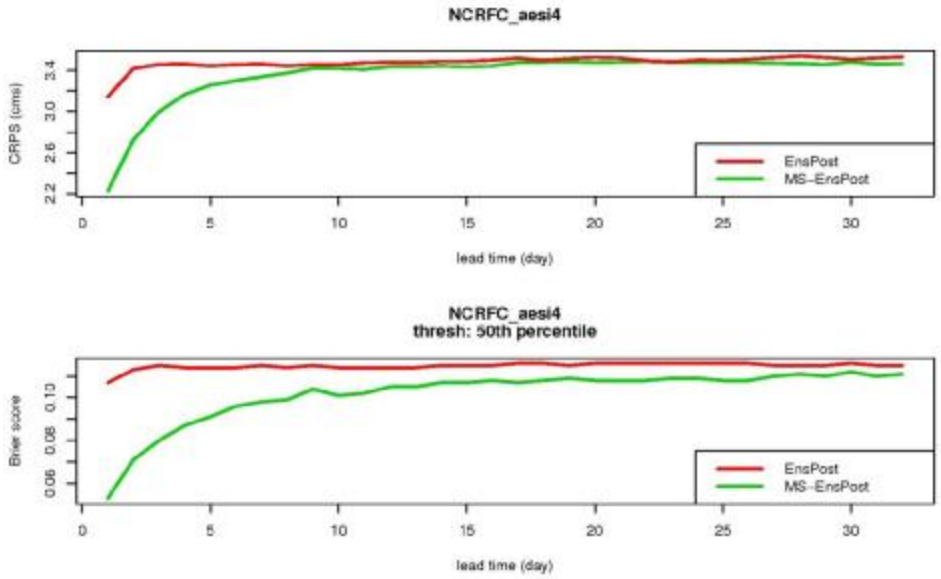


Figure 31: Comparison of CRPS vs. lead time (top) and Brier score (higher 50% of observed flow) vs. lead time (bottom) for AESI4 in NCRFC.

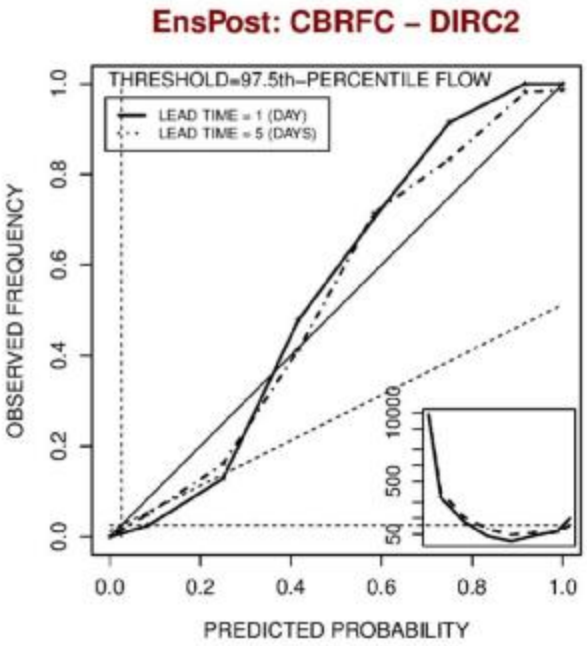


Figure 32: Reliability diagram from EnsPost (higher 2.5% of observed flow) for DIRC2 in Upper Colorado River Basin.

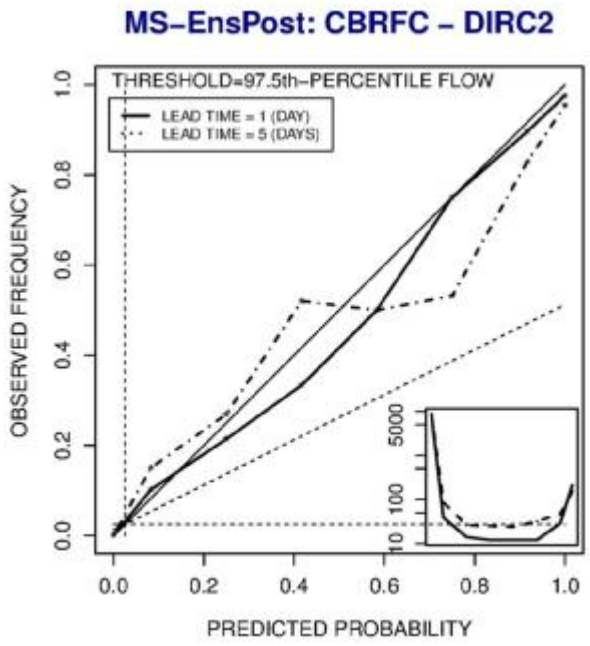


Figure 33: Reliability diagram from MS-EnsPost (higher 2.5% of observed flow) for DIRC2 in Upper Colorado River Basin.

6.2.3 Streamflow predictability

The skill in post-processed ensemble predictions is bounded by the predictability of streamflow explainable by the forcings (Baldwin et al. 2003; Bengtsso and Hodges 2006; Gebregiorgis and Hossain 2011; Li and Ding 2011; Simmons et al. 1995), hydrologic and reservoir models (Hou et al. 2009; Mahanama et al. 2012; Maurer and Lettenmaier 2004; Schlosser and Milly 2002), and statistical assimilation of streamflow via multiscale regression (Bogner et al. 2016; Sharma et al. 2018) used. In this subsection, the predictability of streamflow in different hydroclimatological regions is assessed and characterized based on the ensemble prediction results presented above, and the gains by MS-EnsPost over EnsPost are attributed by assessing the predictability through a skill score (Hou et al. 2009; Westra and Sharma 2010). Fig 34 shows the mean CRPSS of the MS-EnsPost ensemble predictions for all seasons for lead times of 1 to 32 days. The reference forecast is the sample climatology of historical observed flow. To assess seasonal variations, the wet-vs.-dry seasonal results were also examined. They showed that, except for the CBRFC basins, the mean CRPSS does not differ much between the two seasons. As such, only the combined results are presented which are necessarily more reflective of the wet season. For the CBRFC basins, the mean CRPSS is significantly lower for the dry season due to the fact that highly persistent low-flow conditions may be predicted very well with climatology. In Fig 34, the vertical spread in the mean CRPSS curves represents the variations in predictability of streamflow among the different basins within each RFC's service area. It is readily seen that the CBRFC basins, all of which are in the Upper Colorado River Basin, exhibit the smallest variations. The largest variations are observed with the NWRFC basins which encompass the coastal, mountain and intermountain regions of the Pacific Northwest.

For each RFC, there are a small number of basins with conspicuously lower mean CRPSS. They are generally associated with regulated flows which inflate hydrologic uncertainty. Because these basins do not represent natural flows, they are treated separately in the analysis below.

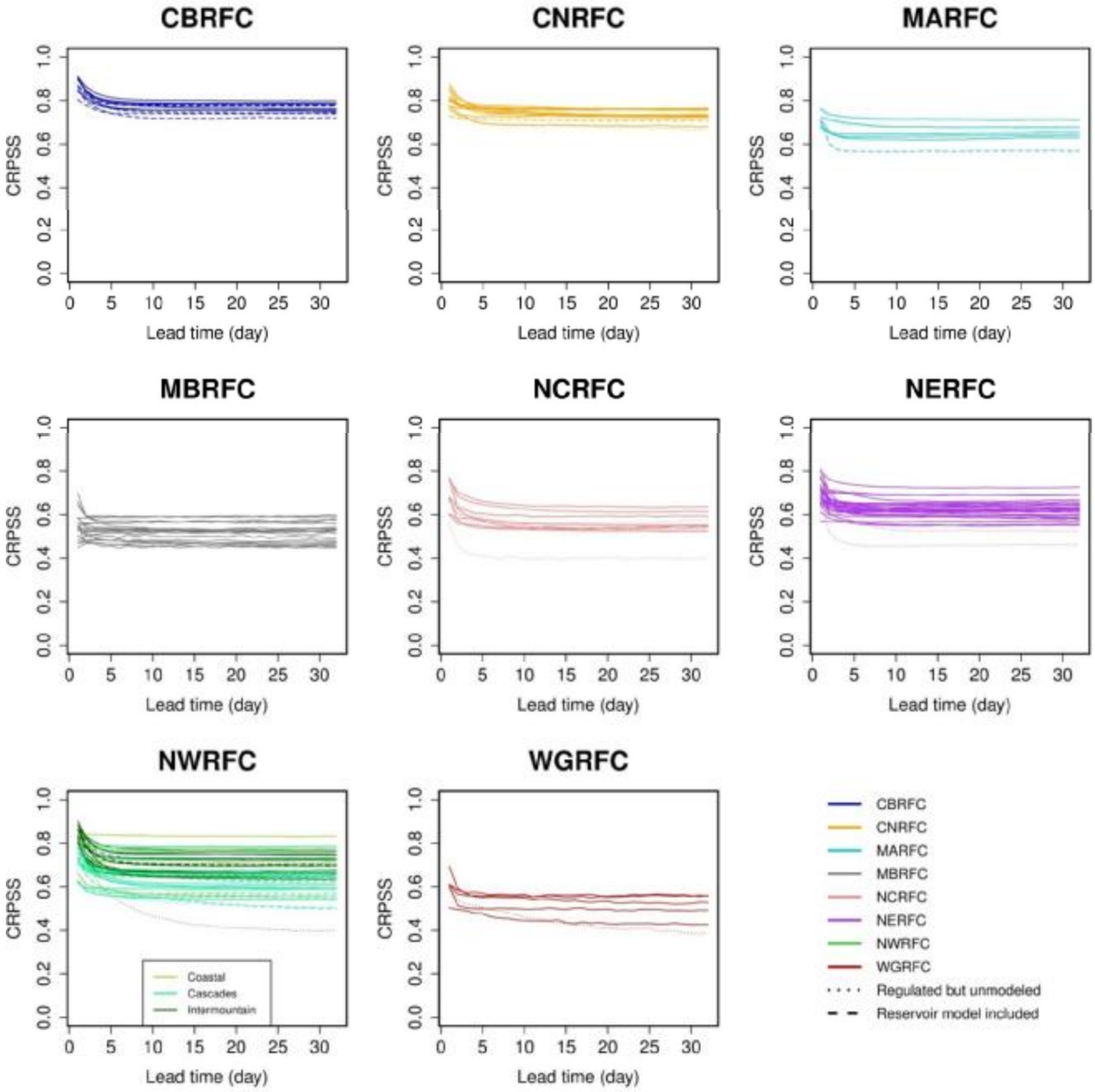


Figure 34: CRPSS of ensemble predictions of daily flow from MS-EnsPost vs. lead time. The reference is sample climatology of historical observed flow.

MS-EnsPost seeks two effects in the mean CRPSS results: an increase in the limiting mean CRPSS from bias correction at very large lead times, $\overline{CRPSS}(|\infty|)$, and an increase in mean CRPSS from multiscale regression at shorter lead times above $\overline{CRPSS}(|\infty|)$. The first and second attributes above are referred to herein as the limiting mean CRPSS and the hydrologic memory scale (Schlosser and Milly 2002), respectively. The larger the limiting CRPSS, the more skillful the bias-corrected ensemble prediction relative to climatology. The larger the hydrologic memory scale, the larger the increase in mean CRPSS due to multiscale regression. The hydrologic memory scale, L_{hm} (days), which represents the predictability of streamflow due to the surface and soil water storages in the basin (Kumar 2011), is defined as:

$$L_{hm} = \int_0^\infty \rho_{\overline{CRPSS}}(|\tau|)d\tau \tag{24}$$

where $\rho_{\overline{CRPSS}}(|\tau|)$ denotes the normalized mean CRPSS at lead time τ (days). The normalization renders mean CRPSS to approach zero at large lead times. One may hence consider $\rho_{\overline{CRPSS}}(|\tau|)$ as correlogram with nugget effect (Norouzi et al. 2018). To illustrate, Fig 35 shows the two attributes in CRPSS.

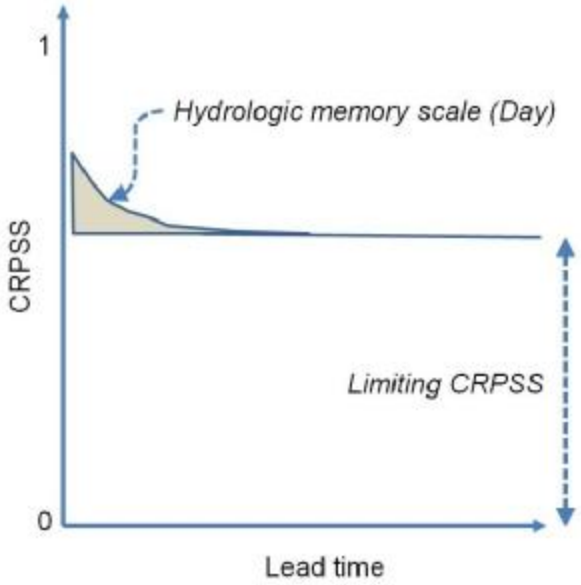


Figure 35: Attribution of changes in CRPSS.

Fig 36 shows the resulting pairs of L_{hm} and $\overline{CRPSS}(|\infty|)$ for all basins in each RFC as obtained from the MS-EnsPost and EnsPost predictions. For each basin, an arrow connects the EnsPost result to the matching MS-EnsPost result. If MS-EnsPost increases the limiting mean CRPSS, the arrow would point upward. If MS-EnsPost increases the hydrologic memory scale, the arrow would point to the right. The longer the arrow is, the larger the improvement or deterioration is. Accordingly, lengthy arrows pointing in the upper-right direction would indicate MS-EnsPost clearly improving over EnsPost. It is seen that, for a number of basins, MS-EnsPost improves limiting mean CRPSS but reduces the hydrologic memory scale, resulting in arrows pointing in the upper-left direction. Examination of the mean CRPS results indicates that, for these basins, the mean CRPS approaches the limiting values very slowly due to very highly correlated errors, thereby artificially inflating the hydrologic memory scale. Accordingly, one may consider MS-EnsPost inferior to EnsPost only if the arrow is pointing in the lower-left direction.

Fig 36 shows that MS-EnsPost outperforms or comparable to EnsPost for all basins, increases limiting CRPSS for almost all basins, and provides significant additional skill via multiscale regression particularly for the CB-, CN-, NC-, and NWRFC basins. From Fig 36, a number of postulations may also be made. The significant increase in L_{hm} by MS-EnsPost for the CB-, CN-, NC-, and NWRFC basins suggests that there exists significant multiscale hydrologic memory to be exploited for operational hydrologic forecasting via data assimilation. The significant increase in $\overline{CRPSS}(|\infty|)$ by MS-EnsPost for the MB- and NERFC basins suggests that there may exist significant room for improving calibration, hydrologic modeling, and input forcings to reduce hydrologic uncertainties. The WGRFC basin results, on the other hand, suggest limited room for improving predictive skill within the existing modeling and

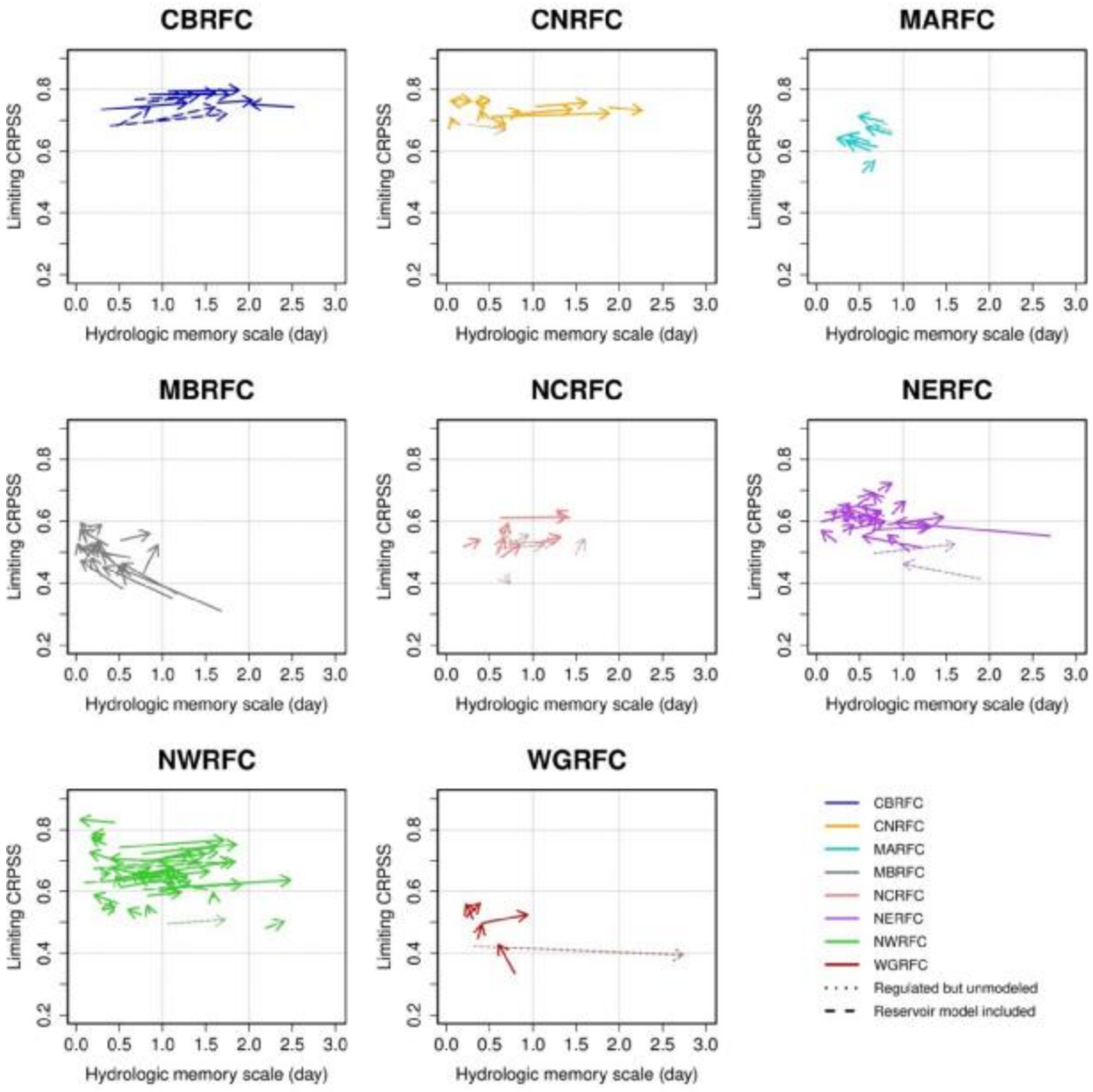


Figure 36: Changes in limiting CRPSS and hydrologic memory scale from those of EnsPost to those of MS-EnsPost (see text for explanation).

forecasting process, and point to improving model physics as well as soil moisture sensing and its assimilation.

The relative importance of $\overline{CRPSS}(|\infty|)$ vs. L_{hm} in assessing predictability necessarily varies with the application at hand. For long-range predictions, $\overline{CRPSS}(|\infty|)$ would be more

important whereas L_{hm} may be just as important for short-range predictions. Hence, it is not readily possible to translate uniquely the two summary attributes into a single measure. One may consider, however, the relative positions of the (L_{hm}, $\overline{CRPSS}(|\infty|)$) pairs for MS-EnsPost (i.e., the tips of the arrows) within the xy-plot in Fig 36, and approximately rank the groups of basins in different RFCs in terms of the collective strength of predictability as measured through MS-EnsPost. The figure indicates that the CB-, CN-, NWRFC basins are the most predictable, followed by the NE-, MA-, and NCRFC basins, and that the MB- and WGRFC basins are the least predictable. The above order reflects what may be garnered visually from Fig 34, and generally follows the decreasing order of the fraction of precipitation as snow, f_s, and mean annual precipitation (see Fig 11 through Fig 13). To illustrate, Fig 37 shows $\overline{CRPSS}(|\infty|)$ vs.

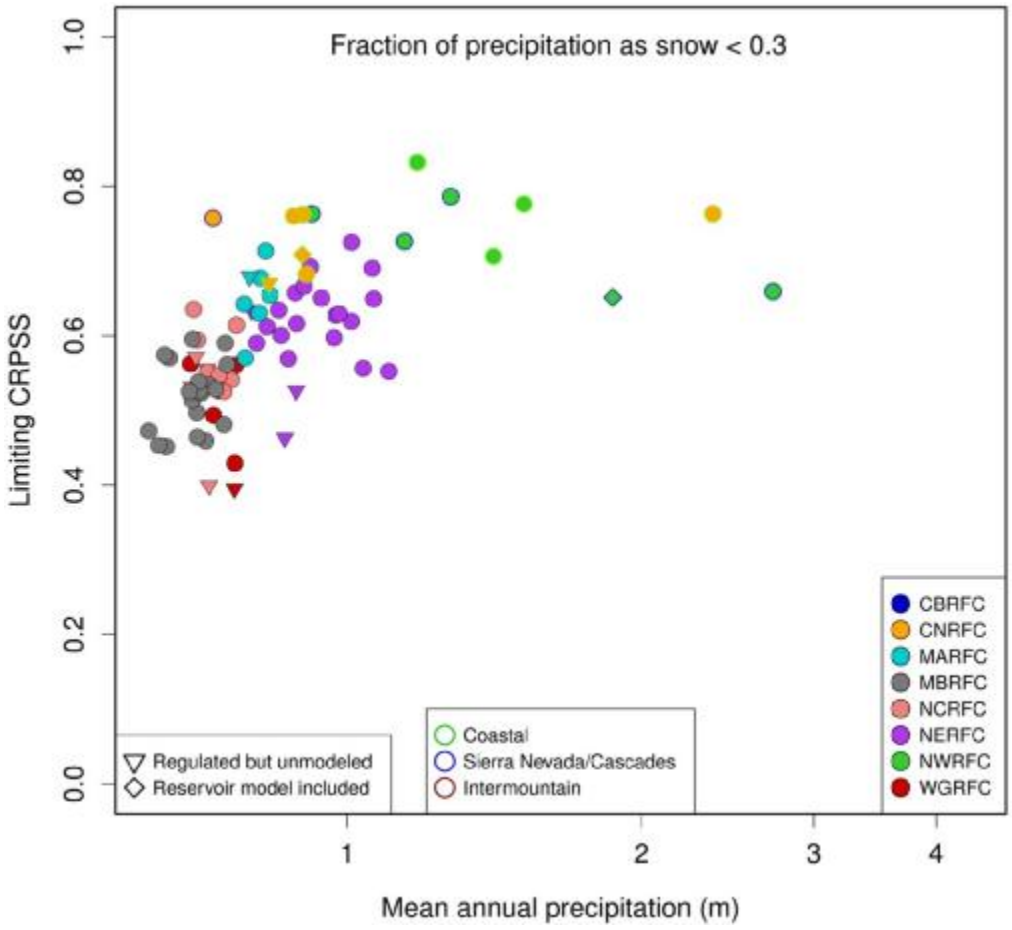


Figure 37: Limiting CRPSS vs. mean annual precipitation for non-snow-driven basins.

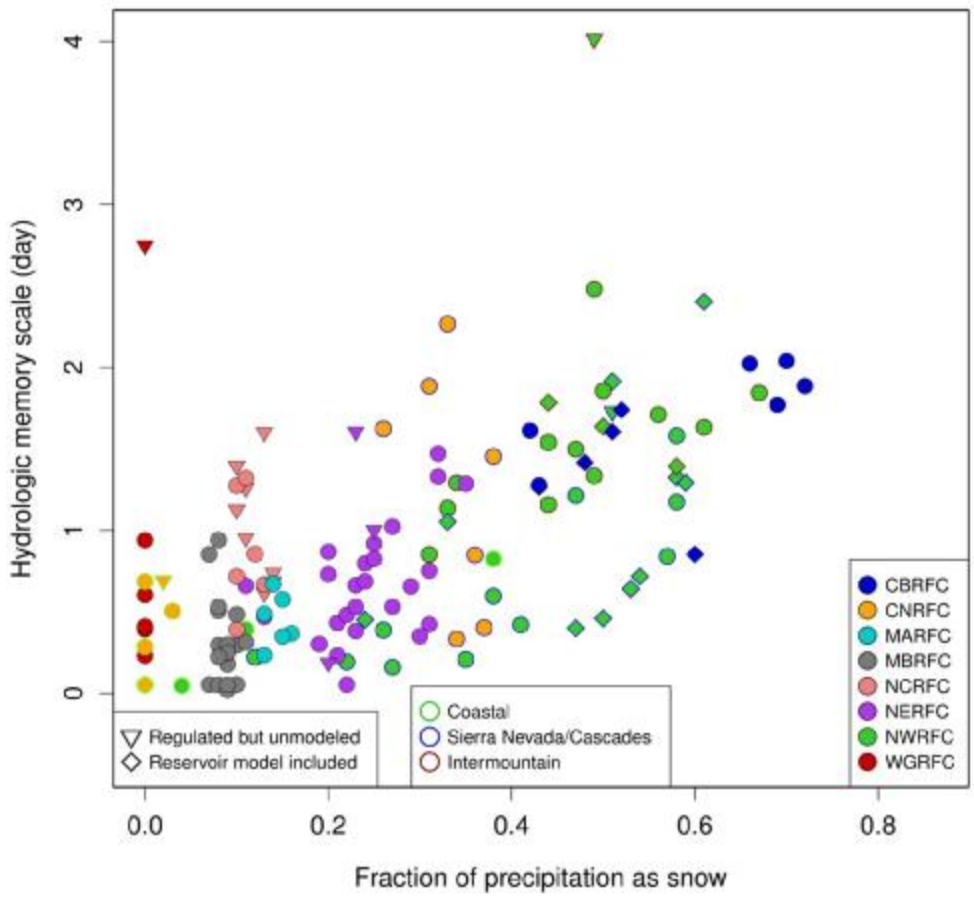


Figure 38: hydrologic memory scale vs. fraction of precipitation as snow.

annual mean precipitation for basins with f_s < 0.3 and Fig 38 shows L_{hm} vs. f_s for all basins (Please see Appendix E for more comparison plots).

Though the scatters are large, $\overline{CRPSS}(|\infty|)$ for non-snow-dominated basins relates well with mean annual precipitation except for the few very wet coastal basins, and L_{hm} relate positively with f_s for all basins.

6.2.4 Analysis of multiscale regression weights

MS-EnsPost uses multiscale regression to assimilate statistically streamflow observations aggregated at different time scales. Because the weight calculated for the observed flow reflects

the strength of memory of the basins for one timestep-ahead prediction of streamflow, the weights succinctly characterize the scale-dependent memory. For interpretation of the weights, it is useful to consider the method-of-moment estimate for the weight associated with the observed flow, ϖ_k, as described in Seo et al. (2006) even though Fisher estimation is used in MS-EnsPost as described in Subsection 3.2.2:

$$\varpi_k = \frac{1+\rho_{o,k}(|1|)-\rho_{c,k}(|1|)-\rho_{c,k}(|0|)}{2\left(1-\rho_{c,k}(|1|)\right)} \quad (25)$$

In the above, $\rho_{o,k}(|l|)$ and $\rho_{c,k}(|l|)$ denote the serial correlation at lag l of observed flow aggregated at the k-th time scale, $a^o_{k,0}$, and the cross correlation at lag l between $a^o_{k,0}$ and the bias-corrected model-simulated flow aggregated at the k-th time scale, $a^b_{k,1}$, respectively. Eq.(25) indicates that, the stronger the persistence in the observed flow is, the larger the weight for $a^o_{k,0}$ is, and that, the more skillful the model-simulated flow is, the smaller the weight is. Figs 39 through 45 and Fig 48 show the scale-dependent weight for all basins as grouped by RFCs. Though the weights are connected across different temporal scales of aggregation for basin identification purposes, they are not to be seen as a form of serial correlation. The main observations may be summarized as follows for each RFC.

