2021

IMPROVING OPERATIONAL HYDROLOGIC FORECASTING VIA CONDITIONAL BIAS-PENALIZED MULTI-SENSOR PRECIPITATION ESTIMATION AND MULTI-MODEL STREAMFLOW PREDICTION

Ali Jozaghi

Recommended Citation

Jozaghi, Ali, "IMPROVING OPERATIONAL HYDROLOGIC FORECASTING VIA CONDITIONAL BIAS-PENALIZED MULTI-SENSOR PRECIPITATION ESTIMATION AND MULTI-MODEL STREAMFLOW PREDICTION" (2021). *Civil Engineering Dissertations*. 444.
https://mavmatrix.uta.edu/civilengineering_dissertations/444

Improving Operational Hydrologic Forecasting via Conditional Bias-Penalized Multi-Sensor Precipitation Estimation and Multi-Model Streamflow Prediction

By

ALI JOZAGHI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at
The University of Texas at Arlington
May, 2021

Arlington, Texas

Supervising Committee:
  Dr. Dong-Jun Seo, Supervising Professor
  Dr. Yu Zhang
  Dr. Michelle Hummel
  Dr. Seann reed

Acknowledgments

Throughout the writing of this dissertation, I have received a great deal of support and assistance. First, it is a genuine pleasure to express my deep sense of gratitude to my advisor Dr. Dong-Jun Seo for his tremendous support during my Ph.D. study. Without his support, I would not be even close to where I am now. I would also like to thank my Ph.D. committee members, Drs. Yu Zhang, Michelle Hummel, and Seann Reed for their comments.

I thank profusely all my current and former colleagues at the Hydrology and Water Resources Laboratory who helped me with this research including, Drs. Sunghee Kim, Seong Jin Noh, Reza Ahmad Limon, Babak Alizadeh, and Haojing Shen, Messrs. Mohammad Nabatian and Vaghef Ghazvinian, and Mrs. Soona Habibi. I would also like to express my gratitude to many others who helped me during my Ph.D. study, including Dr. Edwin Welles of Deltares USA, for helpful comments in model calibration. Finally, it is my privilege to thank my family and friends for all their support and encouragement throughout my research period.

The material presented in this document is based upon work supported in part by the NOAA’s Joint Technology Transfer Initiative Program under Grant NA16OAR4590232, and the NSF under Grant CyberSEES-1442735. This support is gratefully acknowledged.

May 07, 2021

Abstract

Improving Operational Hydrologic Forecasting via Conditional Bias-Penalized Multi-sensor Precipitation Estimation and Bayesian Multi-model Averaging of River Forecasts

Ali Jozaghi, Ph.D.
The University of Texas at Arlington, 2021
Supervising Professor: Dong-Jun Seo

Many multivariate analysis techniques involve minimizing mean square error (MSE) or error variance under unbiasedness. In the presence of observation error, variance minimization tends to introduce negative and positive biases, or conditional bias (CB), over the upper and lower tails of the predictands, respectively. This work describes and evaluates three multivariate merging techniques, 1) adaptive conditional bias-penalized cokriging (CBPCK), 2) conditional bias-penalized Multiple Linear Regression (CBP-MLR), and 3) conditional bias-penalized Bayesian Model Averaging (CBP-BMA) which implements CBP-MLR in place of MLR. CBPCK and CBP-MLR minimize a linearly weighted sum of errors squared and the sum of the Type-II error squared, thereby addressing the type-II CB explicitly. CBPCK is applied to improve multisensor precipitation estimation using rain gauge data and remotely-sensed quantitative precipitation estimates (QPE). The remotely-sensed QPEs used are radar-only and radar-satellite-fused estimates. For comparative evaluation, true validation is carried out over the continental United States (CONUS) for Sep 13-30, 2015, and Oct 7-9, 2016. The hourly gauge data, radar-only QPE, and satellite QPE used are from the Hydrometeorological Automated Data System, Multi-Radar Multi-Sensor System, and Self-Calibrating Multivariate Precipitation Retrieval (SCaMPR),