CBRFC - The weight curves are very similar for most basins. Those for BSWC2, GBYC2 and WCRC2 show the largest differences. The BSWC2 weights indicate that the observed flow is the most persistent and the simulated flow is the most skillful in this group. GBYC2 is influenced by Granby Dam which has a large maximum storage of 539,800 acre feet. The weights for GBYC2 indicate that flow regulation reduces persistence in observed flow at all scales of aggregation at this location. WCRC2 is influenced by Willow Creek Dam which has a significantly smaller maximum storage of 11,177 acre feet. Note that the weight decreases the fastest for this basin as the temporal scale of aggregation increases, an indication that the model-

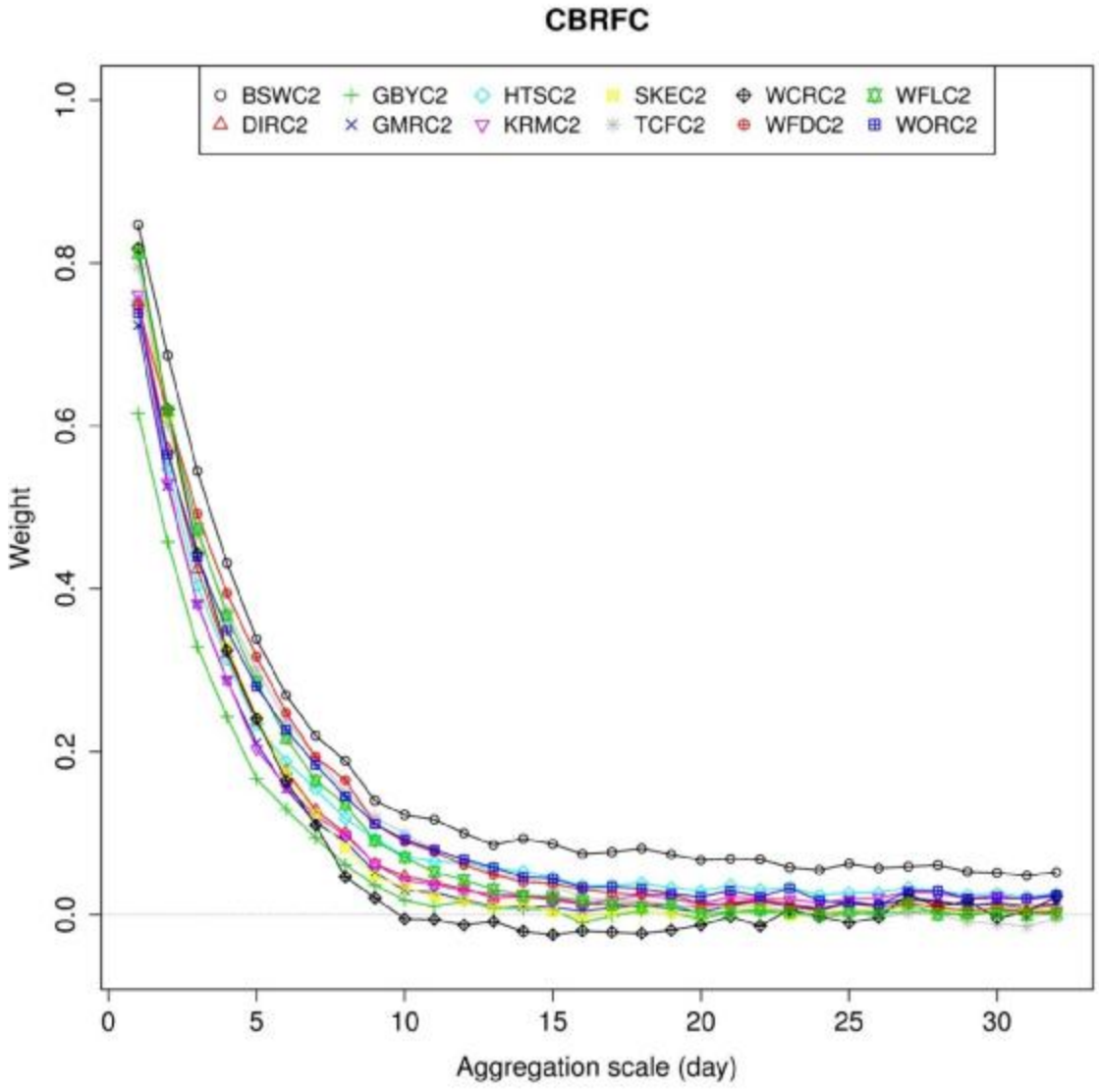


Figure 39: Regression weights vs. aggregation scale for basins in CBRFC.

simulated flow loses skill rather quickly as the aggregation scale increases. DIRC2 and WFDC2 are regulated by Dillon Reservoir and Williams Fork Reservoir which have maximum stroages of 250,000 and 97,000 acre feet, respectively. The weights for these two basins, however, decrease at a slower rate than that for WCRC2.

CNRFC – Unlike the CBRFC basins, the weight curves vary significantly among the basins. Of the 13 basins, only 4 basins show significant memory. SHEC1, LAMC1 and LAMC0

indicate not only very weak memory but also little dependence on aggregation scale. LAMC1 and CEGC1 are impacted by Lake Mendocino and Trinity Lake, respectively, both of which show generally smaller weights at all temporal scales aggregation. It is not clear what the source or sources of reduced memory and skill in model simulation may be for SHEC1. The weight curve suggests that significant unmodeled movement and control of water may exist in this basin. Additional research is needed to ascertain the above.

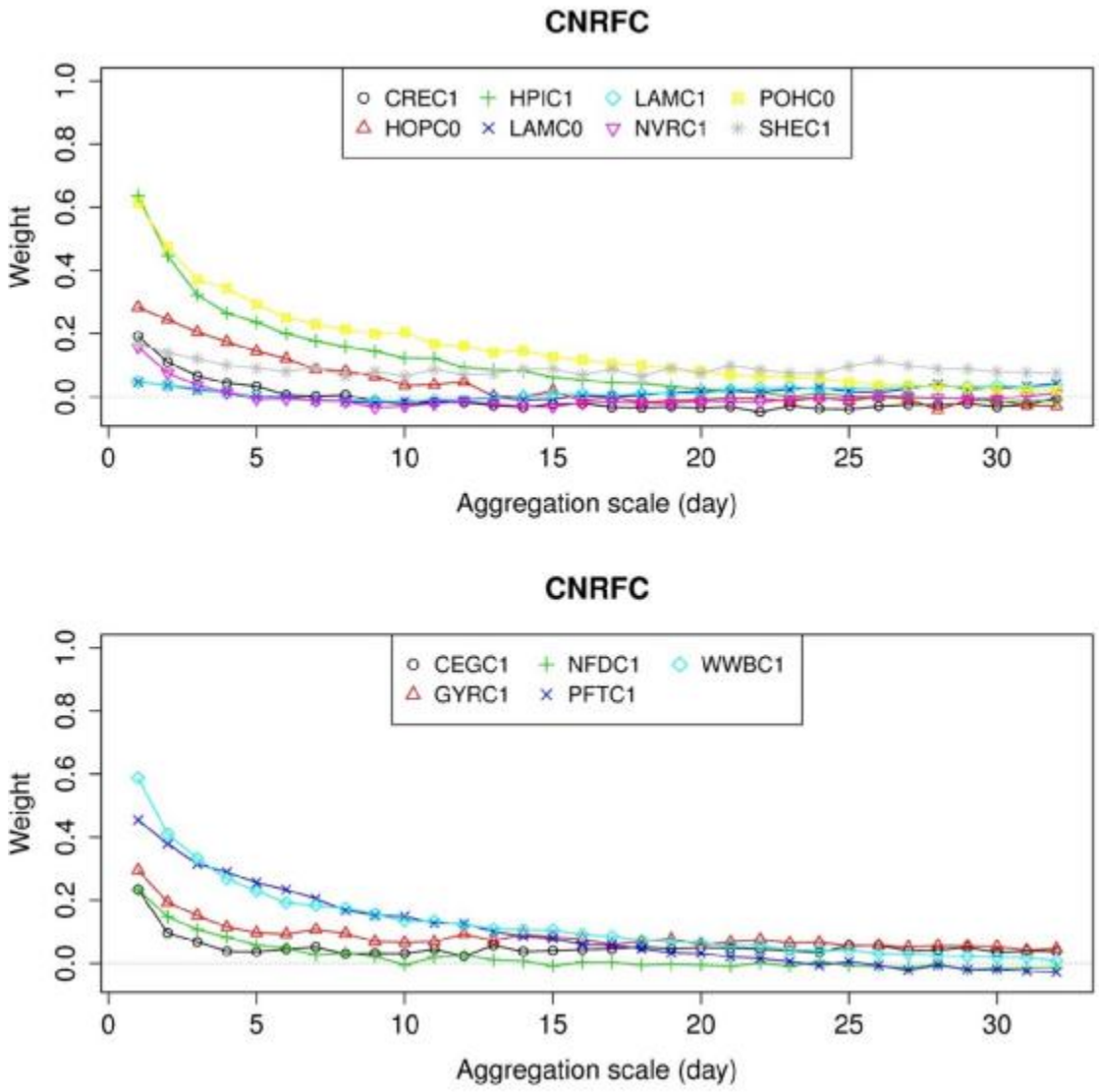


Figure 40: Regression weights vs. aggregation scale for basins in CNRFC.

MARFC – The weight curves show that RTDP1, which is influenced by Raytown Dam, has by far the largest memory in this group. Recall that MS-EnsPost produced the largest improvement over EnsPost for RTDP1 in this group. The weight curves stongly support that MS-EnsPost was able to utilize the strong memory present in this basin. Though all basins in this group are located in the Juniata River Basin, their weight curves exhibit significant variations. The magnitude of hydrologic memory, however, is too small for statistical assimilation to be very effective except for RTDP1.

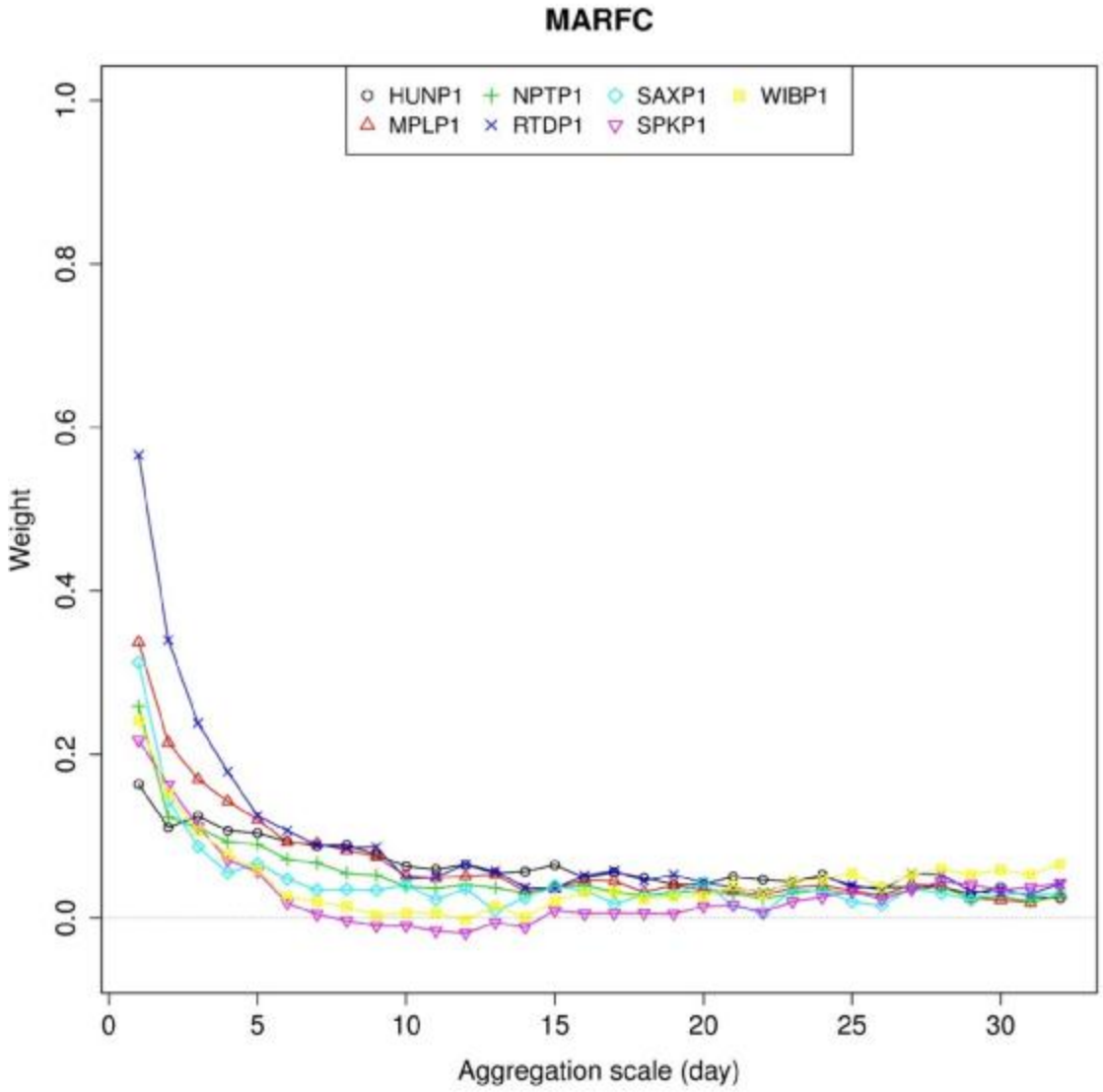


Figure 41: Regression weights vs. aggregation scale for basins in MARFC.

MBRFC – The basins in this group show relative weak hydrologic memory except HMBI4 and SSTM4 the latter of which is influenced by Smithville Reservoir. Additional research is needed to identity the source or sources of strong memory for HMBI4. It is worth noting that several basins in this group indicate little to no memory. They hence present a challenge for statistical post processing as seen in the EnsPost and MS-EnsPost results. HBLN1, in particular, shows negative weights for the smallest aggregation scales, an indication that observed flow at the basin outlet provides little information about the state of the basin. It is

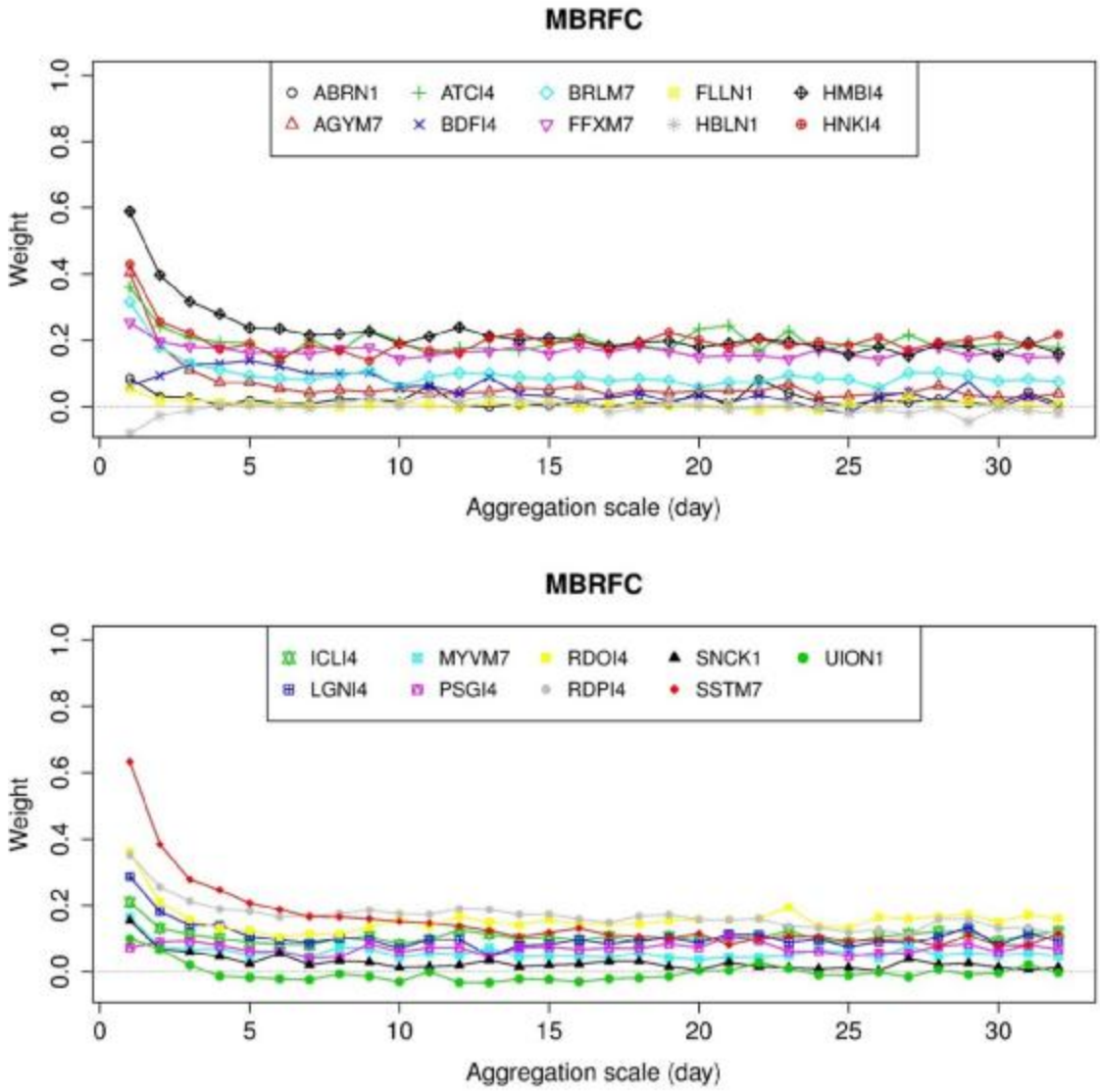


Figure 42: Regression weights vs. aggregation scale for basins in MBRFC.

suspected that multiple sources of hydrologic uncertainty contribute to the apparently lack of hydrologic memory, including ice jams, back water effects, frozen ground, agricultural diversions, and breakout flows which are common in this region.

NCRFC – A number of basins in this group show relatively strong hydrologic memory with OOIA4 and SIGI4 representing the strongest. These two locations are on the South and

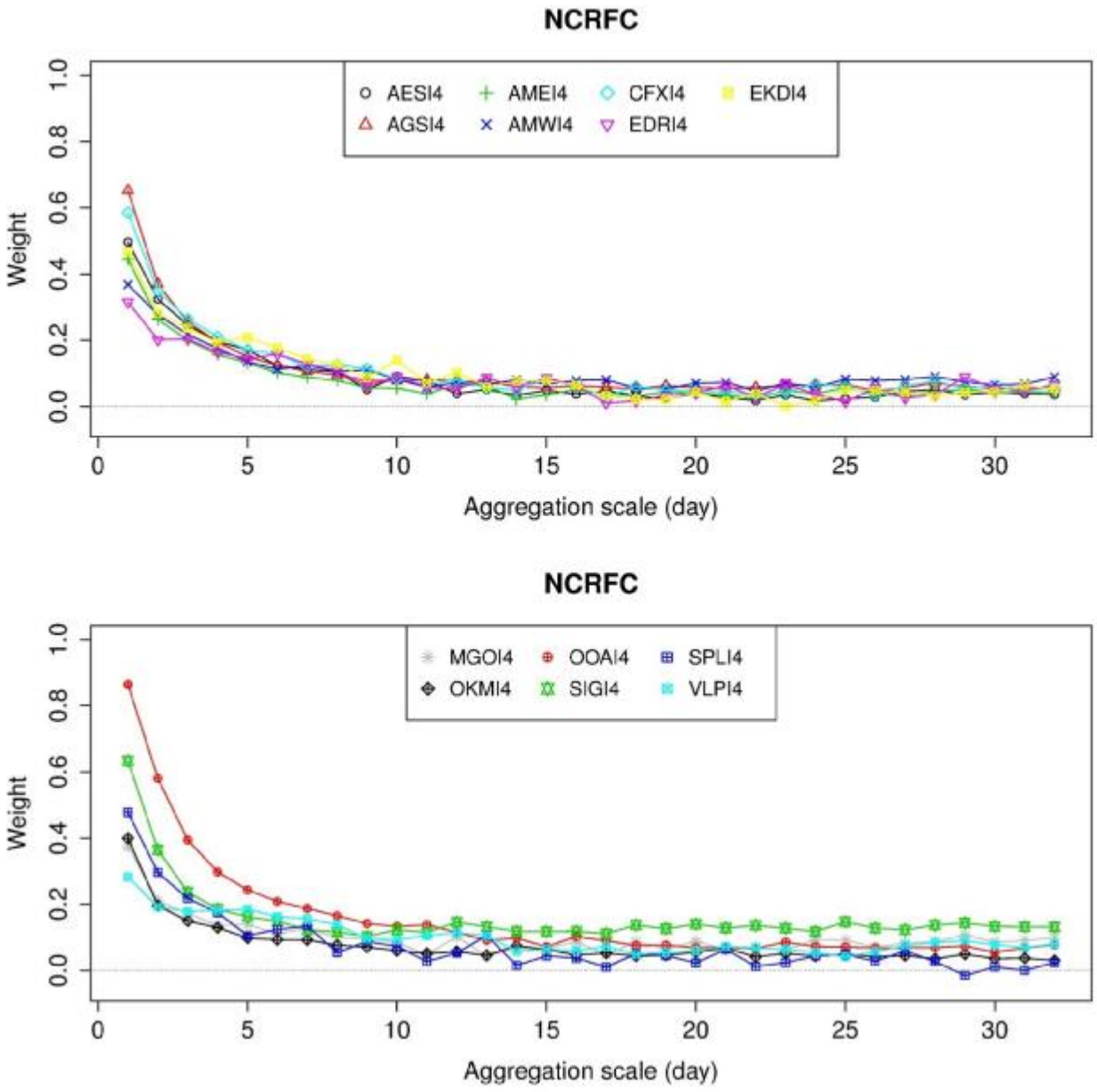


Figure 43: Regression weights vs. aggregation scale for basins in NCRFC.

North Skunk River in IA and drain relatively large areas of 1460.2 and 964.4 km^2, respectively. These rivers are regulated only midly via low-head dams which helps preserve the natural hydrologic memory and allow more skillful hydrologic modeling.

NERFC – The weight curves for the NERFC basins show large variations. A number of them show relatively strong memory with DICM1, FTEN6 and NSTN3 representing the strongest. DICM1 drains a very large area of 3017.4 km^2 in ME which contributes to strong memory due to large storage of water. FTEN6 drains a rather small area of 247.9 km^2 but is impacted appreciably by Great Sacandaga Lake and Indian Lake in NY. NSTN3 drains a large area of 2063.7 km^2 in NH and is impacted by power plants and by First Connecticut and Second Connecticut Lakes and Lake Francis upstream.

NWRFC – A large number of basins in this group show wide-ranging weight curves. Some of them exhibit effects of regulation but in different ways. CAMI1 drains a 1535.3 km^2 area in ID and is impacted by Lost Valley Reservoir upstream. LERI1 drains a 2400.4 km^2 area in ID and is greatly impacted by diversions above the station for irrigation of about 25,500 acres. The water of the Lemhi River and its tributaries is used for irrigation agriculture. Of the river's mainstem tributaries, only 7% are not totally disconnected year round due to diversion for irrigation. RILW1 drains a 489.3 km^2 area in WA and is impacted by Rimrock Lake. WCHW1 drains headwater flows over a 395.5 km^2 area from a glacier on the northwest side of Columbia Peak into the South Fork Sauk River in WA. At this location, the weight curve decreases very slowly as the time scale of aggregation increases due to the long memory in the glacier- and snowmelt processes. The weight curves for the above 4 basins and their varaiations serve to illustrate how MS-EnsPost utilizes scale-dependent hydrologic memory through multiscale regression.

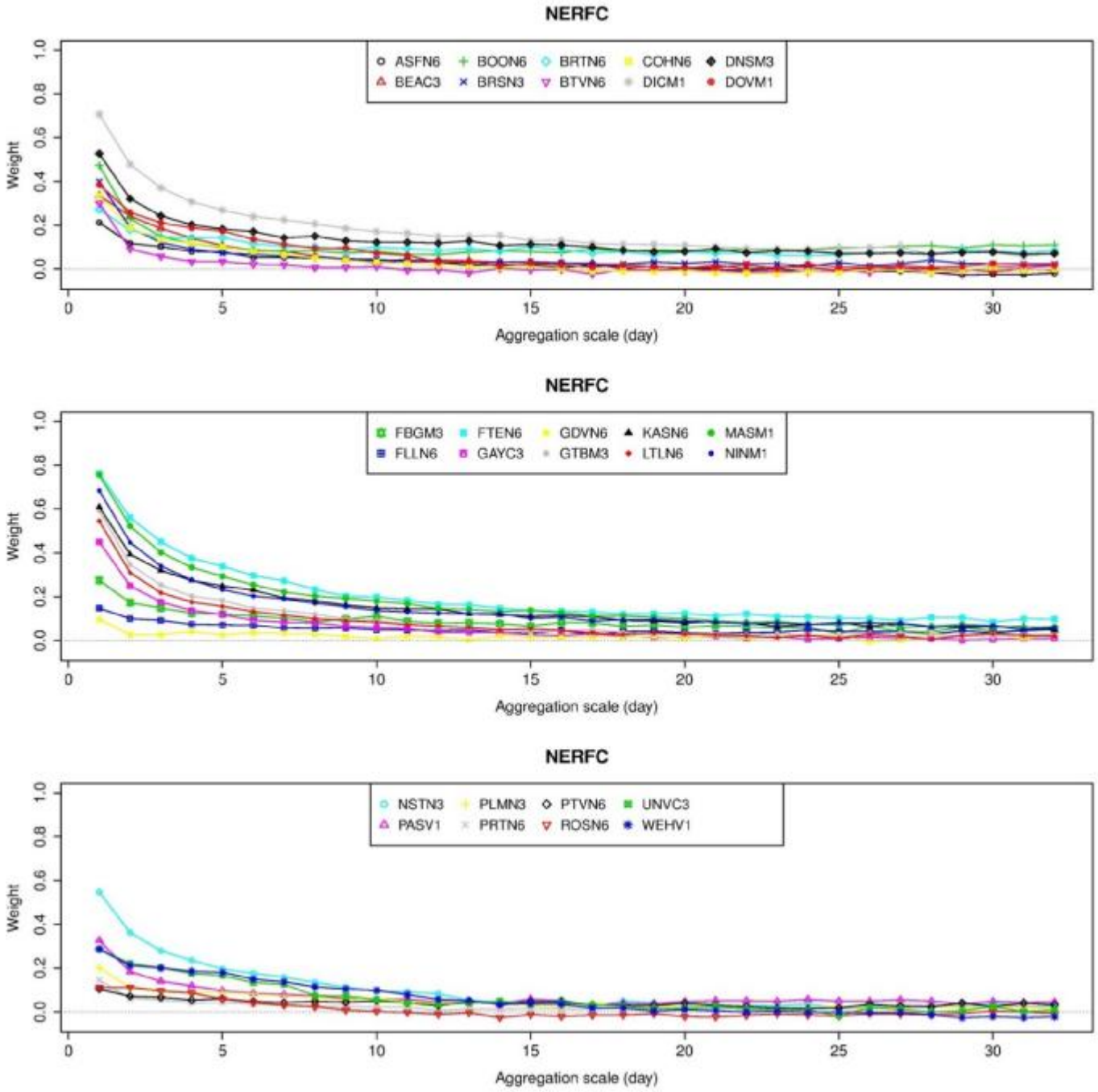


Figure 44: Regression weights vs. aggregation scale for basins in NERFC.

WGRFC – Of the WGRFC basins, RCET2 stands out as having the largest hydrologic memory at all scales of aggregation due to Bardwell Dam upstream. Though the largest in catchment area in this group, DCJT2 has the fastest decreasing memory vs. increasing temporal

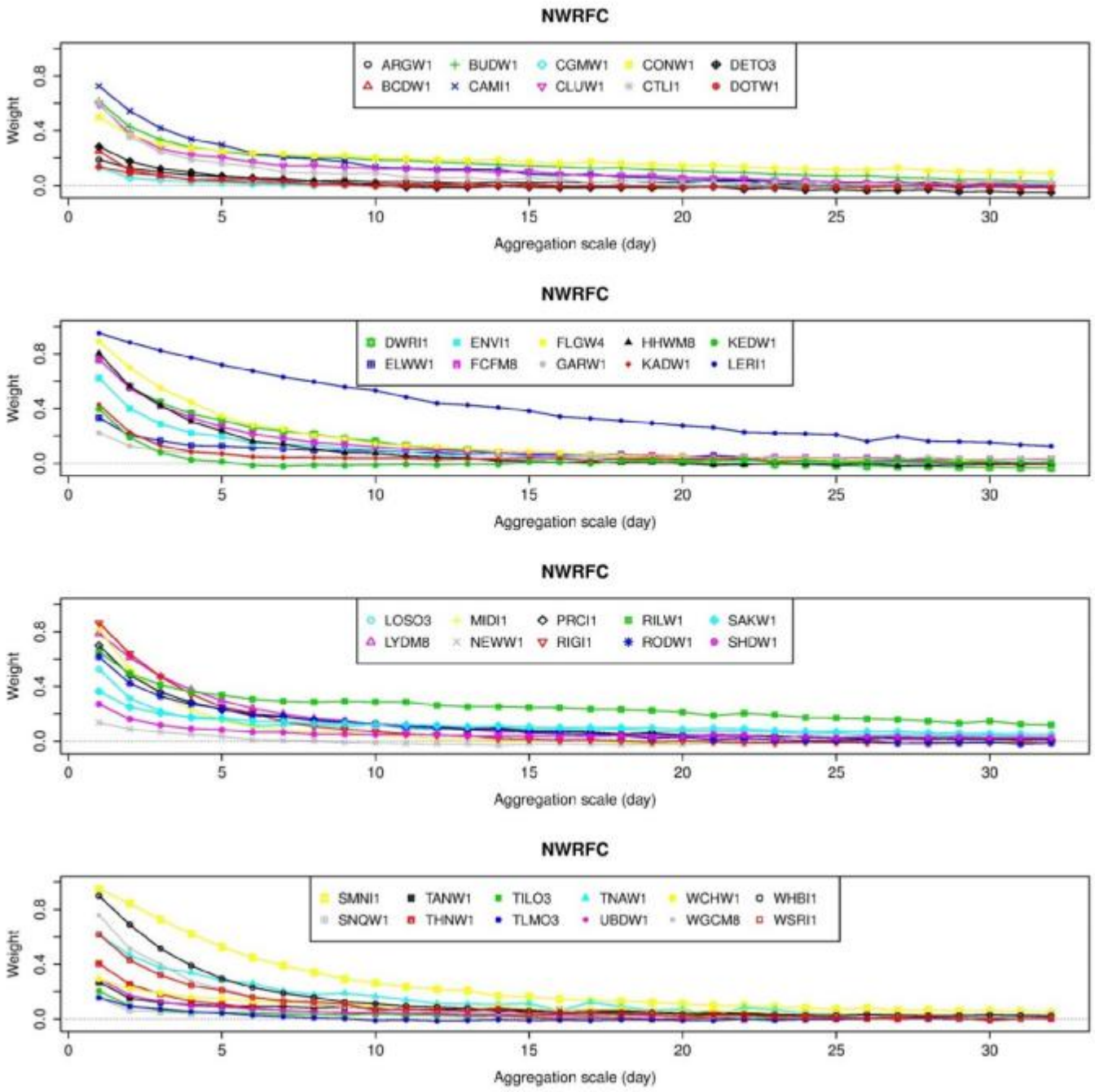


Figure 45: Regression weights vs. aggregation scale for basins in NWRFC.

scale of aggregation due probably to urbanization which quickens drainage of surface runoff. BRPT2 and JAKT2 show smaller memory than GLLT2 or SGET2 at time scales of aggregation of about 4 days or larger. The difference is probably due to the hydroclimatology;

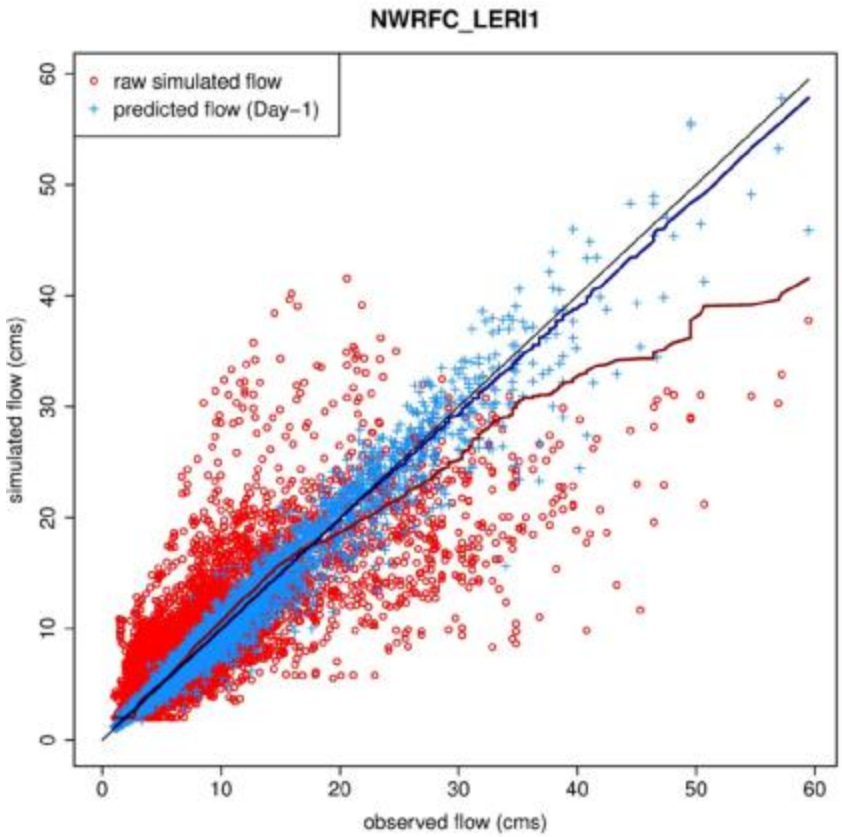


Figure 46: Scatter plot of observed vs. raw simulated (in red) and 1-day-ahead MS-EnsPost predicted flow (in blue) for LERI1 in NWRFC.

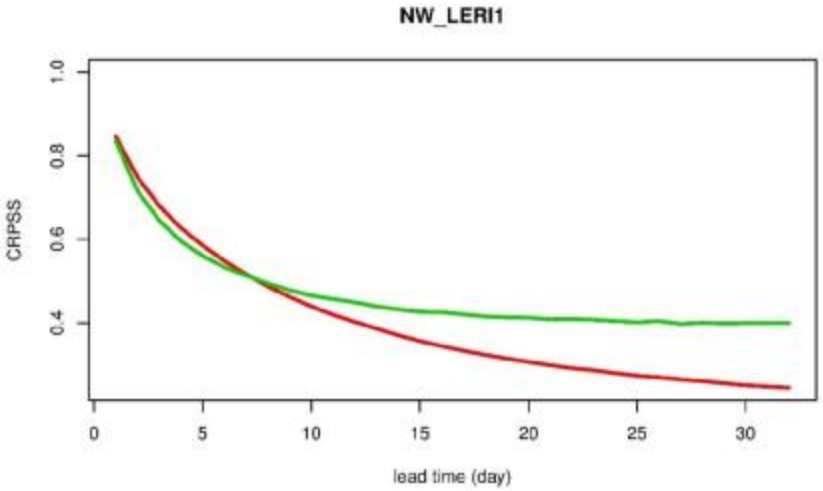


Figure 47: CRPSS from MS-EnsPost (in green) and EnsPost (in red) vs. lead time for LERI1 in NWRFC.

BRPT2 and JAKT2 are drier basins (see Fig 49) and do not hold as much surface and subsurface storages of water as GLLT2 and SGET2 do. The relatively large weights for SGET2 at large temporal scales of aggregation are somewhat puzzling. In this area, there exists a number of small agricultural reservoirs whose storage effects are not modeled (Mike Schultz, personal communication). It is suspected that the resulting detention storages contribute to the increased memory.

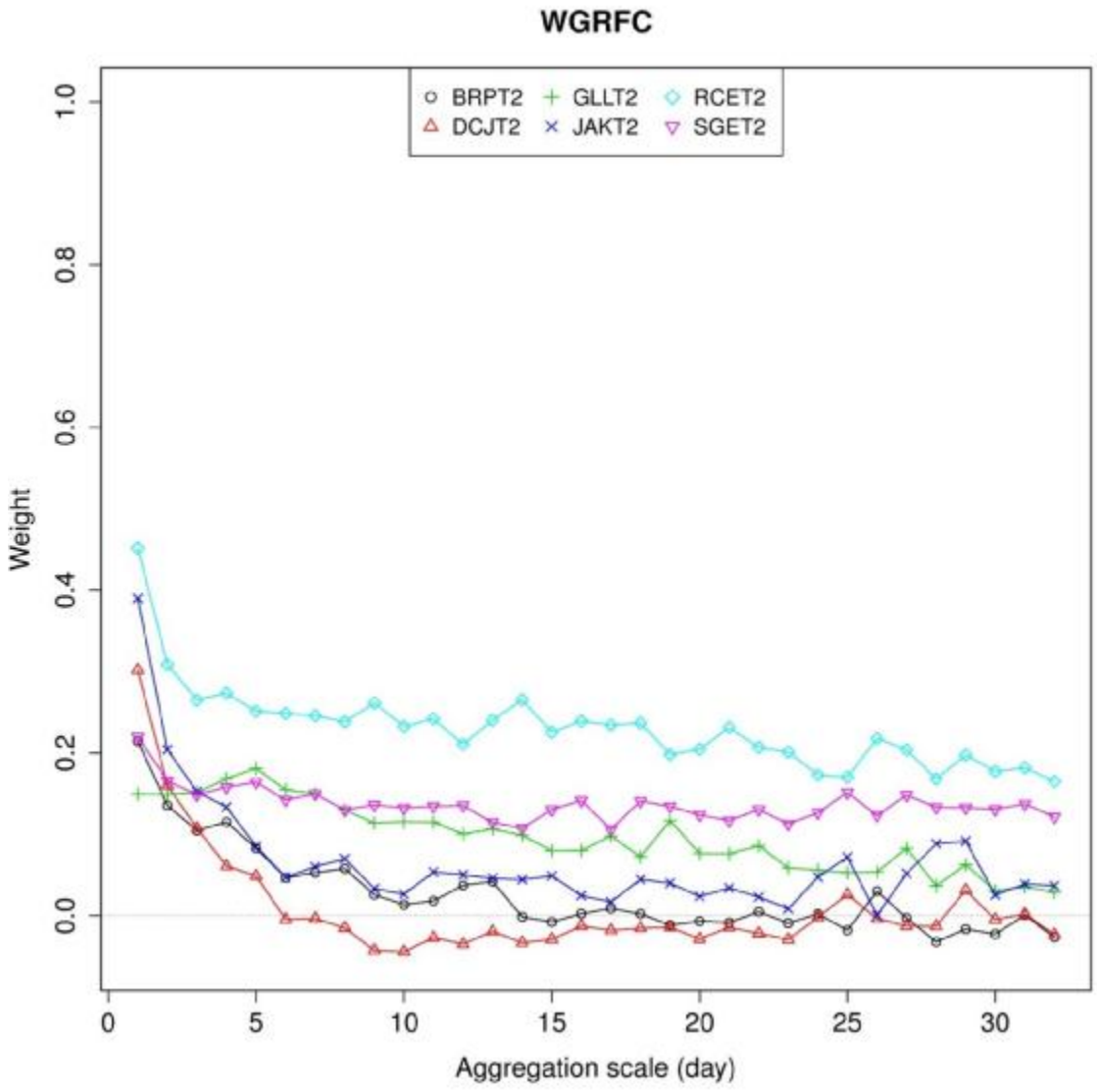


Figure 48: Regression weights vs. aggregation scale for basins in WGRFC.

The above observations indicate that human control of movement and storage of water affects the aggregation scale-dependent memory of a basin in a number of different ways. Whereas flow regulations significantly increase hydrologic memory for some basins, the opposite is also observed for others. They suggest that statistical modeling of regulated flows is likely to require significant complexity that varies greatly from location to location. The results presented in this work demonstate the utility and effectiveness of the multiscale regression approach in characterizing the scale-dependent hydrologic memory and fully utilizing it toward improving streamflow prediction parsimoniously.

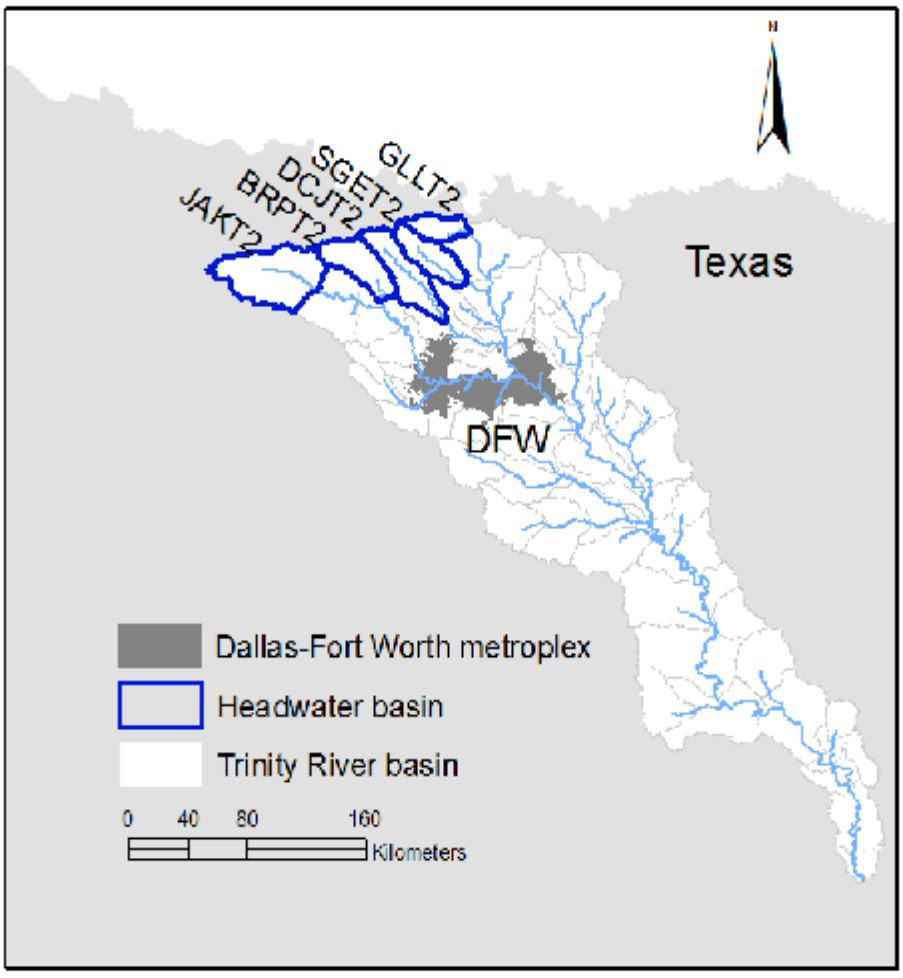


Figure 49: Headwater basins in WGRFC.

6.2.5 Sensitivity to period of record

One of the motivations for MS-EnsPost is to reduce the length requirement for the period of record so that nonstationarity may be considered. Streamflow responses have changed or are changing significantly in many parts of the world due to urbanization and climate change (Milly et al. 2008). Changing conditions force statistical post processors a difficult tradeoff between accounting for nonstationarities by dividing the period of record or modeling trends, which would significantly increase sampling uncertainties, vs. keeping sampling uncertainties smaller but at the expense of introducing biases due to nonstationarities. Owing to the parsimony, one may expect MS-EnsPost to require significantly less data than EnsPost. In this subsection, the relative performance of MS-EnsPost and EnsPost is evaluated under a reduced period of record. For this, the following experiments were carried out for each basin selected. It is important to note that the primary purpose of the experiments was not to identify and test nonstationarity, but to assess relative performance under reduced data availability. As such, the periods of record are devided in equal lengths whether the midpoints represent change points or not. The process comprise the following steps:

1) Divide the the entire period of record into two subperiods of equal length,
2) For each subperiod, carry out leave-two-year-out cross validation for EnsPost and MS-EnsPost, and
3) Comparatively evaluate the EnsPost and MS-EnsPost results for each subperiod.

If one procedure is superior to the other under the increased sampling uncertainty, one may expect the superior procedure to improve over the other for each of the two subperiods. In making the above comparison, one may place more confidence if the streamflow time series are significantly different between the two subperiods. To that end, 19 basins were selected from the

7 RFCs that exhihit the largest differences in the empirical CDFs (ECDF) of observed daily flow between the two periods. Figs 50 through 54 show selected examples. In each figure, the ECDF of observed daily flow for the entire period of record is shown in black, and the ECDFs for the 1st and 2nd subperiods are shown in puple and blue, respectively. In the inset, the right tails of the ECDFs above the 95th percentile are shown to assess the difference in ECDF for high flows. The Kolmogorov-Smirnov test (Kolmogorov 1933; Smirnov 1948) for these basins indicates that one

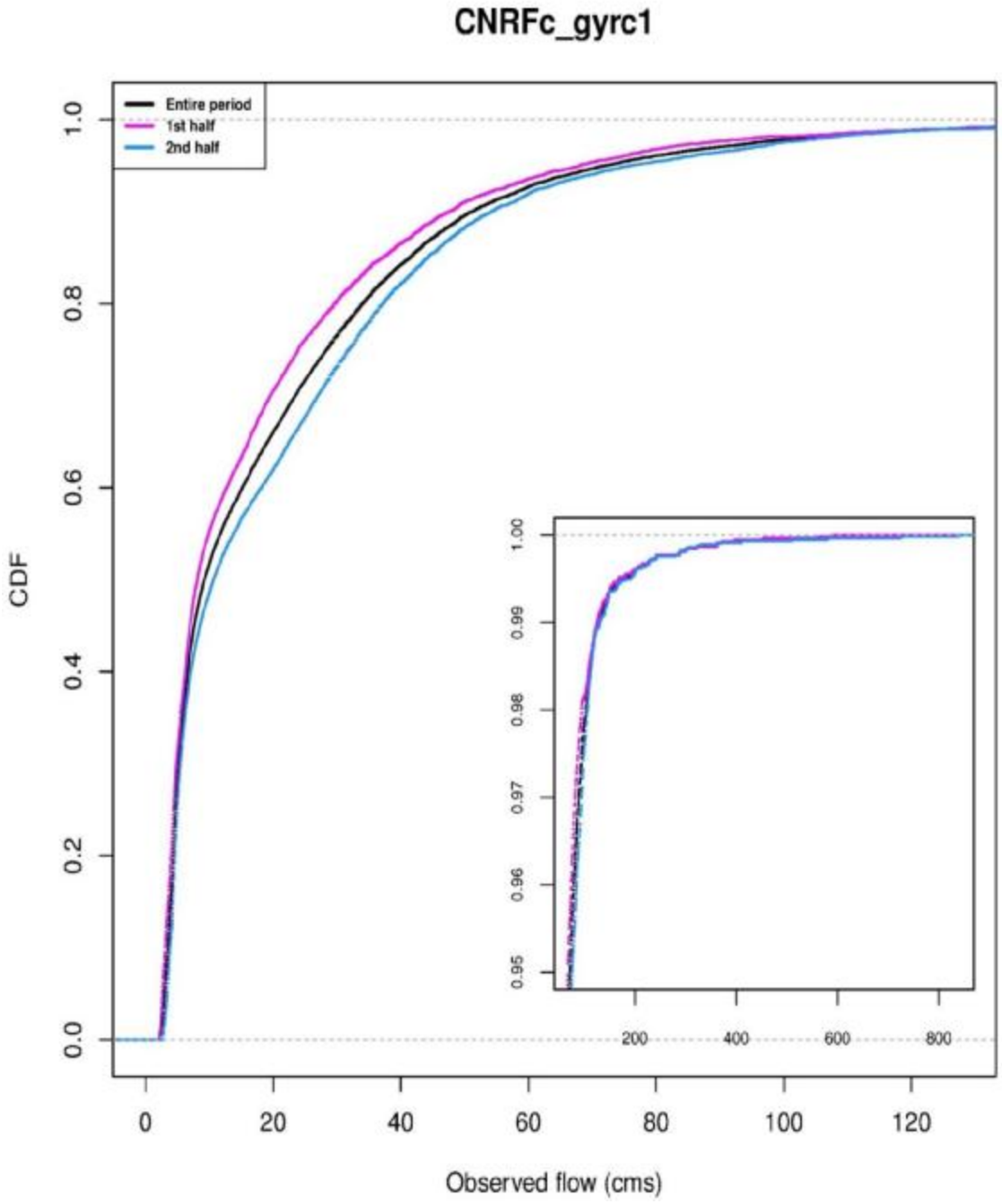


Figure 50: Empirical CDF of observed flow in the entire period of record (in black), first half (in purple) and second half (in blue) for GYRC1 in CNRFC.

may very safely reject the null hypothesis that two ECDFs for the two subperiods come from the same population.

Though not the focus of this subsection, it is worth noting the findings from the visual examination of the ECDFs from the halved time series for all basins. They indicate rather strong

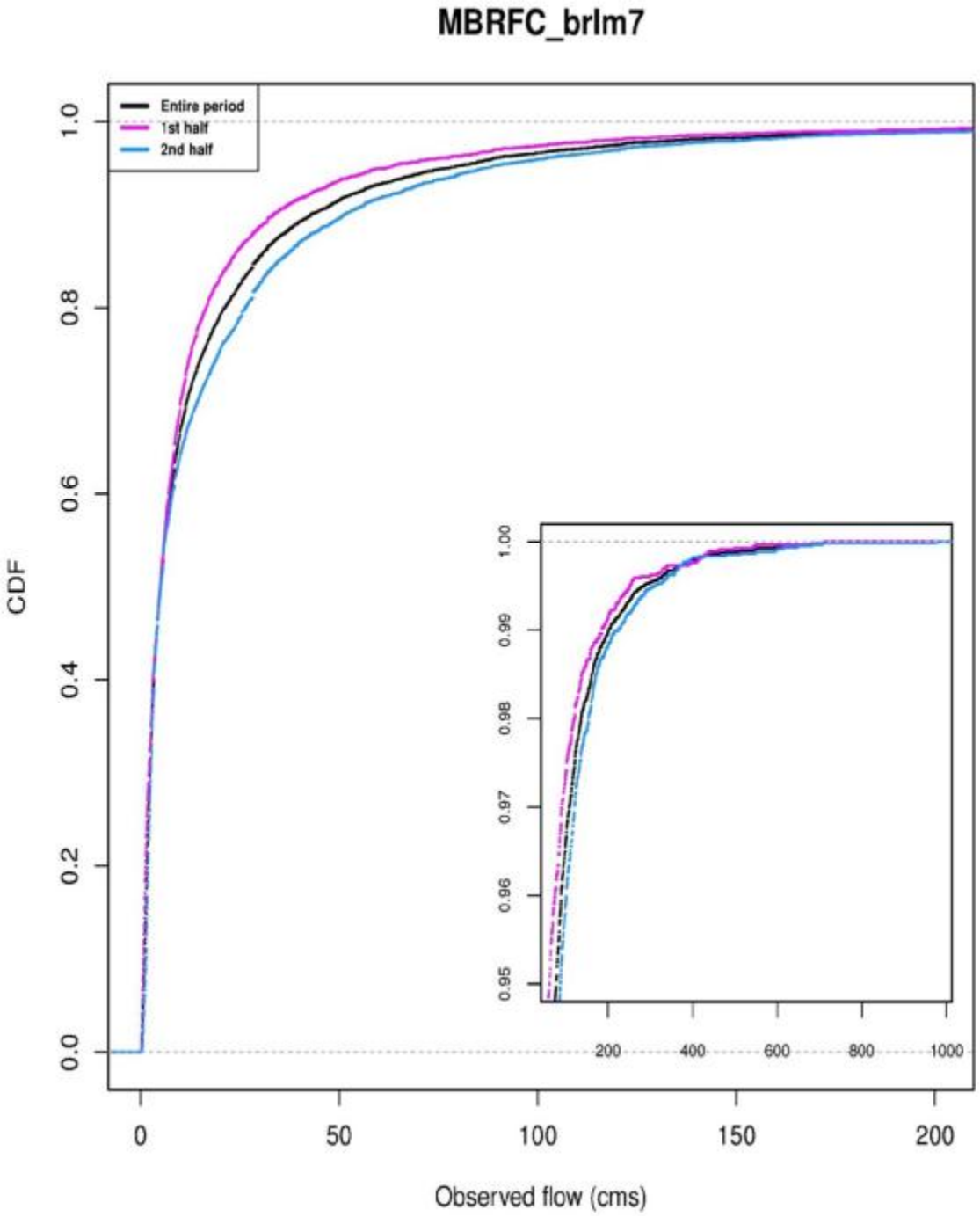


Figure 51: Empirical CDF of observed flow in the entire period of record (in black), first half (in purple) and second half (in blue) for BRLM7 in MBRFC.

trends for a number of groups of basins; most of the MA-, MB- and NCRFC basins have become wetter in the 2nd subperiod, many of the CB- and NWRFC basins have become drier in the 2nd period, and many of the CNRFC basins have become wetter in the 2nd period.