respectively. For radar-satellite fusion, conditional bias-penalized Fisher estimation is used. The reference merging technique compared is ordinary cokriging (OCK) used in the National Weather Service Multisensor Precipitation Estimator. It is shown that, beyond the reduction due to mean-field bias (MFB) correction, both OCK and adaptive CBPCK additionally reduce the unconditional root mean square error (RMSE) of radar-only QPE by 9 to 16% over the CONUS for the two periods, and that adaptive CBPCK is superior to OCK for estimation of hourly amounts exceeding 1 mm. When fused with the MFB-corrected radar QPE, the MFB-corrected SCaMPR QPE for Sep 2015 reduces the unconditional RMSE of the MFB-corrected radar by 4 and 6% over the entire and western half of the CONUS, respectively, but is inferior to the MFB-corrected radar for estimation of hourly amounts exceeding 7 mm. Adaptive CBPCK should hence be favored over OCK for estimation of significant amounts of precipitation despite the higher computational cost, and the SCaMPR QPE should be used selectively in multisensor QPE.

CBP-MLR and CBP-BMA are described and evaluated for improved multi-model streamflow prediction using several operationally produced streamflow forecasts. For comparative evaluation, 10-fold cross-validation is carried out over the NWS Middle Atlantic River Forecast Center's (MARFC) service area for the period of Jan 1, 2017, to Oct 29, 2018. The input streamflow forecasts used are the MARFC single-valued forecast, the Hydrologic Ensemble Forecast System (HEFS) ensemble forecast, the National Water Model (NWM) medium-range single-valued forecast, and the Meteorological Model-based Ensemble Forecast System (MMEFS) ensemble forecasts forced by the Global Ensemble Forecast System (GEFS), the North American Ensemble Forecast System (NAEFS), and the Short-Range Ensemble Forecast System (SREF). Whereas CBP-MLR improves prediction over tails, it degrades the performance near the median. To retain MLR-like performance near median while exploiting the ability of CBP-MLR to improve

prediction over tails, composite MLR (CompMLR), which linearly weight-averages the MLR and CBP-MLR estimates, is also developed and evaluated.

MLR and CBP-MLR do not consider uncertainty in the regression model. They typically choose a single model and fit it to the data. This approach disregards the uncertainty in the model selection which leads to overconfident inferences. To address the model uncertainty while improving performance for large flow, CBP-MLR is implemented in BMA to produce CBP-BMA. The proposed methods are applied for multi-model streamflow prediction using several operationally produced streamflow forecasts as predictors. The results for the MARFC's service area show that the relative performance among different input forecasts varies most significantly with the range of the verifying observed streamflow, and both CompMLR and CBP-BMA are generally superior to the best performing forecasts in the mean squared error sense under widely varying conditions of predictability and predictive skill.

Keywords: QPE, Conditional Bias, CBPCK, CBP-MLR, Streamflow, CompMLR, BMA, CBP-BMA

Table of Contents

List of Illustrations

Chapter 1. General Introduction

Water-related hazards are the most common among all natural hazards and pose large threats to people and their socio-economic well-being (Hassanzadeh et al., 2020; ICHARM, 2009; Noji and Lee, 2005). Of all-natural disasters, water-related disasters account for about 72% of the total economic losses; 26% of these water-related disasters are attributed to floods (ICHARM 2009). Globally, there have been 273 water-related disasters since 1980, the total cost of which approaches $2 trillion. In the US, floods have caused $43.6 billion in damages and claimed 132 lives among all disasters costing billion dollars or more. While this is staggering, it is only a portion of a larger trend for weather and climate disasters over the past 40 years. Due to climate change, deforestation, land-use change, population growth in flood-prone areas, and rising sea levels, the number of people vulnerable to water-related disasters is projected to increase to two billion by 2050 globally, which will inevitably increase water-related damages (ICHARM, 2009; Vogel et al., 2011).

Providing early warnings for water-related disasters, in particular floods, is a very cost-effective way to protect lives and properties and to minimize disruptions to socioeconomic activities. As such, much investment has been made over the years to provide timely and location-specific predictions of water-related hazards. With advances in science and technology, the amount and array of data that may provide helpful information for early hazard warnings are rapidly increasing. Methodological advances are necessary to translate them into actionable information and merge the hydrometeorological and hydrologic data from multiple sources in varying quality and quantity objectively and effectively.