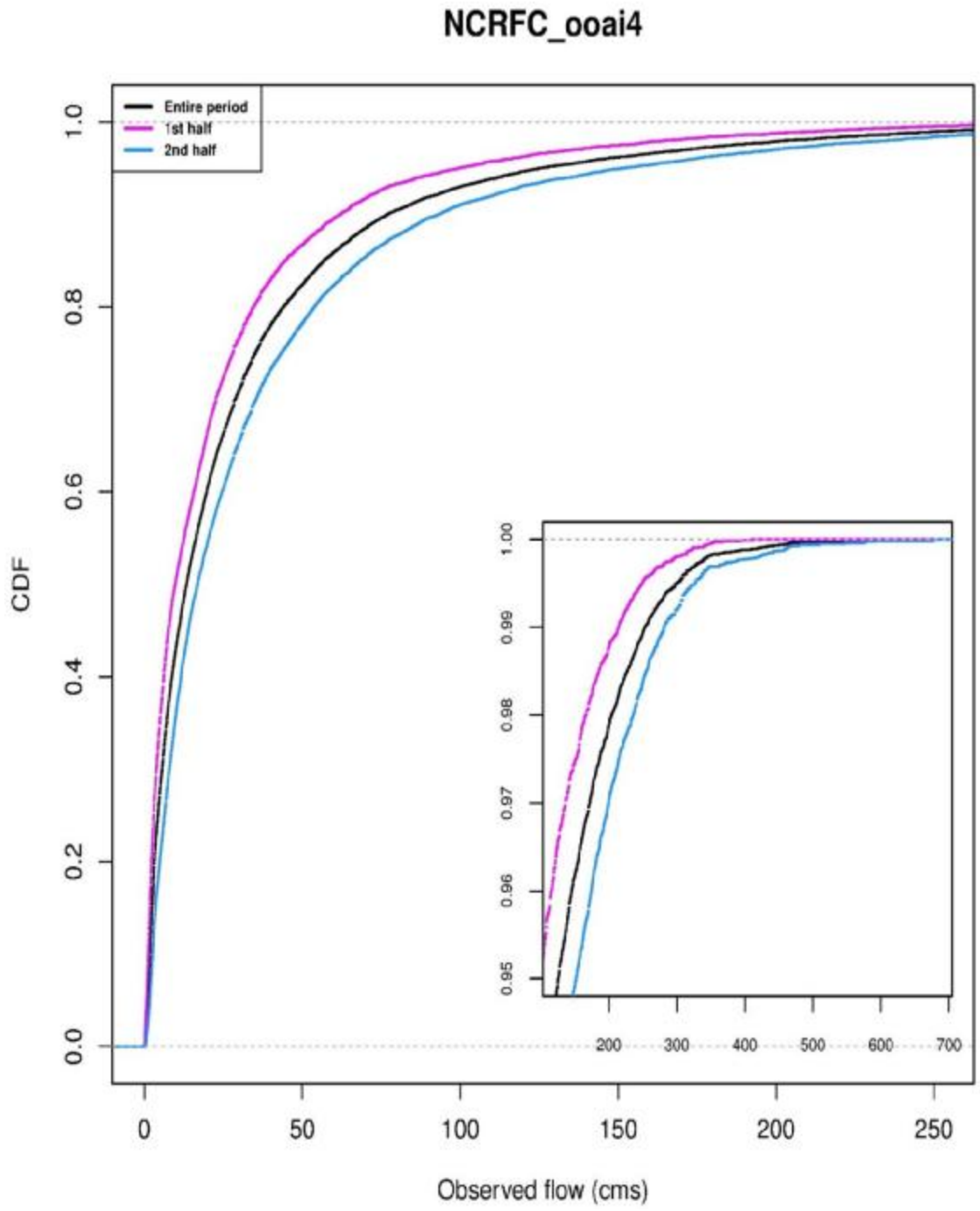


Figure 52: Empirical CDF of observed flow in the entire period of record (in black), first half (in purple) and second half (in blue) for OOAI4 in NCRFC.

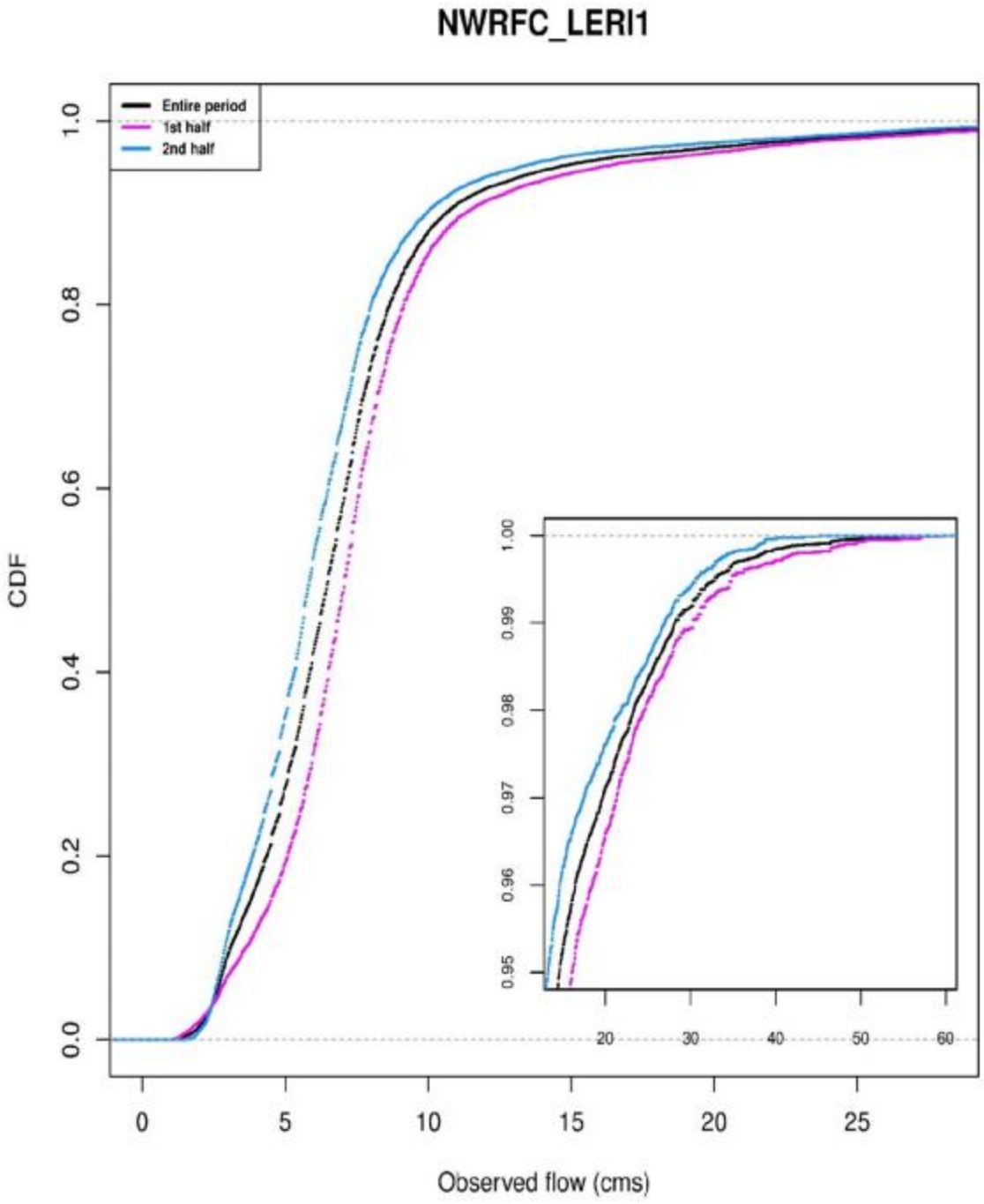


Figure 53: Empirical CDF of observed flow in the entire period of record (in black), first half (in purple) and second half (in blue) for LERI1 in NWRFC.

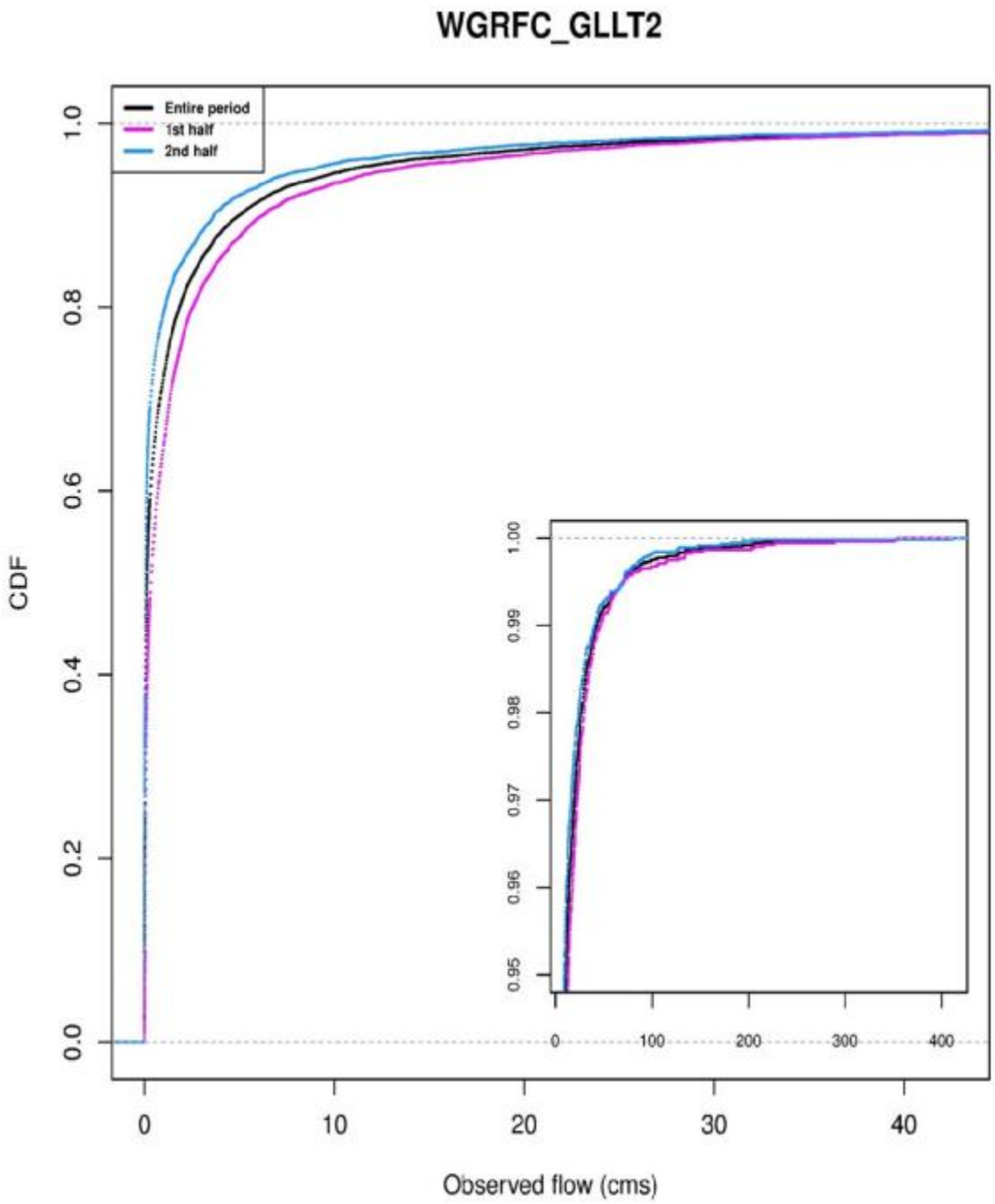


Figure 54: Empirical CDF of observed flow in the entire period of record (in black), first half (in purple) and second half (in blue) for GLLT2 in WGRFC.

Figs 55 and 56 show the worm plots for the selected basins for the two subperiods. As explained in Subsection 6.2.2, the lower and upper ends of each worm are associated with Day-1 and -7 predictions for that basin, respectively. If MS-EnsPost improves over EnsPost for 7 day-

ahead prediction, the worms would stretch downward from the diagonal. The longer the downward stretch, the larger the improvement by MS-EnsPost over EnsPost. If MS-EnsPost does not improve over EnsPost, the worms would lie along the diagonal. Fig 55 shows that, in the first subperiod, MS-EnsPost improves over EnsPost for Day-1 to Day-7 predictions for 13 out of the 19 basins, and performs comparably for the other 6 basins. Fig 56 also shows that, in the second subperiod, MS-EnsPost outperforms EnsPost for 11 out of the 19 basins, and performs comparably for the other 6 basins. For BRPT2 and GLLT2 in WGRFC, EnsPost outperformed MS-EnsPost. As explained above, the performance of MS-EnsnPost is not as good in drier conditions as it is in wetter conditions. This is because the Box-Cox transform is not able to satisfy truncated normality of the error across all ranges in the transformed space when the hydroclimatological conditions became drier in the above two basins.

As noted in Subsection 3.2.3, the Box-Cox parameter, λ (see Eq.(10)), is chosen such that the MS-EnsPost results are better for high flows than for low flows. EnsPost, on the other hand, uses NQT so that, its results are of similar quality for both high and low flows. For forecasting of floods and water supply, however, performance for high flows is much more important than that for low flows. For this reason, we also compare the performance of EnsPost and MS-EnsPost for the two subperiods for the verifying observed flows exceeding the 95^{th} percentile of observed flow. Figs 57 and 58 are the same as Figs 55 and 56, respectively, but for the verifying observed daily flow exceeding the 95^{th} percentile. Fig 57 shows that, in the first subperiod, MS-EnsPost improved over EnsPost for 10 out of the 19 basins for high flows. Similarly, Fig 58 indicates modest to significant improvement by MS-EnsPost over EnsPost for 15 out of the 19 basins. MS-EnsnPost performed comparably to EnsPost for 3 basins. Note that, while the margin of improvement is not very large, MS-EnsPost outperforms EnsPost for BRPT2 and GLLT2 in

WGRFC in Fig 58. This is because the Box-Cox transformation provides better approximation of truncated normality for larger amounts of verifying observed flow.

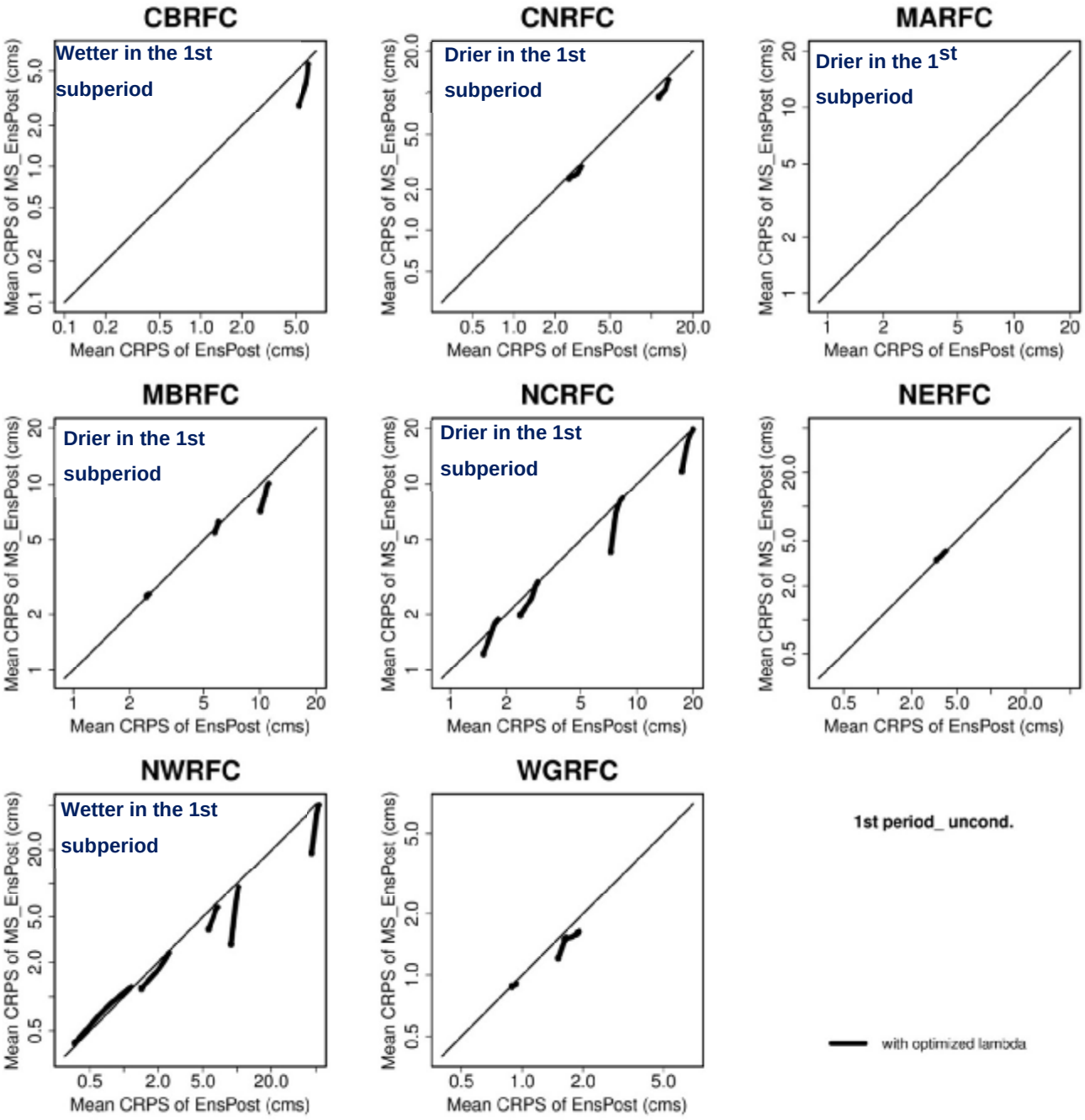


Figure 55: Worm plots (see text for explanation) of mean CRPS of ensemble predictions of daily flow from MS-EnsPost and EnsPost for lead times of 1 to 7 days in the first half of period of record for 19 basins with large differences in empirical CDFs.

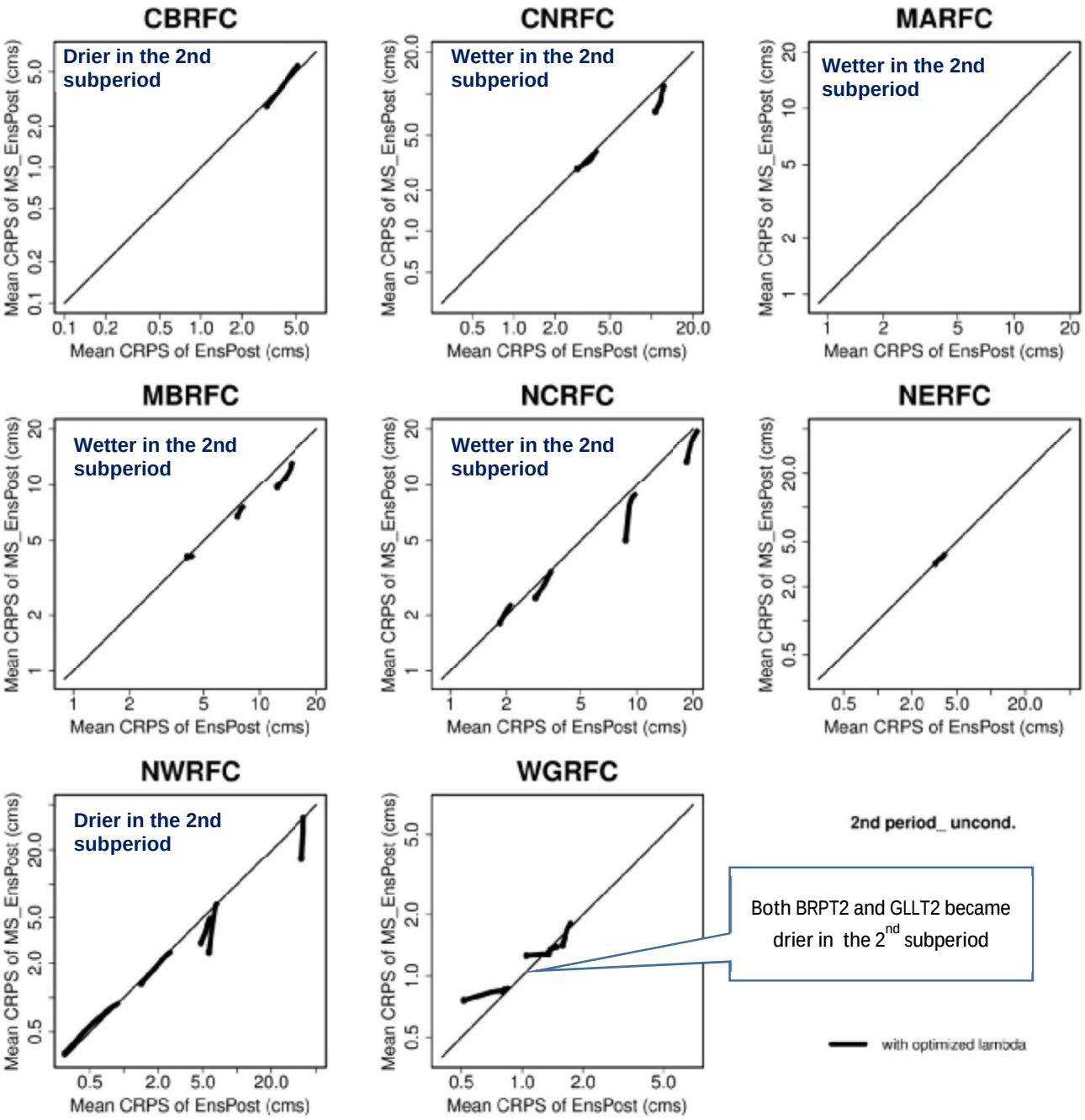


Figure 56:Worm plots (see text for explanation) of mean CRPS of ensemble predictions of daily flow from MS-EnsPost and EnsPost for lead times of 1 to 7 days in the second half of period of record for 19 basins with large differences in empirical CDFs.

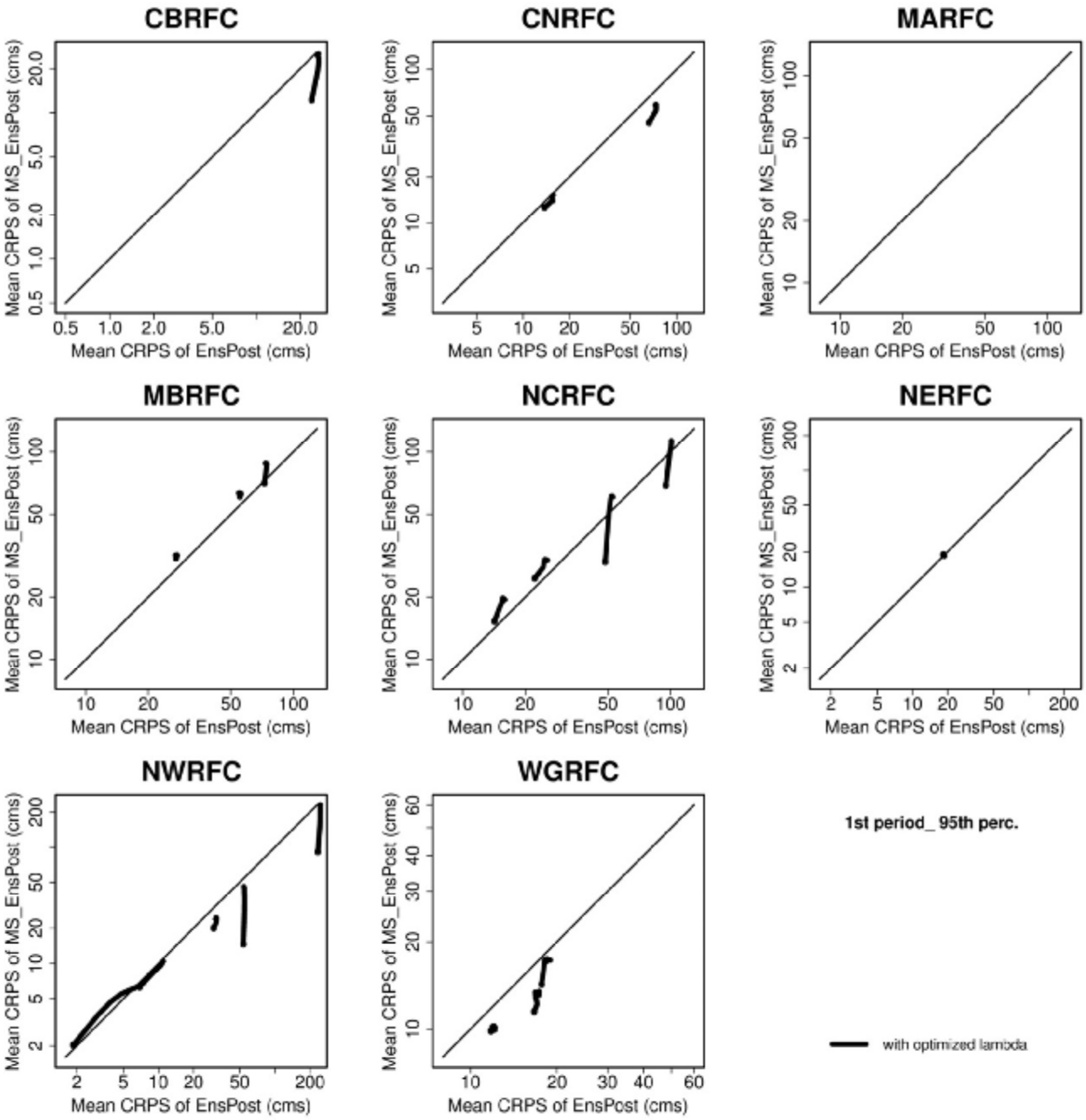


Figure 57: Worm plots (see text for explanation) of mean CRPS (exceeding 95th percentile of observef flow) of ensemble predictions of daily flow from MS-EnsPost and EnsPost for lead times of 1 to 7 days in the first half of period of record for 19 basins with large differences in empirical CDFs.

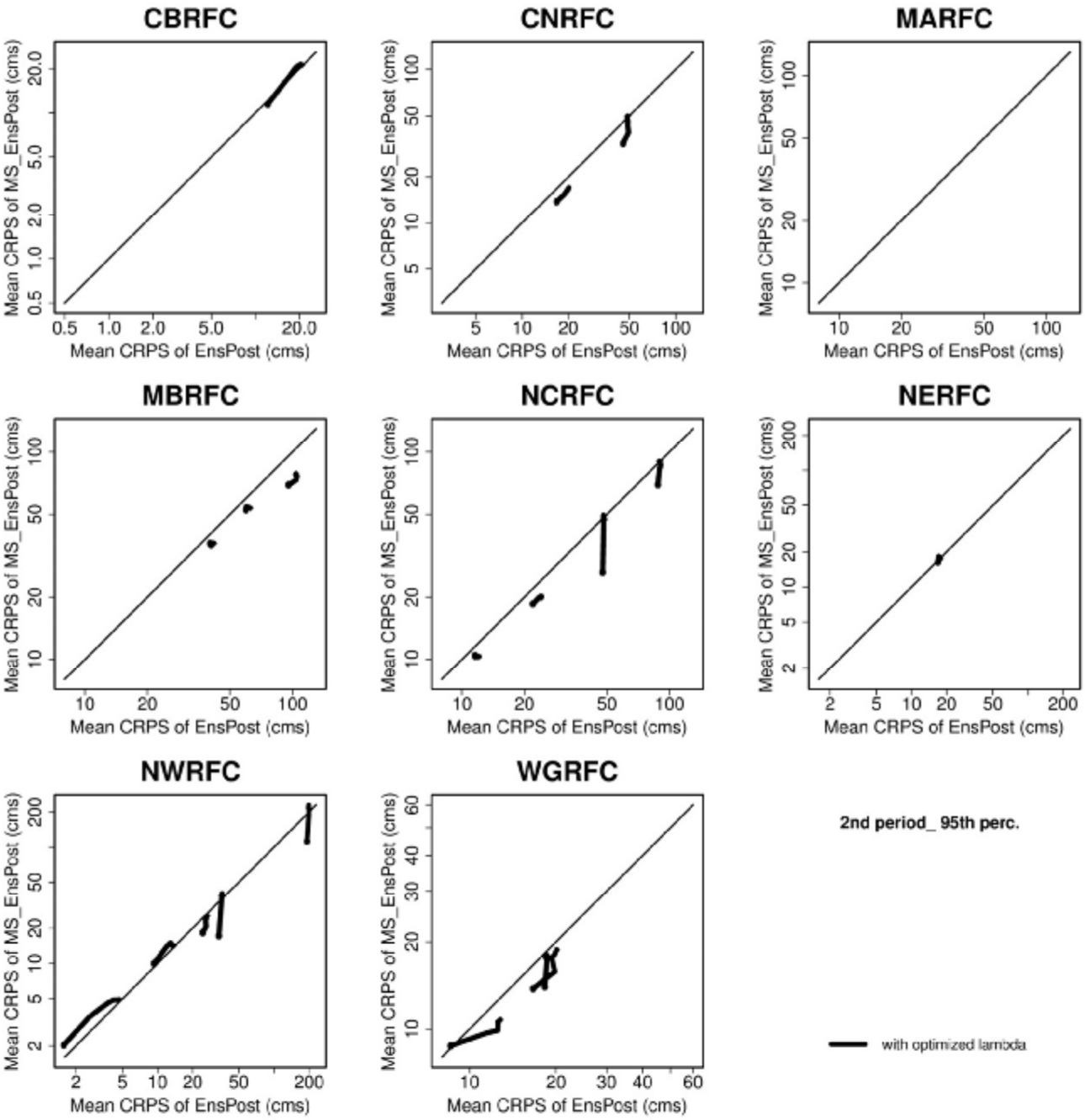


Figure 58: Worm plots (see text for explanation) of mean CRPS (exceeding 95th percentile of observef flow) of ensemble predictions of daily flow from MS-EnsPost and EnsPost for lead times of 1 to 7 days in the second half of period of record for 19 basins with large differences in empirical CDFs.

Chapter 7

Conclusions and future research recommendations

A novel multi-scale post-processor, MS-EnsPost, for ensemble streamflow prediction and a multiscale probability matching (MS-PM) technique for bias correction in streamflow simulation are developed and evaluated. The MS-PM was developed originally to improve the bias correction component of the existing ensemble post-processor, EnsPost, in the NWS's Hydrologic Ensemble Forecast Service. The MS-PM successively applies probability matching (PM) across multiple time scales of aggregation to reduce scale-dependent biases in streamflow simulation. The evaluation results for 34 basins in the service areas of the Colorado Basin (CB-), California-Nevada (CN-), Middle-Atlantic (MA-), and West Gulf (WG-) River Forecast Centers (RFC) show that MS-PM improves over PM for streamflow prediction at a daily time step, that averaging the empirical cumulative distribution functions to reduce sampling uncertainty marginally improves performance, but that the performance of MS-PM quickly reaches a limit with the addition of larger temporal scales of aggregation due to the increasingly large sampling uncertainties.

MS-EnsPost represents a departure from the PM-based approaches so that large sampling uncertainties associated with empirical distribution modeling may be avoided, and that the predictive skill in model-simulated and observed streamflow that may exist over a range of temporal scales may be fully utilized. MS-EnsPost uses data-driven correction of magnitude-dependent biases in model-simulated flow, multiscale regression to utilize observed and simulated flows over a range of temporal scales of aggregation, and ensemble generation based on parsimonious error modeling. MS-EnsPost is evaluated using 139 basins in the service areas

of CB-, CN-, MA-, Missouri Basin (MB-), North Central (NC-), Northeast (NE-), Northwest (NW-), and WGRFC. The main findings are as follows.

MS-EnsPost outperformed EnsPost at all lead times in the root mean square error (RMSE) sense for 137 out of 139 basins. The reduction in RMSE ranged from 5 to 68% for Day-1 to -7 predictions of daily flow. For most basins, the improvement is due to both bias correction and multiscale regression in MS-EnsPost. MS-EnsPost outperformed EnsPost at all lead times in the mean continuous ranked probability score (CRPS) sense for 136 out of 139 basins. The reduction in mean CRPS ranged from 2 to 62% for Day-1 to -7 predictions of daily flow. The improvement is due mostly to improved resolution than reliability in the MS-EnsPost ensembles. The improvement is particularly significant for the Upper Trinity River basins in the WGRFC's service area, an indication that the bias correction and multiscale regression procedures are effective in addressing flow magnitude-dependent biases in raw model-predicted flow, and intermittency of streamflow in the semi-arid region.

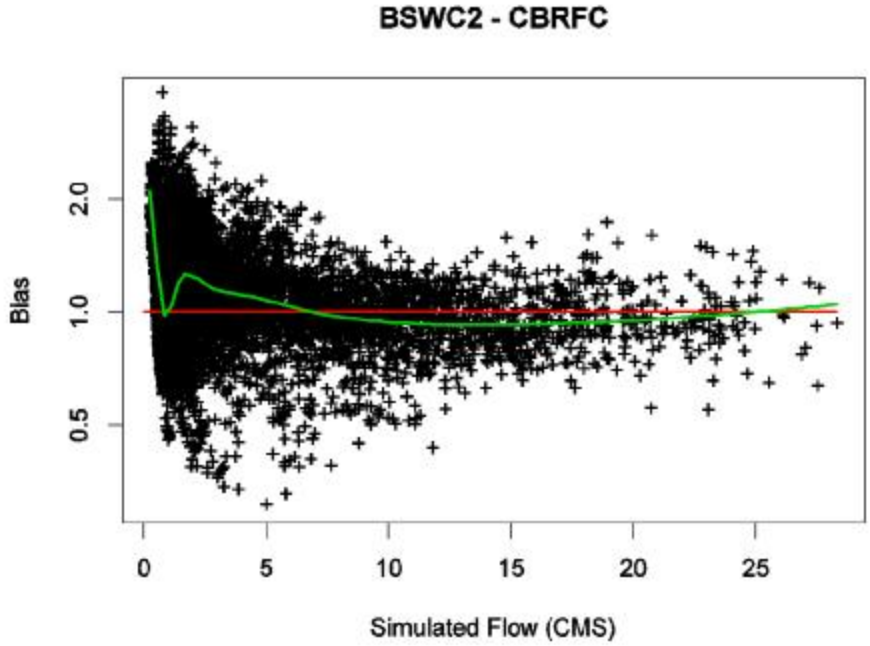
Assessment of predictability measured by the continuous ranked probability skill score (CRPSS) of the MS-EnsPost predictions indicate that, among the basins considered in this work, the CB-, CN-, NWRFC basins are the most predictable, followed by the NE-, MA-, and NCRFC basins, and that the MB- and WGRFC basins are the least predictable. Comparison of the skill scores with hydroclimatic indices indicates that, with the current operational hydrologic modeling process, predictability of streamflow and its limits are strongly modulated by the fraction of mean annual precipitation as snow and, in non-snow-driven basins, mean annual precipitation. The positive impact of MS-EnsPost is particularly significant for a number of basins impacted by flow regulations. Examination of the multiscale regression weights indicates that MS-EnsPost is able to capture and reflect the scale-dependent alterations by flow regulations

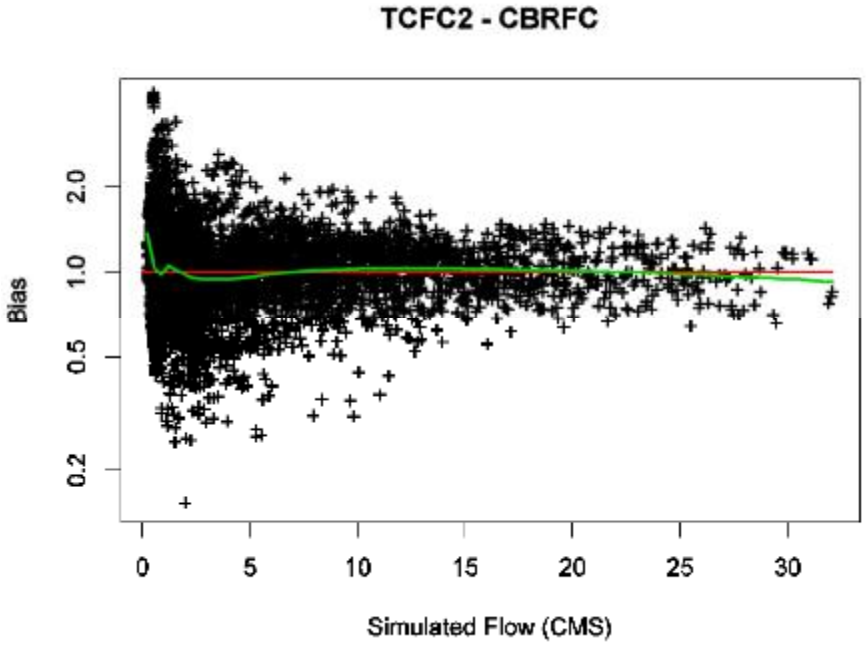
to the predictive skills of observed and model-predicted flow, resulting in improved performance over EnsPost. One of the motivations for MS-EnsPost is to reduce data requirement so that nonstationarity may be considered under changing hydroclimatology. Comparative evaluation of MS-EnsPost with EnsPost indicates that, under reduced data availability, MS-EnsPost generally outperforms EnsPost for those basins exhibiting significant changes in flow regime.

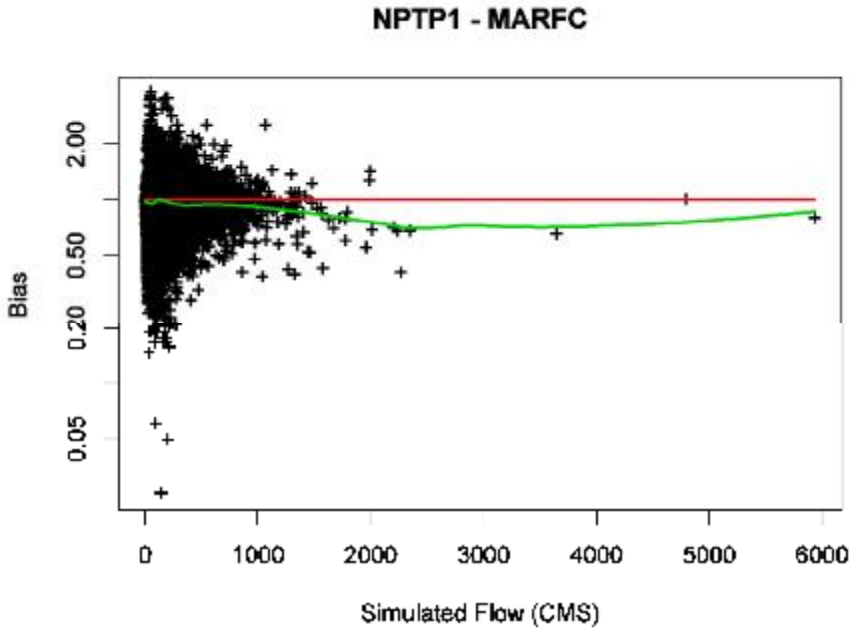
MS-EnsPost uses the Box-Cox transformation rather than the empirical normal quantile transform as in EnsPost. Whereas the former greatly improves parsimony and reduces complexity, the resulting errors may not meet truncated-normality and homoscedasticity as assumed in this work. While the above approximation is able to produce reliable ensembles for moderate to high flows, it tends to reduce reliability of streamflow ensembles in low flow conditions. In this work, the first-order autoregressive process was assumed for time series modeling of the error for simplicity and for direct comparison with EnsPost. The limited results suggest that the use of a general time series model is likely to improve ensemble prediction of time-aggregated flows. In this work, the data reduction experiment under possible nonstationarity was carried out by splitting the period of record into two halves and repeating the comparative evaluation experiments for both halved periods. Additional efforts are needed to render the errors closer to truncated normal and homoscedastic through an improved and more objective process, to improve the temporal dependence modeling of the error for improved prediction of time-integrated flow, and to assess performance under reduced data following rigorous identification of nonstationarity. Whereas the above efforts are not likely to change significantly the overall performance of MS-EnsPost reported herein, they are likely to improve significantly its potential operational worthiness, and yield specific guidance on parameter settings and their refinement.

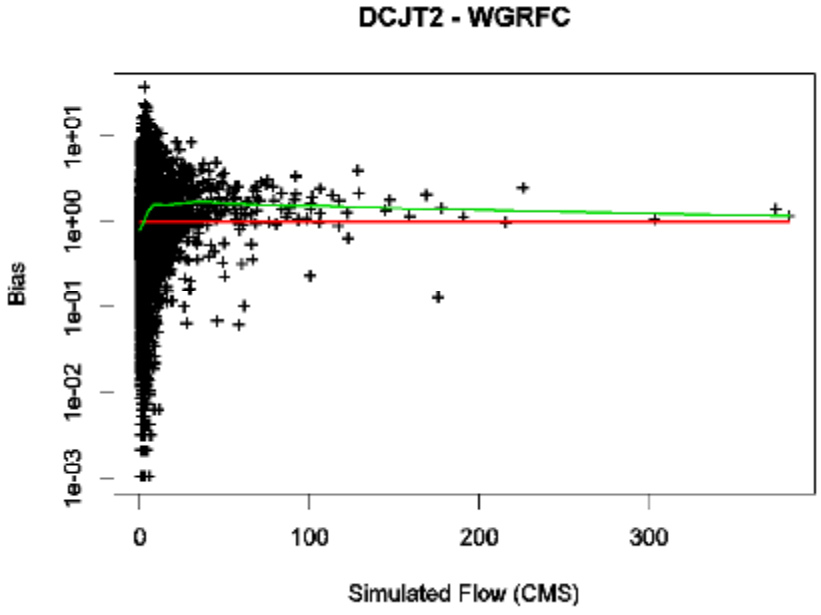
Appendix A

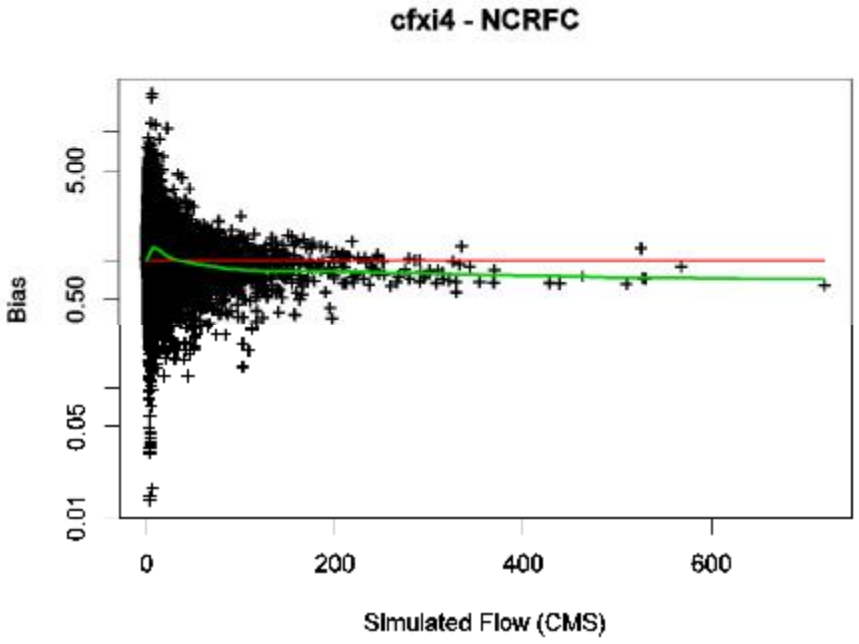
Magnitude-dependent biases for selected basins estimated in MS-EnsPost

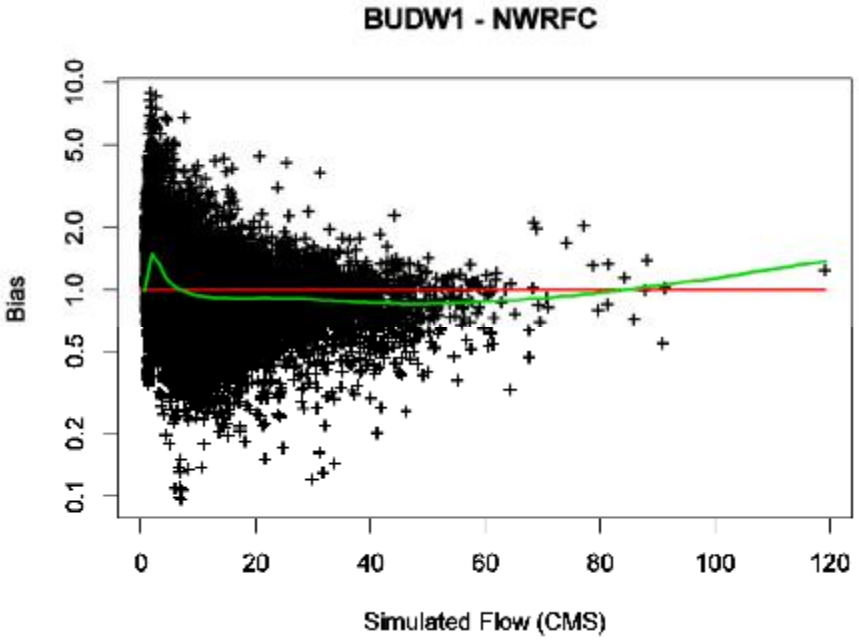

BSWC2 - CBRFC
Bias
0.5
1.0
2.0
0
5
10
15
20
25
Simulated Flow (CMS)


TCFC2 - CBRFC
Bias
0.2
0.5
1.0
2.0
0
5
10
15
20
25
30
Simulated Flow (CMS)


NPTP1 - MARFC
Bias
2.00
0.50
0.20
0.05
0
1000
2000
3000
4000
5000
6000
Simulated Flow (CMS)

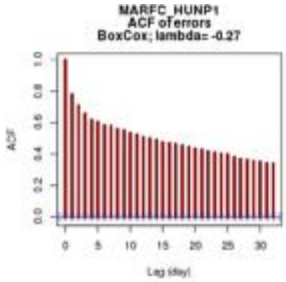

DCJT2 - WGRFC
Bias
1e+01
1e+00
1e-01
1e-02
1e-03
0
100
200
300
Simulated Flow (CMS)

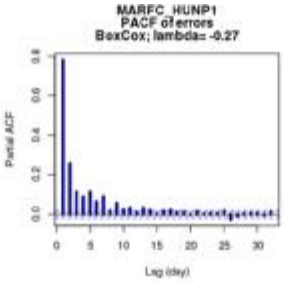

cfxi4 - NCRFC
Bias
0.01
0.05
0.50
5.00
0
200
400
600
Simulated Flow (CMS)


BUDW1 - NWRFC
Bias
0.1
0.2
0.5
1.0
2.0
5.0
10.0
0
20
40
60
80
100
120
Simulated Flow (CMS)

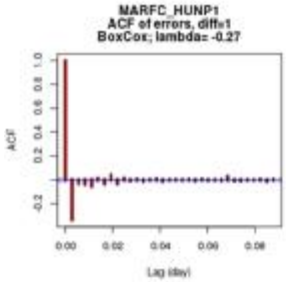
Appendix B

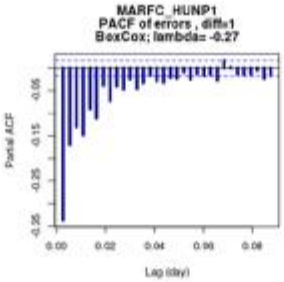
Time series modeling of error for selected basins





KPSS Test for Level Stationarity:
Non-stationary

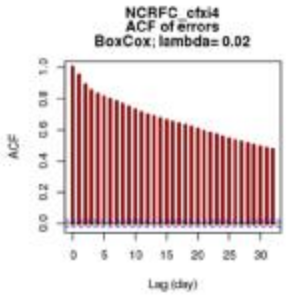


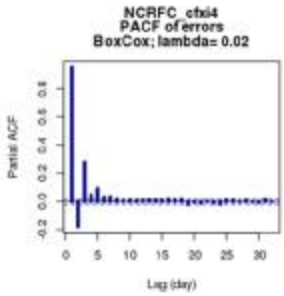


First order differencing

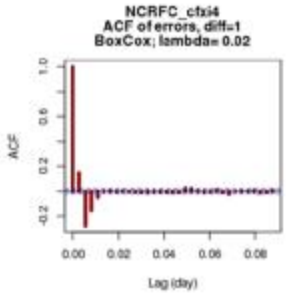
KPSS Test for Level Stationarity:
Stationary

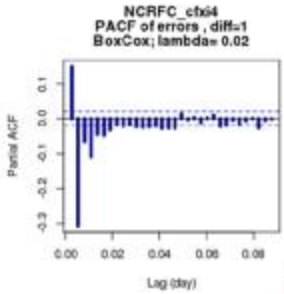
```
> auto.arima(qerrorboxcox2ts)
Series: qerrorboxcox2ts
ARIMA(3,1,3)

Coefficients:
          ar1     ar2     ar3     ma1     ma2     ma3
      -0.7591  0.3006  0.2242  0.2800  -0.7610  -0.237
s.e.   0.0923  0.0730  0.0382  0.0935   0.0419   0.086

sigma^2 estimated as 0.007845:  log likelihood=12991.04
AIC=-25968.07   AICc=-25968.06   BIC=-25915.8
```





KPSS Test for Level Stationarity:
Non-stationary

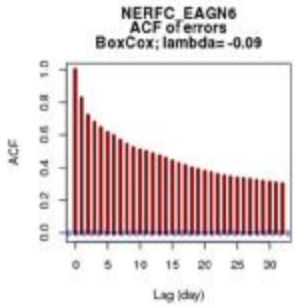


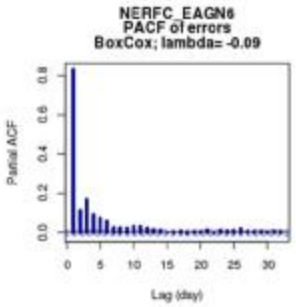


First order differencing

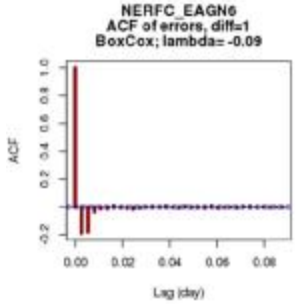
KPSS Test for Level Stationarity:
Stationary

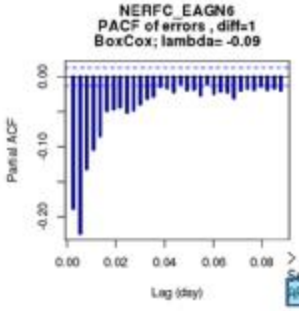
```
> auto.arima(qerrorboxcox2ts)
Series: qerrorboxcox2ts
ARIMA(4,1,2)

Coefficients:
          ar1     ar2      ar3     ar4      ma1     ma2
      0.3946  0.1357  -0.0725  0.0720  -0.2382  -0.5112
s.e.  0.1173  0.1263   0.0576  0.0312   0.1165   0.1081

sigma^2 estimated as 0.03727:  log likelihood=2225.92
AIC=-4437.84   AICc=-4437.83   BIC=-4387.47
```





KPSS Test for Level Stationarity:
Non-stationary

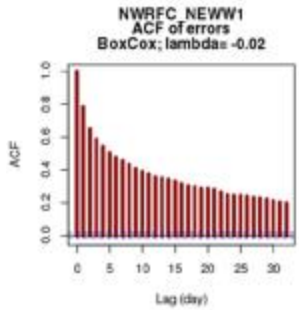


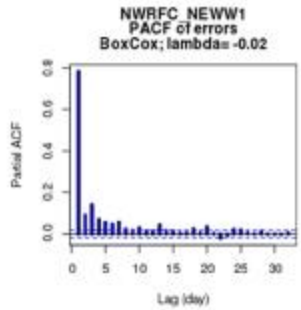


First order differencing

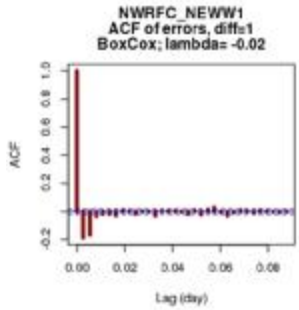
KPSS Test for Level Stationarity:
Stationary

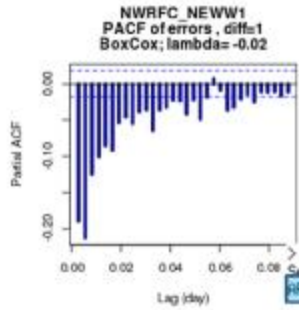
```
> auto.arima(qerrorboxcox2ts)
Series: qerrorboxcox2ts
ARIMA(3,1,2)

Coefficients:
         ar1      ar2     ar3      ma1     ma2
      1.3364  -0.4581  0.0748  -1.6509  0.655
s.e.  0.0277   0.0174  0.0098   0.0276  0.027

sigma^2 estimated as 0.01885:  log likelihood=13203.34
AIC=-26390.79   AICc=-26390.79   BIC=-26342.46
```





KPSS Test for Level Stationarity:
Non-stationary

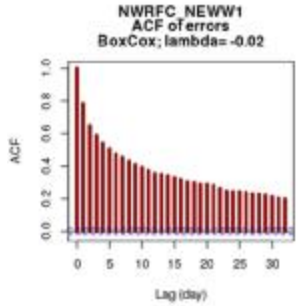


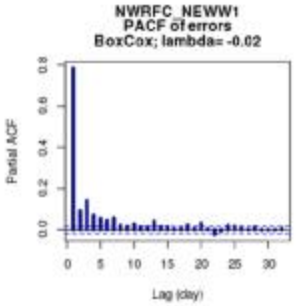


First order differencing

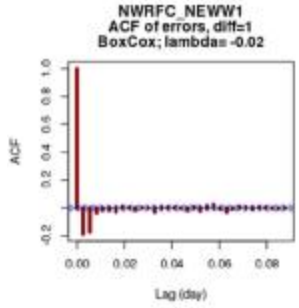
KPSS Test for Level Stationarity:
Stationary

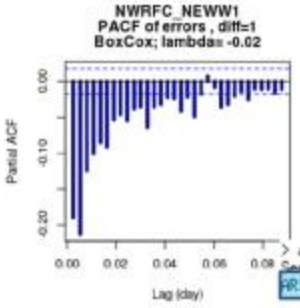
```
> auto.arima(qerrorboxcox2ts)
Series: qerrorboxcox2ts
ARIMA(2,1,4)

Coefficients:
         ar1     ar2      ma1     ma2     ma3     ma4
      0.4303  0.4251  -0.7543  -0.5297  0.1810  0.1113
s.e.  0.0513  0.0465   0.0517   0.0639  0.0144  0.0147

sigma^2 estimated as 0.02065:  log likelihood=6281.08
AIC=-12544.16   AICc=-12544.15   BIC=-12492.58
```