In general, the accuracy of hydrologic predictions depends critically on that of precipitation input. Because decision support for warnings of water-related hazards increasingly calls for high-resolution information for improved spatiotemporal specificity, there is an ever-increasing demand for high-quality, high-resolution quantitative precipitation estimates (QPE). In many parts of the world, multisensor quantitative precipitation estimation (QPE) based on radar QPE products and rain gauge data is a common practice (Habib et al., 2009; Jozaghi et al., 2019; Kim et al., 2018; Kitzmiller et al., 2013; Seo et al., 2010; Tang et al., 2018; Young et al., 2000). While rain gauge networks are usually sparse, ground-based weather radars provide spatially continuous precipitation estimates at high temporal frequency. It is a must to merge rain gauge data and radar estimates in order to obtain precipitation estimates that are more accurate than those obtainable from either sensor alone. However, the coverage of ground-based weather radars is limited to land areas that are mainly free from beam blockages and offshore areas within the effective range of the radar. With the advancement of satellite remote sensing, satellite QPE is now routinely available in many parts of the world and may complement weather radar and rain gauges. The first element of this research addresses advancing multisensor estimation of precipitation using radar QPE and rain gauge data and multisensor fusion using radar and satellite QPE toward improving the accuracy of precipitation input, particularly for heavy-to-extreme events.

Streamflow is arguably the most critical predictand in operational hydrology and water management. With the fast-increasing availability of multiple streamflow forecasts from different sources in many parts of the world (Muhammad et al., 2018), objectively combining multiple forecasts in ways that will yield consistently superior forecast products is of great broad interest. Many machine learning techniques have been used for streamflow prediction, including support vector machines (SVM; Liu et al., 2021; Meng et al., 2019), artificial neural network (Napolitano et al., 2011; Sahoo et al., 2006), gradient-boosting decision tree (Lu et al., 2018), random forest (Schoppa et al., 2020), multiple linear regression (MLR; Jozaghi et al. 2021a), and Bayesian Model Averaging (BMA; Jozaghi et al., 2021b, 2020; Wang et al., 2017). The second and third elements of this work address advancing multimodel streamflow prediction using multiple linear regression (MLR; Freedman, 2005) and BMA (Duan et al., 2007a; Jennifer A. Hoeting et al., 1999; Madigan and Raftery, 1994; Raftery, 1995; Raftery et al., 2005), particularly for floods.

Many of the techniques for multisensor estimation or multimodel prediction involve minimizing mean square error (MSE) or error variance under unbiasedness. However, when the predictors are subject to significant uncertainty, variance minimization is subject to conditional bias (CB), often resulting in under- and overprediction of the predictand over the upper and lower tails, respectively (Fuller, 1987; Hausman, 2001; Seber and Wild, 1989). One of the most critical societal functions of hydrologic prediction is minimizing or reducing surprises, i.e., the Type-II errors. Because extreme events are rare, warning them requires an accurate forecast of the variables of interest over the tail end of their distributions, where the Type-II CB is particularly important. With urbanization and climate change, calibrating hydrologic models, including uncertainty models, is becoming an increasingly tenuous practice based on long periods of records (Alizadeh, 2019; Alizadeh et al., 2020b). In parts of the US, there is also growing evidence that the uncertainty in the initial conditions of operational hydrologic models may be increasing (Cotter, 2015). The record-breaking flooding of the Blanco River in central TX in May 2015 (Furl et al., 2018), which was attributed to rapid urbanization whose potential impact was assessed and predicted as early as 2007 (Curran, 2008), is one of many recent examples that encapsulate the challenges in hydrologic prediction in changing conditions. The above picture suggests that, for prediction of extremes, every datum, be it an observation or a model output, in the upper tail may potentially carry significant information content, and it is hence essential that multisensor or multimodel techniques be able to account for the Type-II error which may be obscured by small sample size.