KPSS Test for Level Stationarity:
Non-stationary





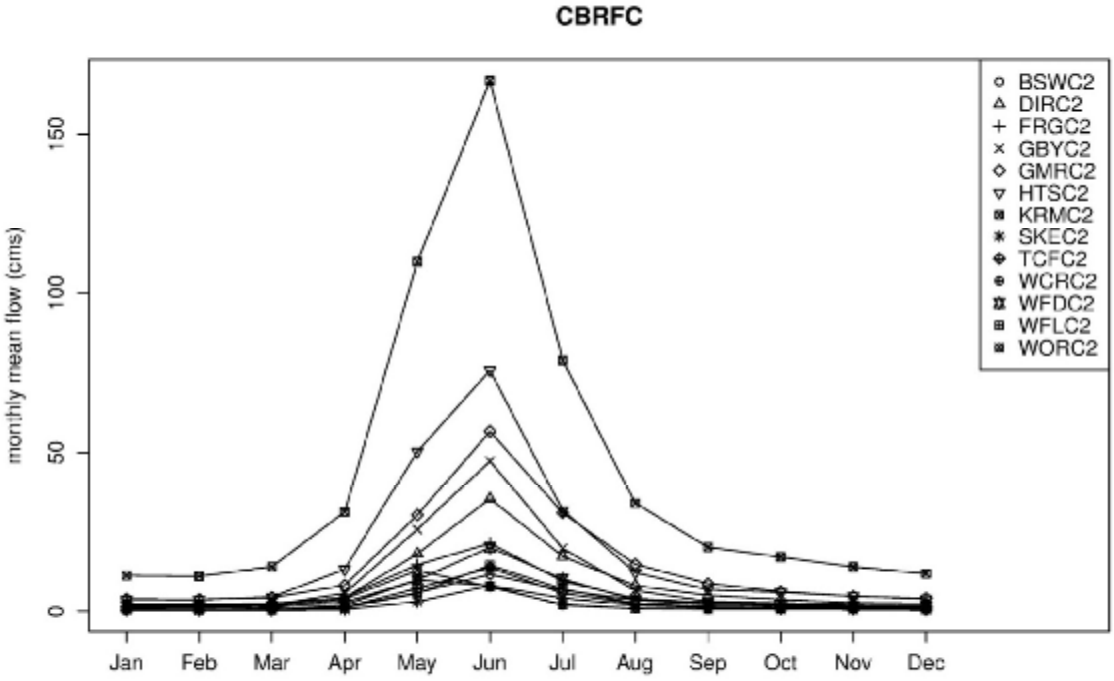
First order differencing

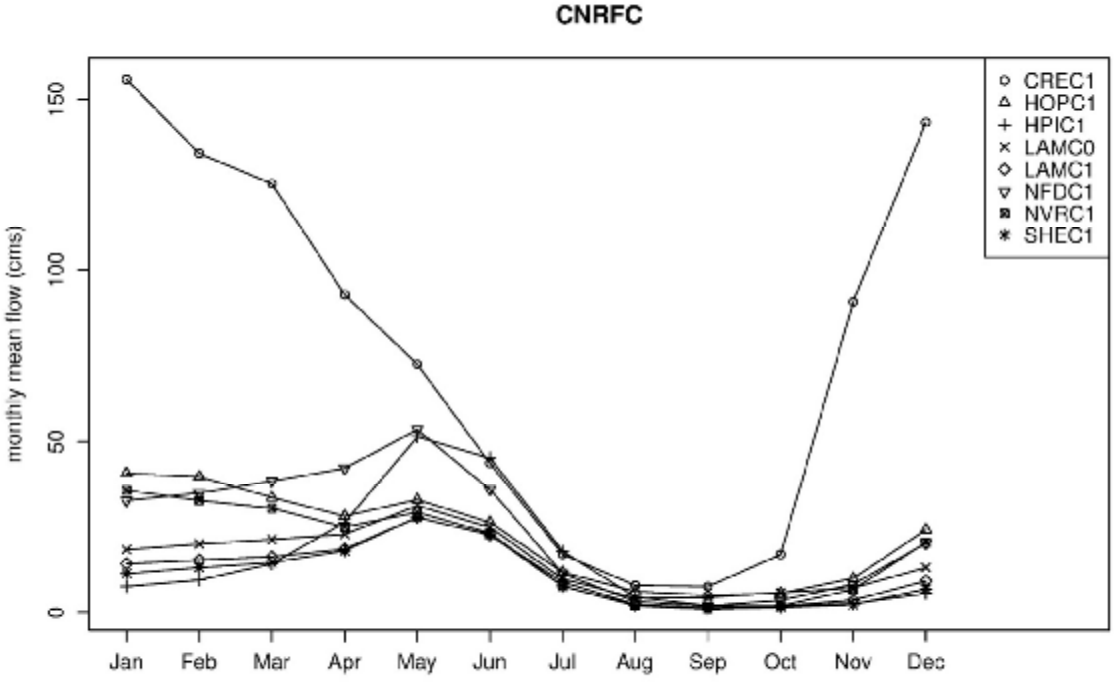
KPSS Test for Level Stationarity:
Stationary

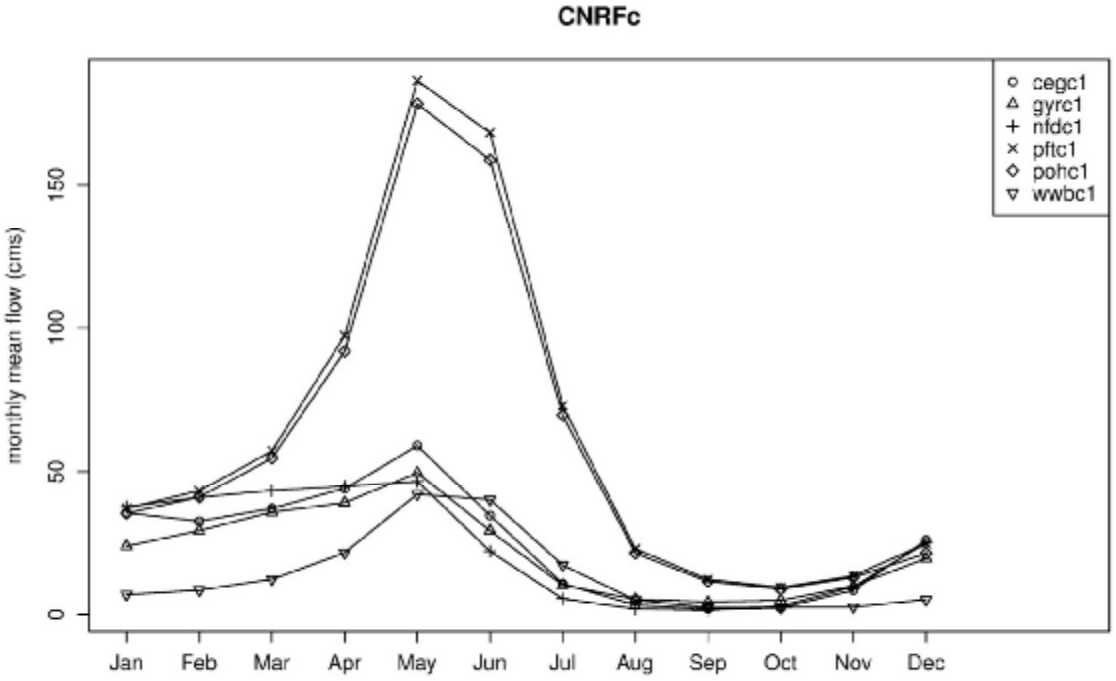
```
> auto.arima(qerrorboxcox2ts)
Series: qerrorboxcox2ts
ARIMA(2,1,4)

Coefficients:
          ar1     ar2      ma1      ma2     ma3     ma4
      0.4303  0.4251  -0.7543  -0.5297  0.1810  0.1113
s.e.  0.0513  0.0465   0.0517   0.0639  0.0144  0.0147

sigma^2 estimated as 0.02065:  log likelihood=6281.08
AIC=-12544.36   AICc=-12544.35   BIC=-12492.58
```

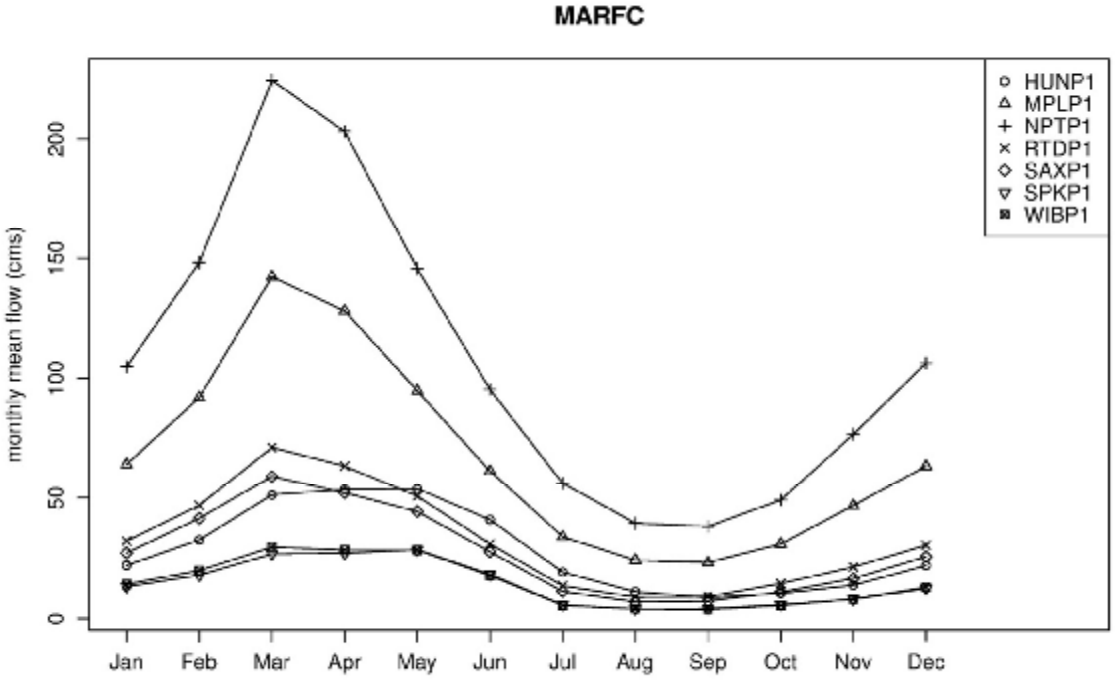
Appendix C

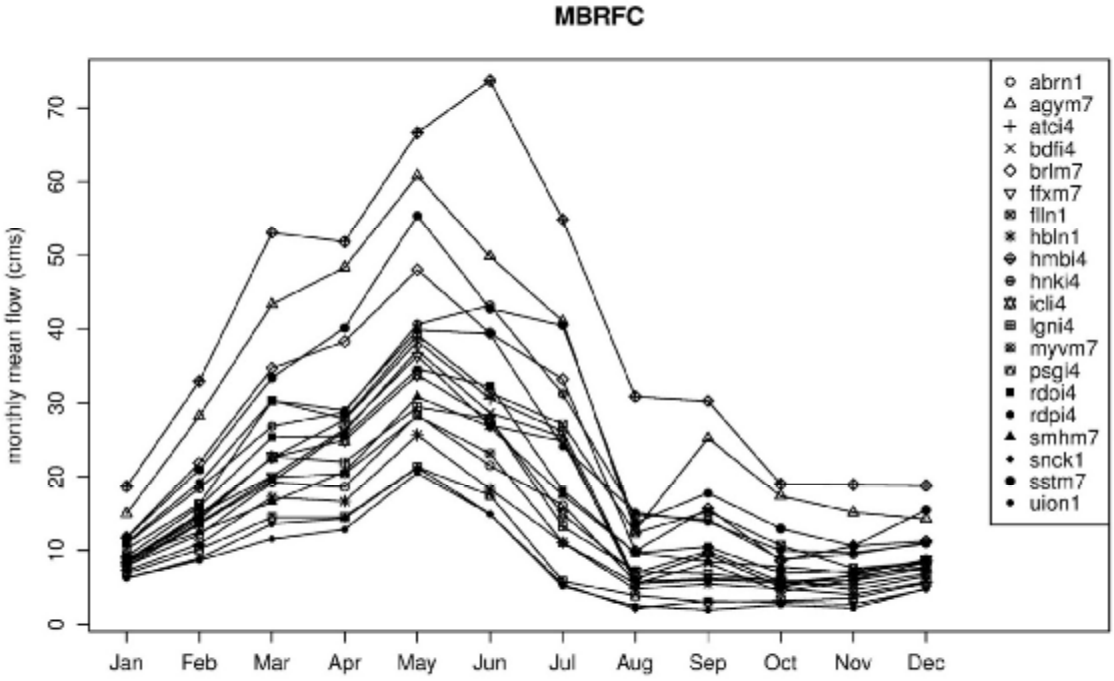
Monthly Mean Flow

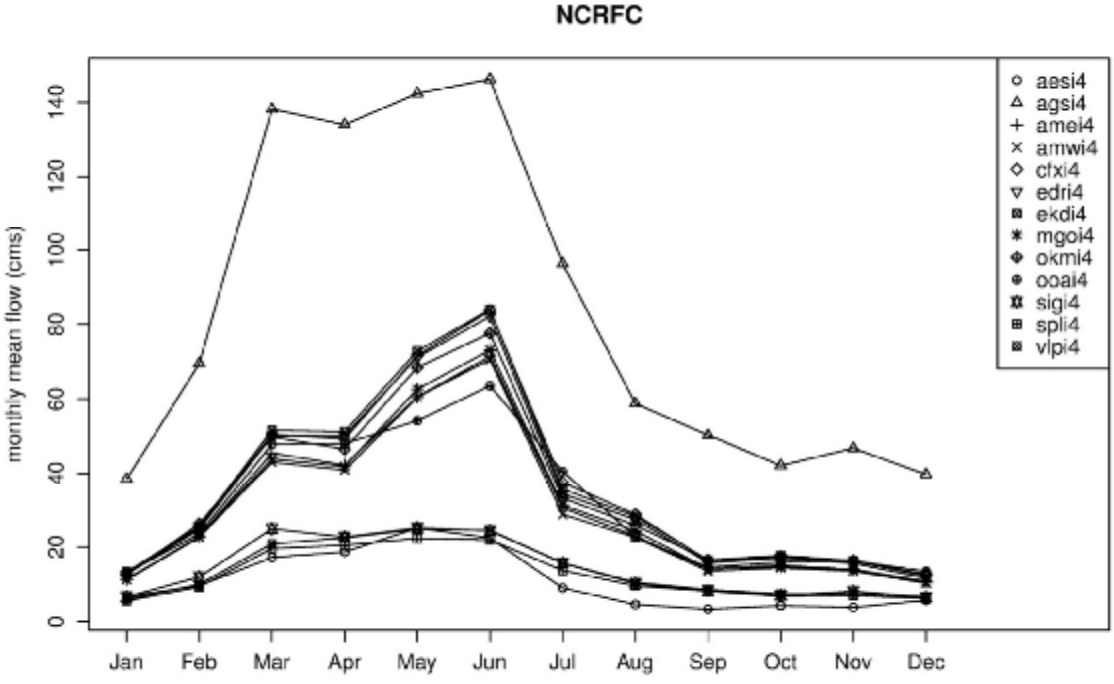

CBRFC
monthly mean flow (cms)
0
50
100
150
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
BSWC2
DIRC2
FRGC2
GBYC2
GMRC2
HTSC2
KRMC2
SKEC2
TCFC2
WCRC2
WFDC2
WFLC2
WORC2


CNRFC
monthly mean flow (cms)
0
50
100
150
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
CREC1
HOPC1
HPIC1
LAMC0
LAMC1
NFDC1
NVRC1
SHEC1

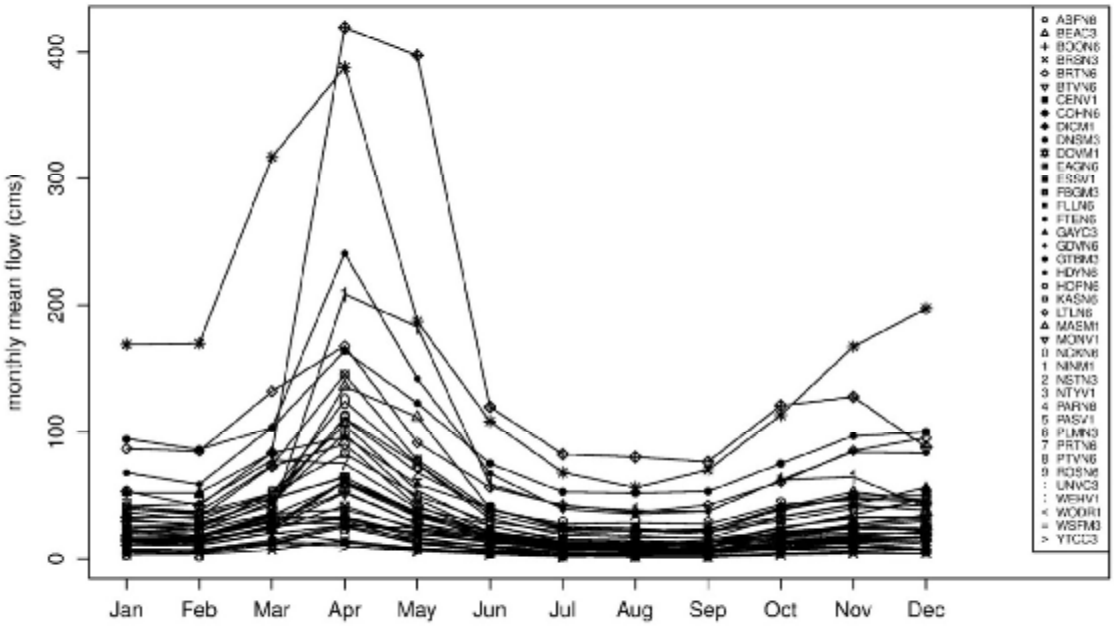

CNRFc
monthly mean flow (cms)
0
50
100
150
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
cegc1
gyrc1
nfdc1
pftc1
pohc1
wwbc1


MARFC
monthly mean flow (cms)
0
50
100
150
200
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
HUNP1
MPLP1
NPTP1
RTDP1
SAXP1
SPKP1
WIBP1

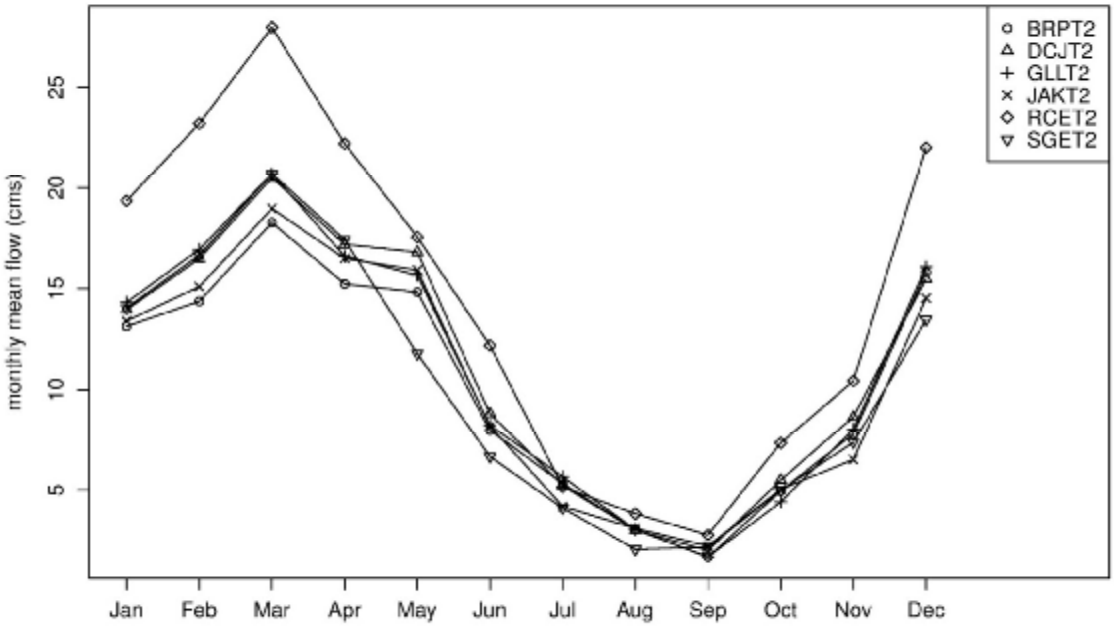

MBRFC
monthly mean flow (cms)
0
10
20
30
40
50
60
70
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
abrn1
agym7
atci4
bdfi4
brlm7
ffxm7
flln1
hbln1
hmbl4
hnki4
icli4
lgni4
myvm7
psgi4
rdoi4
rdpi4
smhm7
snck1
sstm7
uion1


NCRFC
monthly mean flow (cms)
0
20
40
60
80
100
120
140
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
aesi4
agsi4
amei4
amwi4
cfxi4
edri4
ekdi4
mgoi4
okmi4
ooai4
sigi4
spli4
vlpi4

NERFC

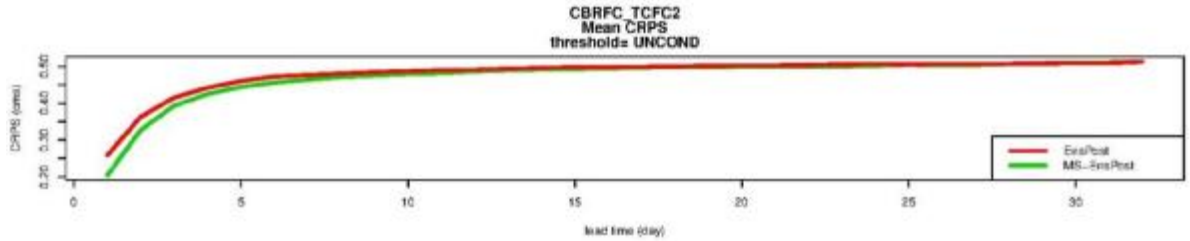

monthly mean flow (cms)
0
100
200
300
400
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
ABFN8
BEAC3
BOON6
BRSN3
BRTN6
BTVN6
CENV1
COHN6
DICM1
DNSM3
DOVM1
EAGN6
ESSV1
FBGM3
FLLN6
FTEN6
GAYC3
GDMN6
GTRM3
HDYN6
HOFN6
KASN6
LTLN6
MASM1
MONV1
NCKN6
NINM1
NSTN3
NTYV1
PARN6
PASV1
PLMN3
PRTN6
PTYN6
RDSN6
UNVC3
WBHV1
WOOR1
WSFM3
YTCC3

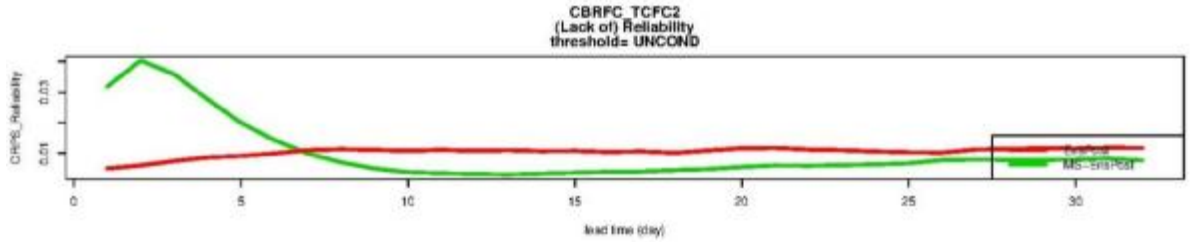
WGRFC

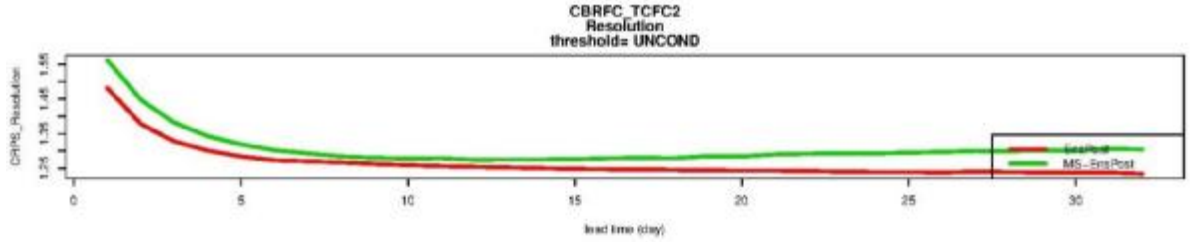

monthly mean flow (cms)
5
10
15
20
25
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
BRPT2
DCJT2
GLLT2
JAKT2
RCET2
SGET2

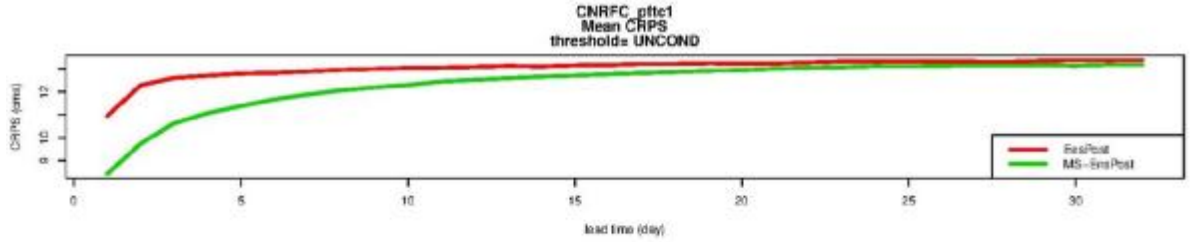
Appendix D

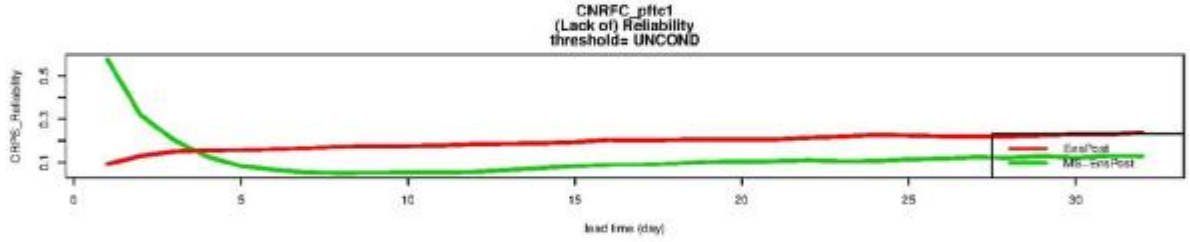
Mean CRPS and its decomposition into reliability and resolution vs. lead time for selected basins

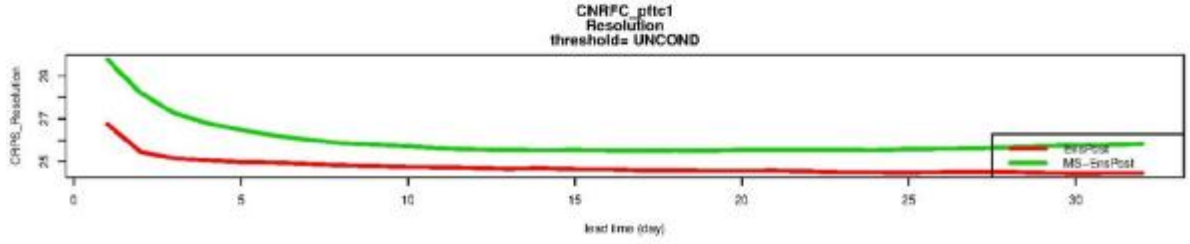

CBRFC_TCFC2
Mean CRPS
threshold= UNCOND
CRPS (cms)
lead time (day)
EnsPost
MS-EnsPost

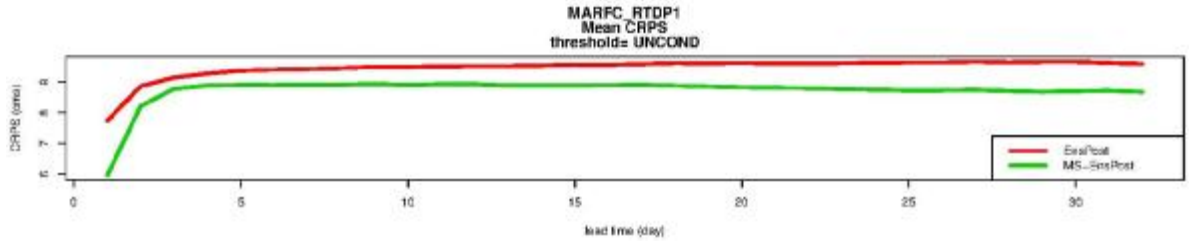

CBRFC_TCFC2
(Lack of) Reliability
threshold= UNCOND
CRPS_Reliability
lead time (day)
EnsPost
MS-EnsPost

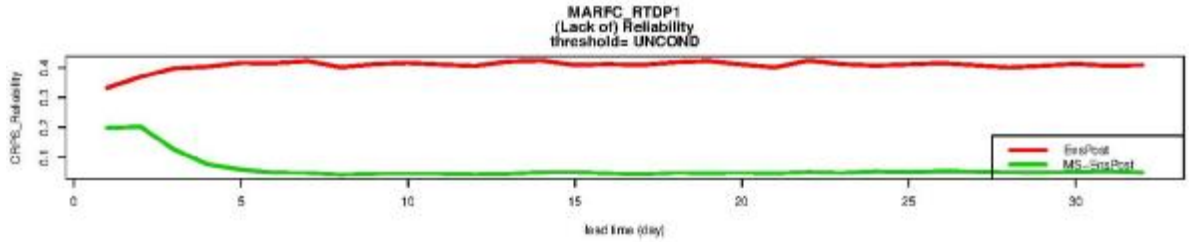

CBRFC_TCFC2
Resolution
threshold= UNCOND
CRPS_Resolution
lead time (day)
EnsPost
MS-EnsPost

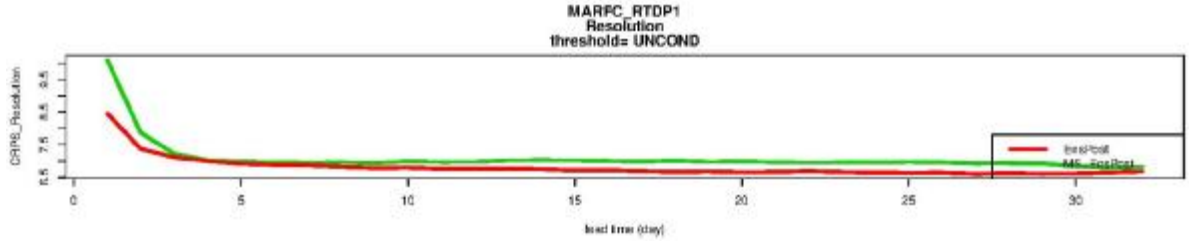

CNRFC_pftc1
Mean CRPS
threshold= UNCOND
CRPS (cms)
lead time (day)
EnsPost
MS-EnsPost

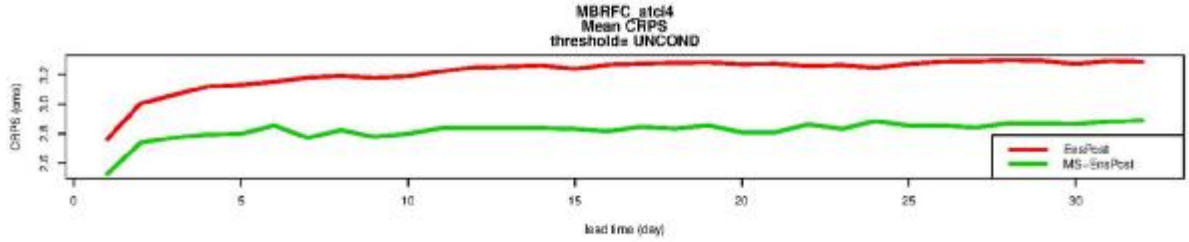

CNRFC_pftc1
(Lack of) Reliability
threshold= UNCOND
CRPS_Reliability
lead time (day)
EnsPost
MS-EnsPost

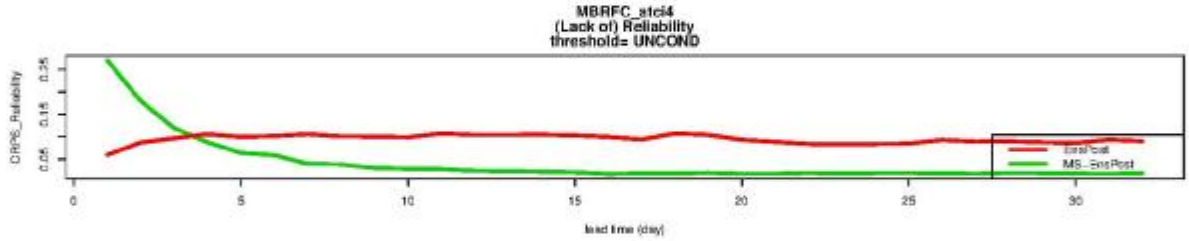

CNRFC_pftc1
Resolution
threshold= UNCOND
CRPS_Resolution
lead time (day)
EnsPost
MS-EnsPost

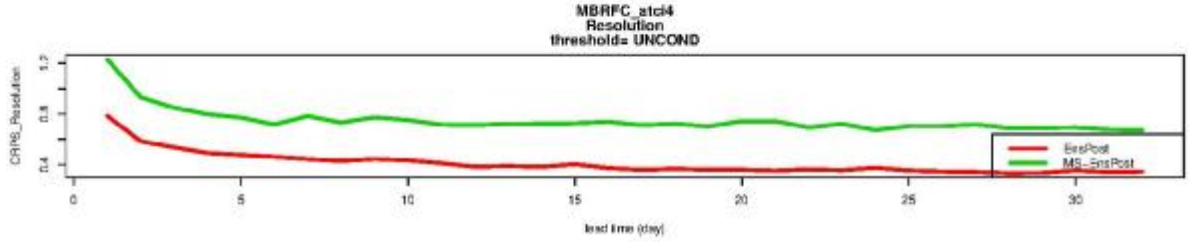

MARFC_RTDP1
Mean CRPS
threshold= UNCOND
CRPS (cms)
lead time (day)
EnsPost
MS-EnsPost

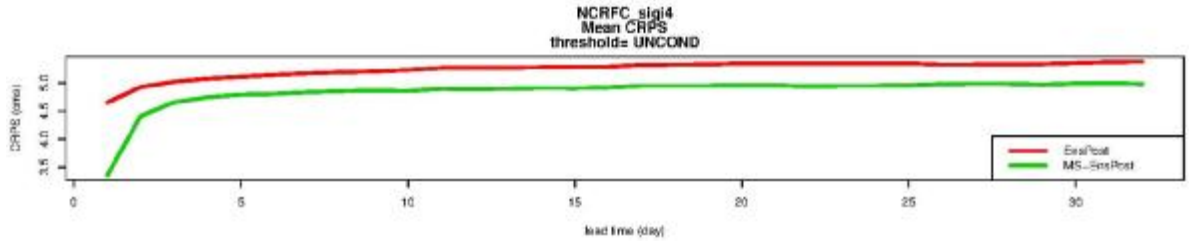

MARFC_RTDP1
(Lack of) Reliability
threshold= UNCOND
CRPS_Reliability
lead time (day)
EnsPost
MS-EnsPost

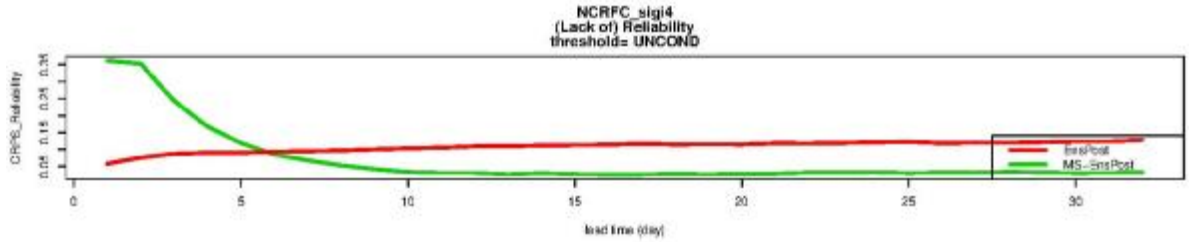

MARFC_RTDP1
Resolution
threshold= UNCOND
CRPS_Resolution
lead time (day)
EnsPost
MS-EnsPost

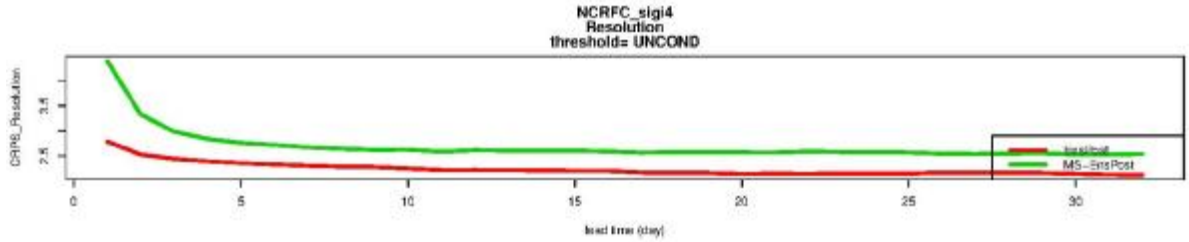

MBRFC_atcl4
Mean CRPS
threshold= UNCOND
CRPS (cms)
lead time (day)
EnsPost
MS-EnsPost

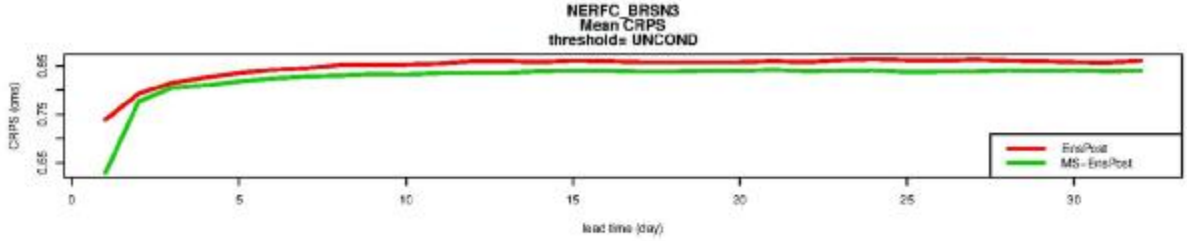

MBRFC_atcl4
(Lack of) Reliability
threshold= UNCOND
CRPS_Reliability
lead time (day)
EnsPost
MS-EnsPost

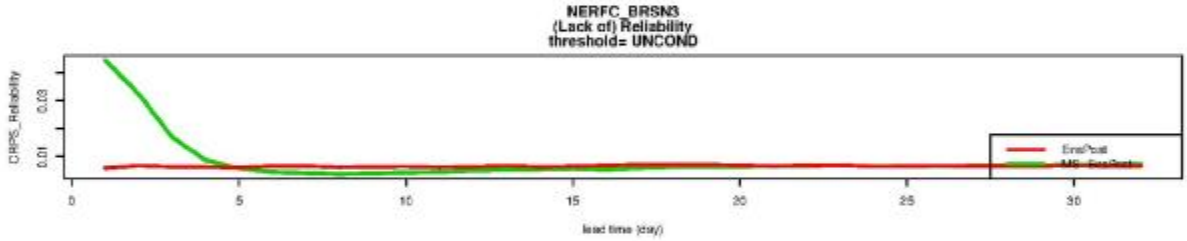

MBRFC_atcl4
Resolution
threshold= UNCOND
CRPS_Resolution
lead time (day)
EnsPost
MS-EnsPost

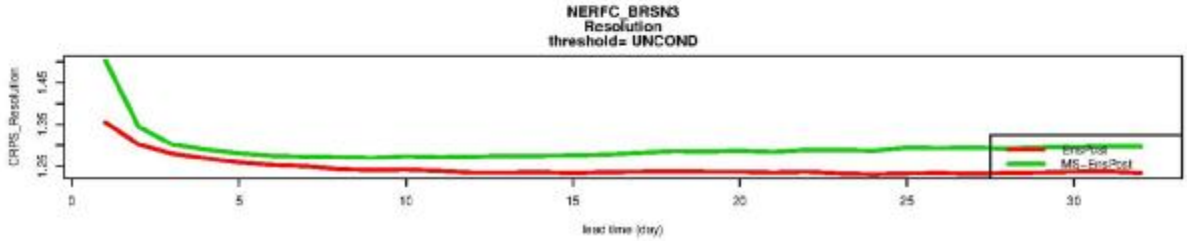

NCRFC_sigi4
Mean CRPS
threshold= UNCOND
CRPS (cms)
lead time (day)
EnsPost
MS-EnsPost

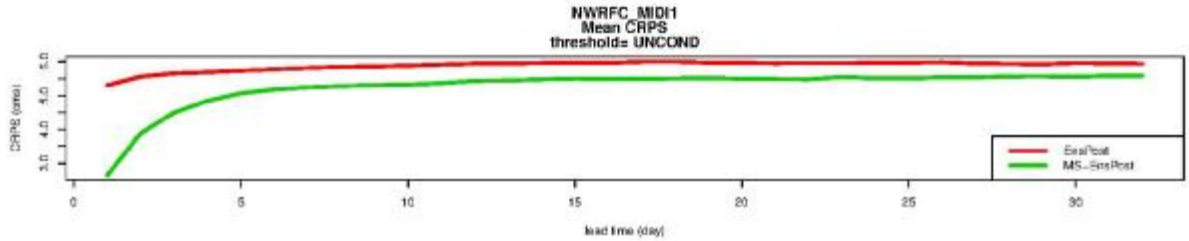

NCRFC_sigi4
(Lack of) Reliability
threshold= UNCOND
CRPS_Reliability
lead time (day)
EnsPost
MS-EnsPost

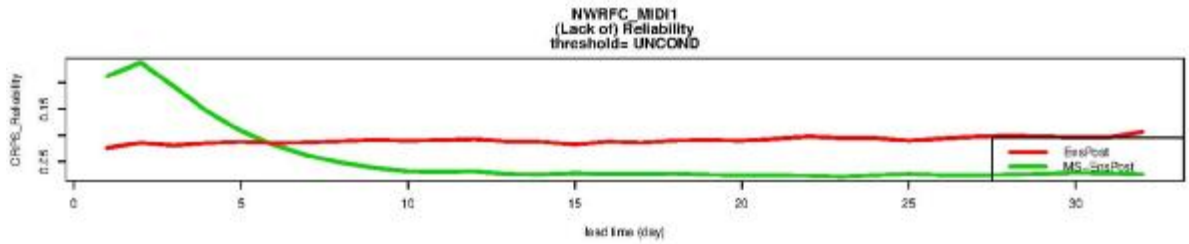

NCRFC_sigi4
Resolution
threshold= UNCOND
CRPS_Resolution
lead time (day)
EnsPost
MS-EnsPost

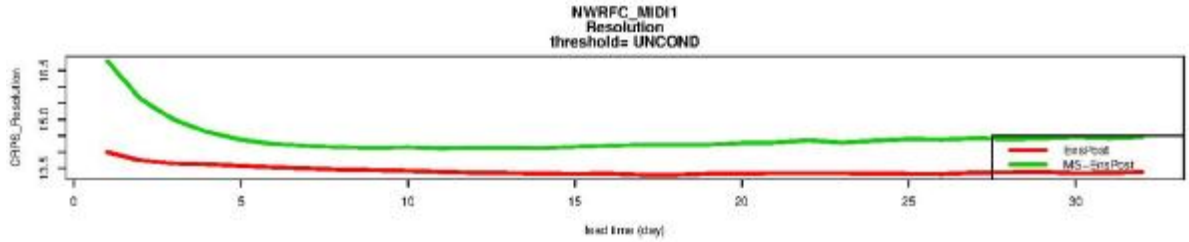

NERFC_BRSN3
Mean CRPS
threshold= UNCOND
CRPS (cms)
lead time (day)
EnsPost
MS-EnsPost

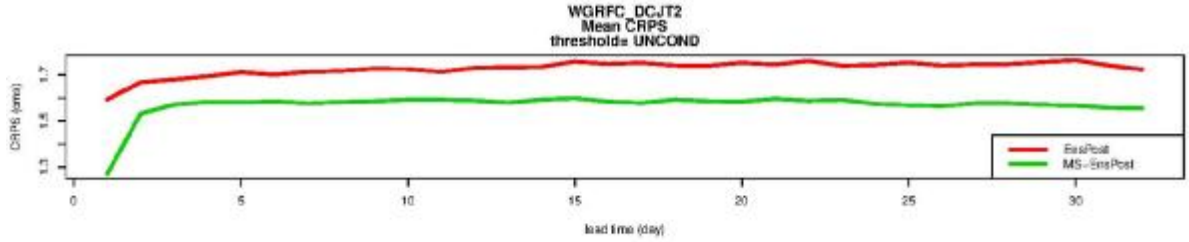

NERFC_BRSN3
(Lack of) Reliability
threshold= UNCOND
CRPS_Reliability
lead time (day)
EnsPost
MS-EnsPost

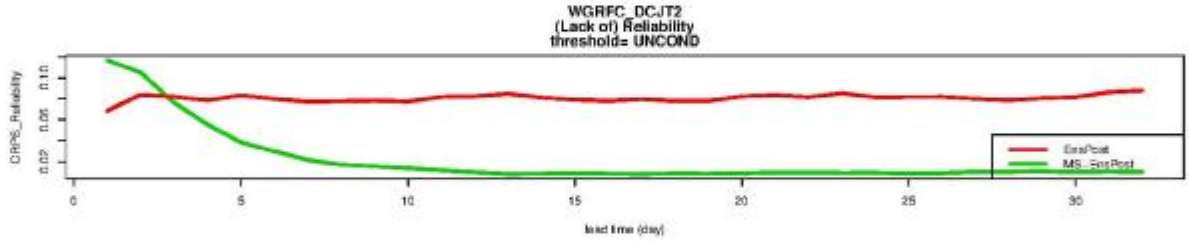

NERFC_BRSN3
Resolution
threshold= UNCOND
CRPS_Resolution
lead time (day)
EnsPost
MS-EnsPost

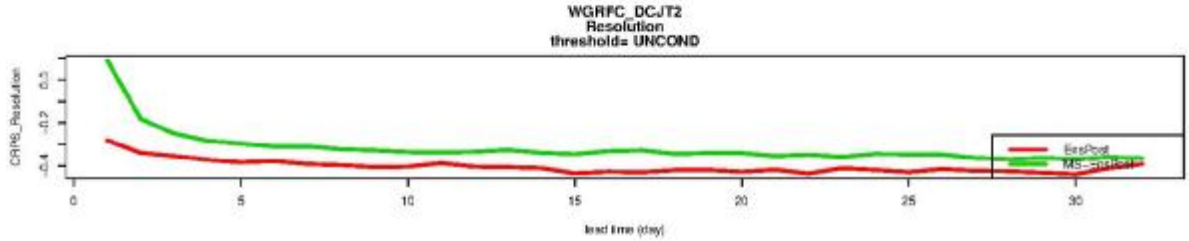

NWRFC_MIDI1
Mean CRPS
threshold= UNCOND
CRPS (cms)
lead time (day)
EnsPost
MS-EnsPost


NWRFC_MIDI1
(Lack of) Reliability
threshold= UNCOND
CRPS_Reliability
lead time (day)
EnsPost
MS-EnsPost


NWRFC_MIDI1
Resolution
threshold= UNCOND
CRPS_Resolution
lead time (day)
EnsPost
MS-EnsPost

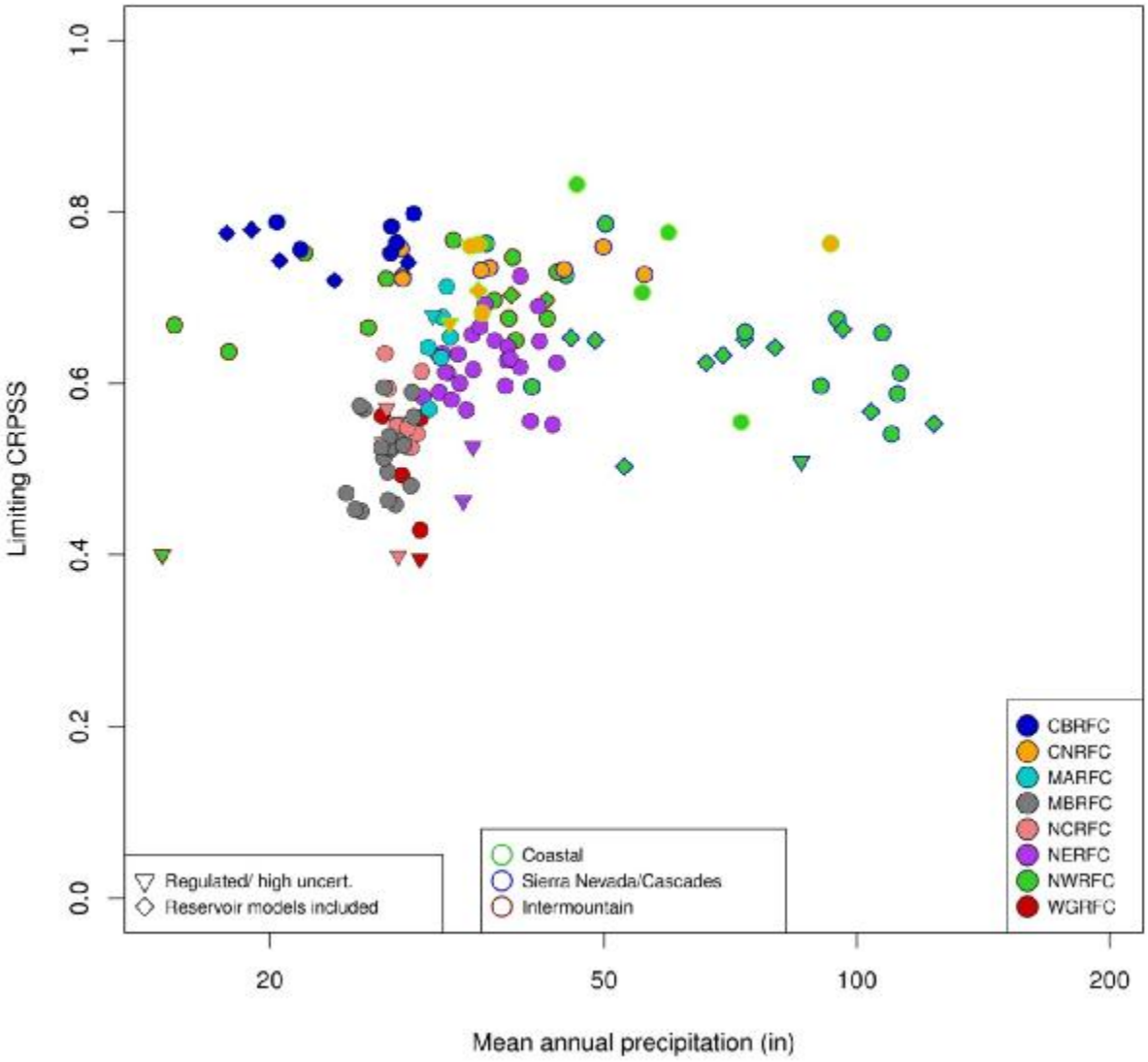

WGRFC_DCJT2
Mean CRPS
threshold= UNCOND
CRPS (cms)
lead time (day)
EnsPost
MS-EnsPost

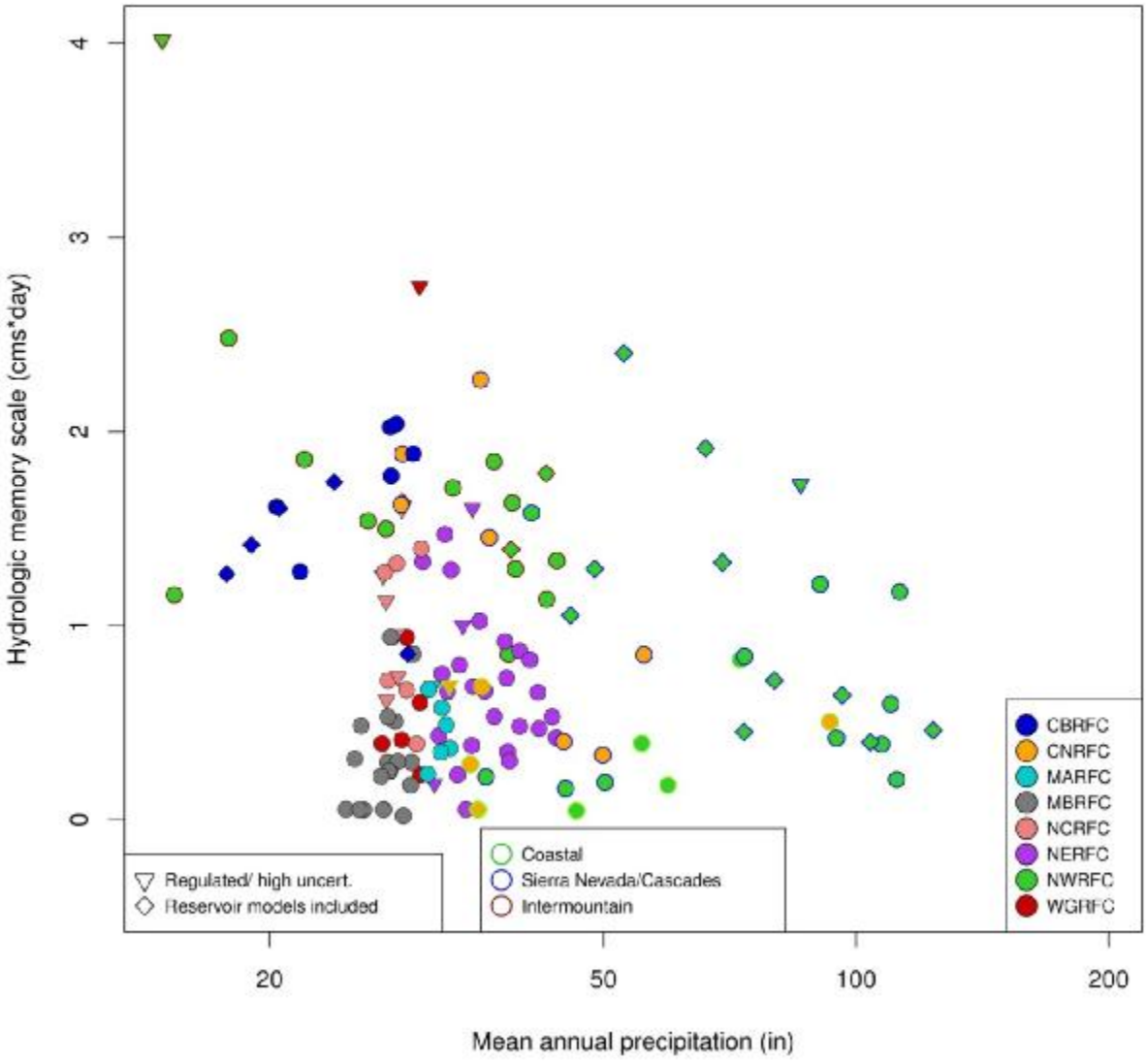

WGRFC_DCJT2
(Lack of) Reliability
threshold= UNCOND
CRPS_Reliability
lead time (day)
EnsPost
MS-EnsPost

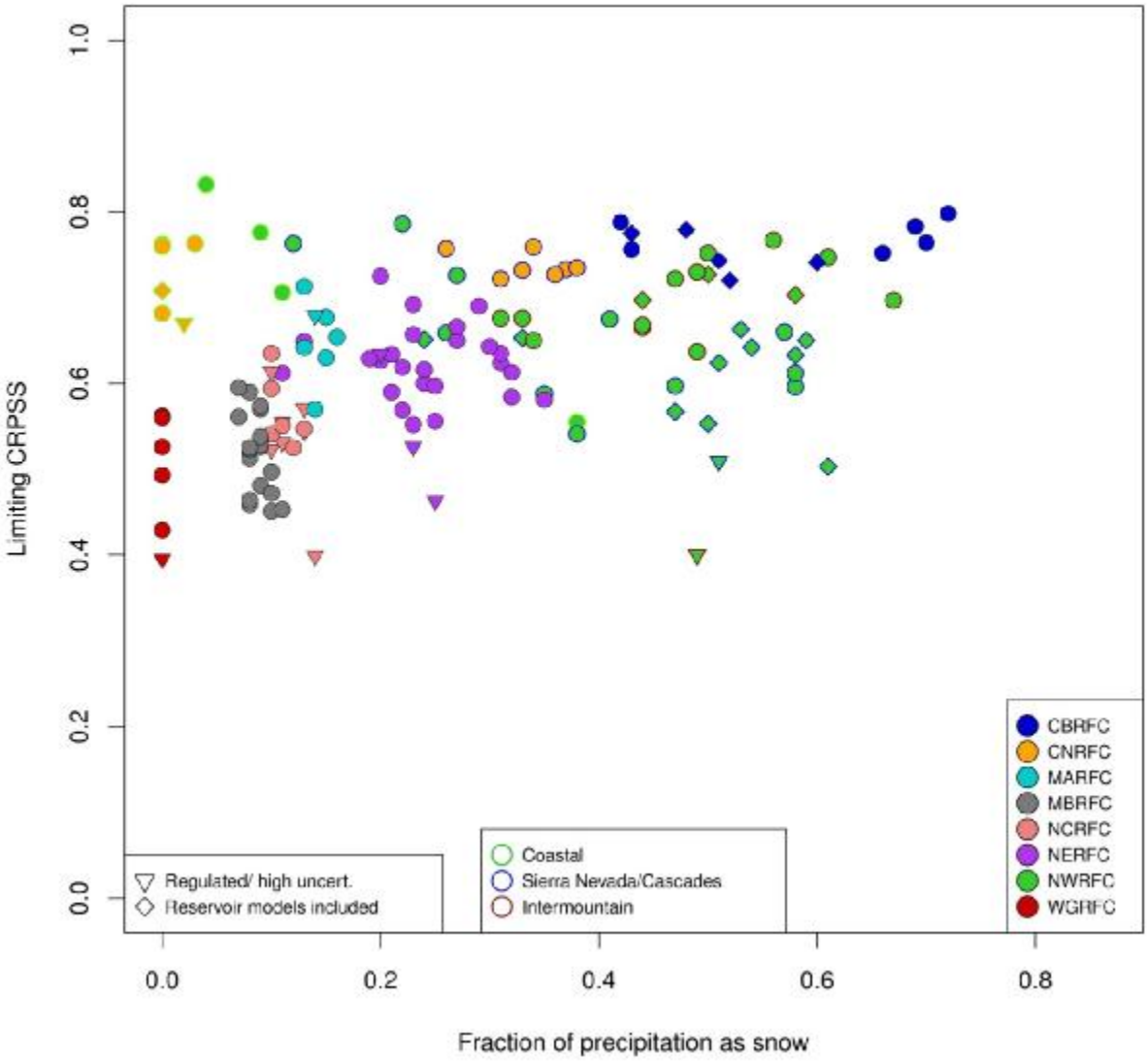

WGRFC_DCJT2
Resolution
threshold= UNCOND
CRPS_Resolution
lead time (day)
EnsPost
MS-EnsPost

Appendix E

CRPSS attributes vs. hydroclimatic indices


Limiting CRPSS
0.0
0.2
0.4
0.6
0.8
1.0
20
50
100
200
Mean annual precipitation (in)
Regulated/ high uncert.
Reservoir models included
Coastal
Sierra Nevada/Cascades
Intermountain
CBRFC
CNRFC
MARFC
MBRFC
NCRFC
NERFC
NWRFC
WGRFC


Hydrologic memory scale (cms*day)
0
1
2
3
4
20
50
100
200
Mean annual precipitation (in)
Regulated/ high uncert.
Reservoir models included
Coastal
Sierra Nevada/Cascades
Intermountain
CBRFC
CNRFC
MARFC
MBRFC
NCRFC
NERFC
NWRFC
WGRFC


Limiting CRPSS
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
Fraction of precipitation as snow
Regulated/ high uncert.
Reservoir models included
Coastal
Sierra Nevada/Cascades
Intermountain
CBRFC
CNRFC
MARFC
MBRFC
NCRFC
NERFC
NWRFC
WGRFC

References

Adams III, T., 2016: Flood forecasting in the United States NOAA/National Weather Service. *Flood Forecasting: A global perspective*, Academic Press,, 249-310, https://doi.org/10.1016/B978-0-12-801884-2.00010-4.

Ajami, N. K., Q. Duan, and S. Sorooshian, 2007: An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resour. Res.*, **43**, W01403, https://doi.org/10.1029/2005WR004745.

Anderson, E. A., 1973: National Weather Service river forecast system: Snow accumulation and ablation model. National Oceanographic and Atmospheric Administration, Silver Springs, Md., Tech. Memo NWS HYDRO-17

Baldwin, M. P., D. B. Stephenson, D. W. Thompson, T. J. Dunkerton, A. J. Charlton, and A. O'Neill, 2003: Stratospheric memory and skill of extended-range weather forecasts. *Science*, **301,** 636-640, https://doi.org/10.1126/science.1087143.

Bengtsso, L., and K. I. Hodges, 2006: A note on atmospheric predictability. *Tellus A: Dynamic Meteorology and Oceanography*, **58,** 154-157, https://doi.org/10.1111/j.1600-0870.2006.00156.x.

Bennett, C., R. Stewart, and J. Lu, 2014: Autoregressive with exogenous variables and neural network short-term load forecast models for residential low voltage distribution networks. *Energies*, **7,** 2938-2960, https://doi.org/10.3390/en7052938.

Berghuijs, W.R., Sivapalan, M., Woods, R.A. and Savenije, H.H., 2014: Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. *Water Resour. Res*, **50(7)**, 5638-5661, https://doi.org/10.1002/2014WR015692.

Bjørnar Bremnes, J., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132,** 338-347, https://doi.org/10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2.

Blöschl, G., and M. Sivapalan, 1995: Scale issues in hydrological modelling: a review. *Hydrol. Process.*, **9,** 251-290, https://doi.org/10.1002/hyp.3360090305.

Bogner, K., K. Liechti, and M. Zappa, 2016: Post-processing of stream flows in Switzerland with an emphasis on low flows and floods. *Water*, **8,** 115, https://doi.org/10.3390/w8040115.

Borgomeo, E., J. W. Hall, F. Fung, G. Watts, K. Colquhoun, and C. Lambert, 2014: Risk-based water resources planning: Incorporating probabilistic nonstationary climate uncertainties. *Water Resour. Res.*, **50,** 6850-6873, https://doi.org/10.1002/2014WR015558.

Box, G., and D. Cox, 1964: An analysis of transformations. *J. R. Stat. Soc. Series B (Methodol.)* **26**, 211-252.

Box, G. E., and G. M. Jenkins, 1976: Time series analysis: forecasting and control. San Francisco, CA: Holden Day, 575 pp.

Breidenbach, J.P., M. A. Fortune, D.-J. Seo, and P. Tilles, 2001: Multisensor precipitation estimation for use by River Forecast Centers during heavy rainfall events. [Available online at http://ams. confex.com/ams/annual2001/techprogram/paper_18658.htm.]

Breidenbach, J.P., M. A. Fortune, D.-J. Seo, and M. Fortune, 2002: Multisensor precipitation estimation for use by the National Weather Service River Forecast Centers. Preprints, 16th Conf. on Hydrology, Orlando, FL, Amer. Meteor. Soc., 2.5. [Available online at http://ams. confex.com/ams/annual2002/techprogram/paper_26937.htm.]

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2 .

Brown, J. D., and D.-J. Seo, 2010: A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. *J. Hydrometeor.*, **11,** 642-665, https://doi.org/10.1175/2009JHM1188.1.

Budyko, M. I., D. H. Miller, and D. H. Miller, 1974: *Climate and life.* Vol. 508, Academic press New York.

Burnash, R. J., R. L. Ferral, and R. A. McGuire, 1973: A generalized streamflow simulation system, conceptual modeling for digital computers. *U.S. Department of Commerce National Weather Service and State of California Department of Water Resources.*

Butts, M.B., Payne, J.T., Kristensen, M. and Madsen, H., 2004: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *J. Hydrol.*, **298**, 242-266, https://doi.org/10.1016/j.jhydrol.2004.03.042.

Carlu, M., F. Ginelli, V. Lucarini, and A. Politi, 2019: Lyapunov analysis of multiscale dynamics: the slow bundle of the two-scale Lorenz 96 model. *Nonlin. Processes Geophys.*, **26**, 73-89, https://doi.org/10.5194/npg-26-73-201

Chapman, D. G., 1956: Estimating the parameters of a truncated gamma distribution. *The Ann. Math. Statist.*,**27,** 498-506. [Available online at https://www.jstor.org/stable/2237007]

Chow, V. T., D. R. Maidment, and L. W. Mays, 1988: Unit Hydrograph. In: Applied Hydrology McGraw-Hill Series In Water Resources and Environmental Engineering. *McGraw Hill,* 100-118, https://doi.org/10.1029/89EO00083.

Cloke, H., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375,** 613-626, https://doi.org/10.1016/j.jhydrol.2009.06.005.

Coccia, G., and E. Todini, 2011: Recent developments in predictive uncertainty assessment based on the model conditional processor approach. *Hydrol. Earth Syst. Sci.*, **15,** 3253-3274, https://doi.org/10.5194/hess-15-3253-2011.

Damon, J., and S. Guillas, 2002: The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics*, **13,** 759-774, https://doi.org/10.1002/env.527.

Day, G.N., 1985: Extended streamflow forecasting using NWSRFS. J. Water Resour. Plann. Manage., 111(2), 157-170, https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157).

Demargne, J., and Coauthors, 2014: The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Amer. Meteor. Soc.*, **95,** 79-98, https://doi.org/10.1175/BAMS-D-12-00081.1

Demeritt, D., S. Nobert, H. Cloke, and F. Pappenberger, 2010: Challenges in communicating and using ensembles in operational flood forecasting. *Meteorol. Appl.*, **17,** 209-222, https://doi.org/10.1002/met.194.

Doherty, J., and D. Welter, 2010: A short exploration of structural noise. *Water Resour. Res.*, **46**, W05525, https://doi.org/10.1029/2009WR008377.

Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian, 2007: Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.*, **30,** 1371-1386, https://doi.org/10.1016/j.advwatres.2006.11.014.

Dunne, T., and R. D. Black, 1970: Partial area contributions to storm runoff in a small New England watershed. *Water Resour. Res.*, **6,** 1296-1311, https://doi.org/10.1029/WR006i005p01296.

Engman, E. T., and A. S. Rogowski, 1974: A partial area model for storm flow synthesis. *Water Resour. Res.*, **10,** 464-472, https://doi.org/10.1029/WR010i003p00464.

Erickson, M., 1996: Medium-range prediction of PoP and Max/Min in the era of ensemble model output. *Conference on weather analysis and forecasting, Norfolk VA. Am. Meteorol. Soc.*, J35-J38.

Fedora, M., and R. Beschta, 1989: Storm runoff simulation using an antecedent precipitation index (API) model. *J. Hydrol.*, **112,** 121-133, https://doi.org/10.1016/0022-1694(89)90184-4.

Fowler, K., M. Peel, A. Western, and L. Zhang, 2018: Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function. *Water Resour. Res.*, **54,** 3392-3408, https://doi.org/10.1029/2017WR022466.

Freer, J., K. Beven, and B. Ambroise, 1996: Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resour. Res.*, **32,** 2161-2173, https://doi.org/10.1029/95WR03723.

Freeze, R. A., 1972: Role of subsurface flow in generating surface runoff: 2. Upstream source areas. *Water Resour. Res.*, **8,** 1272-1283, https://doi.org/10.1029/WR008i005p01272.

Gan, T. Y., E. M. Dlamini, and G. F. Biftu, 1997: Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. *J. Hydrol.*, **192,** 81-103, https://doi.org/10.1016/S0022-1694(96)03114-9.

Gebregiorgis, A., and F. Hossain, 2011: How much can a priori hydrologic model predictability help in optimal merging of satellite precipitation products? *J. Hydrometeor.*, **12,** 1287-1298, https://doi.org/10.1175/JHM-D-10-05023.1.

Georgakakos, K. P., D.-J. Seo, H. Gupta, J. Schaake, and M. B. Butts, 2004: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *J. Hydrol.*, **298,** 222-241, https://doi.org/10.1016/j.jhydrol.2004.03.037.

Gijsbers, P., L. Cajina, C. Dietz, J. Roe, and E. Welles, 2009: CHPS-an NWS development to enter the interoperability era. *AGU Fall Meeting Abstracts*.[Available online at http://adsabs.harvard.edu/abs/2009AGUFMIN11A1041G]

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133,** 1098-1118, https://doi.org/10.1175/MWR2904.1.

Groves, D.G., Yates, D., Tebaldi, C., 2008: Developing and applying uncertain global climate change projections for regional water management planning, *Water Resour. Res.*, **44**, W12413, https://doi.org/10.1029/2008WR006964

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye, 2012: Towards a comprehensive assessment of model structural adequacy. *Water Resour. Res.*, **48**, W08301, https://doi.org/10.1029/2011WR011044.

Hall, J., and E. Borgomeo, 2013: Risk-based principles for defining and managing water security. *Philos. Trans. Royal Soc. A.*, **371,** 20120407, https://doi.org/10.1098/rsta.2012.0407.

Hall, J. W., and Coauthors, 2019: Risk-based water resources planning in practice: a blueprint for the water industry in England. *Water Environ. J.*, 1-14, https://doi.org/10.1111/wej.12479.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126,** 711-724, https://doi.org/10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132,** 1434-1447, https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

Hartman, R., M. Fresch, and E. Wells, 2015: National Weather Service (NWS) Implementation of the Hydrologic Ensemble Forecast Service. *AGU Fall Meeting Abstracts.* [http://adsabs.harvard.edu/abs/2015AGUFM.H52A..01H]

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2002: Verification of probabilistic streamflow forecasts, *IIHR Report No. 427, IIHR-Hydroscience & Engineering and Dept. of Civil and Environ. Eng., The Univ. of Iowa, Iowa City, IA*, 125pp.[Available online at https://www.iihr.uiowa.edu/wp-content/uploads/2013/06/IIHR427.pdf]

Hashino, T., A. Bradley, and S. Schwartz, 2006: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci*, **3,** 561-594.[Available online at https://hal.archives-ouvertes.fr/hal-00298675]

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15,** 559-570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Horton, R. E., 1933: The role of infiltration in the hydrologic cycle. *Trans. Amer. Geophys. Union*, **14,** 446-460, https://doi.org/10.1029/TR014i001p00446 .

Hou, D., K. Mitchell, Z. Toth, D. Lohmann, and H. Wei, 2009: The effect of large-scale atmospheric uncertainty on streamflow predictability. *J. Hydrometeor.*, **10,** 717-733, https://doi.org/10.1175/2008JHM1064.1.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast verification: a practitioner's guide in atmospheric science.* John Wiley & Sons, 292 pp.

Kim, S., and Coauthors, 2016: Integrating Ensemble Forecasts of Precipitation and Streamflow into Decision Support for Reservoir Operations in North Central Texas. *AGU Fall Meeting Abstracts*. [Available online at http://adsabs.harvard.edu/abs/2016AGUFM.H53I..08K]

Kim, S., and Coauthors, 2018: Assessing the skill of medium-range ensemble precipitation and streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS) for the Upper Trinity River Basin in North Texas. *J. Hydrometeor.*, **19,** 1467-1483, https://doi.org/10.1175/JHM-D-18-0027.1.

Kim, S. M., B. L. Benham, K. M. Brannan, R. W. Zeckoski, and J. Doherty, 2007: Comparison of hydrologic calibration of HSPF using automatic and manual methods. *Water Resour. Res.*, **43**, W01402, https://doi.org/10.1029/2006WR004883.

Koenker, R., and G. Bassett, 1978: Regression quantiles. *Econometrika*, **46**, 33–50, https://doi.org/10.2307/1913643.

Kolmogorov, A., 1933: Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, **4**, 83-91.

Krause, P., D. Boyle, and F. Bäse, 2005: Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.*, **5,** 89-97, https://doi.org/10.5194/adgeo-5-89-2005.

Krzysztofowicz, R., 1999: Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resour. Res.*, **35,** 2739-2750, https://doi.org/10.1029/1999WR900099.

Krzysztofowicz, R., and K. S. Kelly, 2000: Hydrologic uncertainty processor for probabilistic river stage forecasting. *Water Resour. Res.*, **36,** 3265-3277, https://doi.org/10.1029/2000WR900108.

Krzysztofowicz, R., and H. D. Herr, 2001: Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation-dependent model. *J. Hydrol.*, **249,** 46-68, https://doi.org/10.1016/S0022-1694(01)00412-7.

Kumar, P., 2011: Typology of hydrologic predictability. *Water Resour. Res.*, **47**, W00H05, https://doi.org/10.1029/2010WR009769.

Lee, H., Y. Liu, J. Brown, J. Ward, A. Maestre, M.A. Fresch, H. Herr, and E. Wells, 2018: Validation of ensemble streamflow forecasts from the Hydrologic Ensemble Forecast Service (HEFS). *AGU Fall Meeting*, Washington, D.C.

Legates, D. R., and G. J. McCabe Jr, 1999: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.*, **35,** 233-241, https://doi.org/10.1029/1998WR900018.

Li, J., and R. Ding, 2011: Temporal–spatial distribution of atmospheric predictability limit by local dynamical analogs. *Mon. Wea. Rev.*, **139,** 3265-3283, https://doi.org/10.1175/MWR-D-10-05020.1.

Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di, 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, **4,** e1246, https://doi.org/10.1002/wat2.1246.

Li, Z., J. C. McWilliams, K. Ide, and J. D. Farrara, 2015: A multiscale variational data assimilation scheme: Formulation and illustration. *Mon. Wea. Rev.*, **143,** 3804-3822, https://doi.org/10.1175/MWR-D-14-00384.1.

Limon, R., 2019: Improving multi-reservoir water supply system operation using ensemble forecasting and global sensitivity analysis, Dissertation, The University of Texas at Arlington, 164 pp, http://hdl.handle.net/10106/28115.

Loague, K., and J. E. VanderKwaak, 2004: Physics-based hydrologic response simulation: Platinum bridge, 1958 Edsel, or useful tool. *Hydrol. Process.*, **18,** 2949-2956, https://doi.org/10.1002/hyp.5737.

Madadgar, S., H. Moradkhani, and D. Garen, 2014: Towards improved post-processing of hydrologic forecast ensembles. *Hydrol. Process.*, **28,** 104-122, https://doi.org/10.1002/hyp.9562.

Mahanama, S., B. Livneh, R. Koster, D. Lettenmaier, and R. Reichle, 2012: Soil moisture, snow, and seasonal streamflow forecasts in the United States. *J. Hydrometeor.*, **13,** 189-203, https://doi.org/10.1175/JHM-D-11-046.1.

Marimo, P., Kaplan, T.R., Mylne, K., Sharpe, M., 2015: Communication of uncertainty in temperature forecasts, Wea. Forecasting, 30, 5-22, https://doi.org/10.1175/WAF-D-14-00016.1.

Maurer, E. P., and D. P. Lettenmaier, 2004: Potential effects of long-lead hydrologic predictability on Missouri River main-stem reservoirs. *J. Clim.*, **17,** 174-186, https://doi.org/10.1175/1520-0442(2004)017<0174:PEOLHP>2.0.CO;2.

McMillan, H., Jackson, B., Clark, M., Kavetski, D., Woods, R., 2011: Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, J. Hydrol., **400**, 83-94, https://doi.org/10.1016/j.jhydrol.2011.01.026.

Mendoza, P., A. Wood, E. Clark, B. Nijssen, M. Clark, M.-H. Ramos, and N. Voisin, 2016: Improving medium-range ensemble streamflow forecasts through statistical post-processing. *AGU Fall Meeting Abstracts*.[Available online at http://adsabs.harvard.edu/abs/2016AGUFM.H51F1547M]

Milly, P. C., J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer, 2008: Stationarity is dead: Whither water management? *Science*, **319,** 573-574, https://doi.org/10.1126/science.1151915.

Mizukami, N., and Coauthors, 2017: Towards seamless large-domain parameter estimation for hydrologic models. *Water Resour. Res.*, **53,** 8020-8040, https://doi.org/10.1002/2017WR020401.

Montanari, A., and A. Brath, 2004: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resour. Res.*, **40**, W01106, https://doi.org/10.1029/2003WR002540.

Nash, J. E., and J. V. Sutcliffe, 1970: River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.*, **10,** 282-290, https://doi.org/10.1016/0022-1694(70)90255-6.

National Research Council of the National Academies, 2006: Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts. *The National Academies Press*, 124 pp.

National Weather Service Office of Water Prediction (NWS OWP/NOAA), 2008a: Joint Reservoir Regulation (RES-J) Model. Accessed 30 Jan, 2017. http://www.nws.noaa.gov/oh/hrl/general/chps/Models/ Joint_Reservoir_Regulation.pdf..

National Weather Service Office of Water Prediction (NWS OWP/NOAA), 2008b: Single Reservoir Regulation (RES-SNGL) Model. Accessed 30 Jan, 2017. http://www.nws.noaa.gov/oh/hrl/general/chps/Models/ Single_Reservoir_Regulation.pdf.

National Weather Service, 2015: Ensemble Postprocessor (EnsPost) User's Manual, Office of Hydrologic Development, Silver Spring, MD, Accessed 04 April, 2015. [Available online at http://www.nws.noaa.gov/oh/hrl/general/indexdoc.html.]

Norouzi, A., H. Habibi, B. Nazari, S. J. Noh, D.-J. Seo, and Y. Zhang, 2018: Toward parsimonious modeling of frequency of areal runoff from heavy-to-extreme precipitation in large urban areas under changing conditions: a derived moment approach. *Stoch. Environ. Res. Risk Assess.* 1-19. [Available online at https://link.springer.com/article/10.1007/s00477-019-01698-8]

Piani, C., J. Haerter, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.,* **99,** 187-192. [Available online at https://link.springer.com/article/10.1007/s00704-009-0134-9]

Perica, S., 1998: Integration of meteorological forecasts/climate outlooks into an ensemble streamflow prediction system. 78th Annual AMS Meeting, Phoenix, AZ, [Available online at https://ams.confex.com/ams/pdfpapers/54775.pdf]

Perica, S., Marcus, M., Schaake, J. and Seo, D., 1999: Accounting for hydrologic model errors in ensemble streamflow prediction. Preprint volume, 14th Conf. on Hydrol., Dallas, TX, Jan 10-15.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133,** 1155-1174, https://doi.org/10.1175/MWR2906.1.

Rakovec, O., and Coauthors, 2018: Multiscale and multivariate evaluation of water fluxes and states over European river basins. *J. Hydrometeor.*, **17**, 287-307, https://doi.org/10.1175/JHM-D-15-0054.1.

Regonda, S., and Seo, 2008: Statistical post processing streamflow ensembles to improve reliability over a wide range of time scales. *2nd CPPA PIs meeting,* , Silver Spring, MD, Aug.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks, 2010: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resour. Res.*, **46**, W05521, https://doi.org/10.1029/2009WR008328.

Robert, C. P., 1995: Simulation of truncated normal variables. *Stat. Comput.*, **5,** 121-125. [Available online at https://link.springer.com/article/10.1007/BF00143942]

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, **55,** 16-30, https://doi.org/10.3402/tellusa.v55i1.12082.

Samaniego, L., R. Kumar, and S. Attinger, 2010: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resour. Res.*, **46**, W05523, https://doi.org/10.1029/2008WR007327.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX: the hydrological ensemble prediction experiment. *Bull. Amer. Meteor. Soc.*, **88,** 1541-1548, https://doi.org/10.1175/BAMS-88-10-1541.

Schlosser, C. A., and P. C. D. Milly, 2002: A model-based investigation of soil moisture predictability and associated climate predictability. *J. Hydrometeor.*, **3,** 483-501, https://doi.org/10.1175/1525-7541(2002)003<0483:AMBIOS>2.0.CO;2.

Schweppe, F. C., 1973: *Uncertain dynamic systems.* Prentice Hall.

Seo, D.-J., H. Herr, and J. Schaake, 2006: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci. Discussions*, **3,** 1987-2035. [Available online at https://www.hydrol-earth-syst-sci-discuss.net/3/1987/2006/hessd-3-1987-2006.pdf]

Seo, D.-J., S. Kim, B. Alizadeh, M. Ghazvinian, and H. Lee, 2019: Improving precipitation ensembles for heavy-to-extreme events and streamflow post-processing for short-to-long ranges. Final Report to the NOAA/NWS/Office of Water Prediction.

Sharma, S., R. Siddique, S. Reed, P. Ahnert, P. A. Mendoza Zúñiga, and A. Mejia, 2018: Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system , *Hydrol. Earth Syst. Sci*., **22**, 1831-1849, https://doi.org/ 10.5194/hess-22-1831-2018.

Simmons, A., R. Mureau, and T. Petroliagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Q.J.R. Meteorol. Soc.*, **121,** 1739-1771, https://doi.org/10.1002/qj.49712152711.

Sittner, W. T., C. E. Schauss, and J. C. Monro, 1969: Continuous hydrograph synthesis with an API-type hydrologic model. *Water Resour. Res.*, **5,** 1007-1022, https://doi.org/10.1029/WR005i005p01007.

Smirnov, N., 1948: Table for estimating the goodness of fit of empirical distributions. *The Ann. Math. Statist.*, **19**, 279-281.

Smith, T., L. Marshall, and B. McGlynn, 2014: Calibrating hydrologic models in flow-corrected time. *Water Resour. Res.*, **50,** 748-753, https://doi.org/10.1002/2013WR014635.

Stephenson, D., C. Coelho, F. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: a unified framework for the combination of multi-model weather and climate predictions. *Tellus A: Dynamic Meteorology and Oceanography*, **57,** 253-264, https://doi.org/10.3402/tellusa.v57i3.14664.

Subbey, S., Christie, M., Sambridge, M., 2004: Prediction under uncertainty in reservoir modeling, J. of Petrol. Sci. Eng., **44**, 143-153, https://doi.org/10.1016/j.petrol.2004.02.011.

Wells, E., 2017: National Weather Service (NWS) Hydrologic Ensemble Forecast Service. [Available online at https://ispuw.uta.edu/nsf/downloads/2017_Workshop/Presentations/1_3_NWS_Wells.pdf]

Westra, S., and A. Sharma, 2010: An upper limit to seasonal rainfall predictability? *J. Clim.*, **23,** 3332-3351, https://doi.org/10.1175/2010JCLI3212.1.

Wilks, D. S., 2006: Comparison of ensemble-MOS methods in the Lorenz'96 setting. *Meteorol. Appl.*, **13,** 243-256, https://doi.org/10.1017/S1350482706002192.

Wood, A. W., and J. C. Schaake, 2008: Correcting errors in streamflow forecast ensemble mean and spread. *J. Hydrometeor.*, **9,** 132-148, https://doi.org/10.1175/2007JHM862.1.

Zhao, L., Q. Duan, J. Schaake, A. Ye, and J. Xia, 2011: A hydrologic post-processor for ensemble streamflow predictions. *Adv. Geosci.*, **29,** 51-59, https://doi.org/10.5194/adgeo-29-51-2011.

Zhu, Y., and Y. Luo, 2015: Precipitation calibration based on the frequency-matching method. *Wea. Forecasting*, **30,** 1109-1124, https://doi.org/10.1175/WAF-D-13-00049.1.