A distinguishing aspect of information fusion for water-related hazards vs. that for many other applications is that for the former, performance in the upper tail is far more critical than that near median. Whereas the issue of CB has been widely recognized in statistics and econometrics, particularly in the context of least-squares regression (Fuller, 1987; Hausman, 2001; Seber and Wild, 1989), only a limited number of reports may be found elsewhere (Ciach et al., 2000; David et al., 1984; Emery, 2006; Guertin, 1984; Olea, 2012; Pan, 1998). Moreover, to the best of the author's knowledge, none addresses Type-2 CB explicitly. Recently, Seo (2013) introduced a linear geostatistical estimation technique, conditional bias-penalized kriging, which minimizes the sum of squared error and Type-II error squared to address specifically the detrimental effects of Type-II CB on spatial prediction. The technique has since been cast also in the context of Kalman filter for temporal prediction to yield CB-penalized Kalman filter (CBPKF; Seo et al. 2018a, b; Shen et al. 2019) and, in ensemble form, CB-penalized ensemble KF (CBEnKF; Lee et al., 2019; Shen et al., 2020)

With the above backdrop, the overarching objective of this work is to advance information fusion for multisensor estimation of heavy-to-extreme precipitation and multimodel streamflow prediction, particularly for floods. This work consists of three elements. The first element newly develops adaptive conditional bias-penalized cokriging (ACBPCK) for merging radar QPE and rain gauge observations and applies Simple Estimation (Rafieeinasab et al., 2015) for the fusion of radar and satellite QPE products for improved spatial prediction of heavy-to-extreme precipitation. The second element casts CB-penalized optimal estimation in the context of MLR to newly develop CB-penalized MLR (CBP-MLR) for improved multimodel prediction of large flows. MLR and CBP-MLR are then combined to yield composite MLR (CompMLR), which constitutes a parsimonious nonlinear regression comprising two linear regressions. In the third element, CBP-MLR is used in place of MLR in BMA with Bayesian regression linear model with Zellner's g-prior (Zellner, 1986) to yield CBP-BMA for improved prediction of large flows. This dissertation is organized as follows. Ch 2 is a reproduction of Jozaghi et al. (2019) and describes the adaptive CBPCK and fusion algorithms and presents the evaluation results. Ch 3 is a reproduction of Jozaghi et al. (2021a) and describes CBP-MLR and CompMLR for multimodel streamflow prediction and apply to the NWS Middle Atlantic River Forecast Center’s (MARFC) service area. Ch 4 describes the CBP-BMA, which is then applied to the MARFC’s service area for multimodel streamflow prediction. Ch 5 provides the general conclusions and future research recommendations.

Chapter 2. CBPCK: Theory and Application to Multisensor Precipitation Estimation

2.1 Introduction

Accurate real-time quantitative precipitation estimation (QPE) is a prerequisite for precise water forecasting. With the widespread use of weather radar systems, multisensor QPE using ground-based radar and rain gauge data is routine in many parts of the world. Given the increasing availability of various real-time satellite QPE products (AghaKouchak et al. 2011; Habib et al. 2009, 2012; Huffman et al. 2007, 2017; Joyce et al. 2004; Okamoto et al. 2005; Sorooshian et al. 2000; Turk and Miller 2005; Vicente et al. 1998), effective utilization of satellite data for multisensor QPE is an increasingly important topic (Ashouri et al. 2014; Gourley et al. 2011; Kalinga and Gan 2010; Prat and Nelson 2013). Numerous efforts have been made to reduce systematic and random errors in multisensor QPE via bias correction and multivariate analysis (Kondragunta et al. 2005; Nelson et al. 2010; Prat et al. 2014, 2015; Seo et al. 2010; Smith and Krajewski 1992; Smith et al. 1996; Vasiloff et al. 2007). Here, by bias correction, we mean multiplicative or additive correction applied to the raw gridded precipitation data at an effective spatiotemporal scale larger than a single grid box or a time step. By multivariate analysis, we mean data assimilation with multivariate observation equation with or without a dynamical model. Many bias correction and multivariate analysis techniques involve minimizing mean square error (MSE) or error variance under unbiasedness (Ciach et al. 2000; Goudenhoofdt and Delobbe 2009; Gandin 1963; Seo and Breidenbach 2002; Seo 1998a,b; Smith et al. 2006). It is well known in statistics and econometrics that, in the presence of observation error, variance minimization tends to introduce negative and positive biases, or conditional bias (CB), over the upper and lower tails of the predictands, respectively (Fuller, 1987; Seber and Wild, 1989; Hausman, 2001). In regression, the above effect is referred to as regression dilution, which results in attenuation bias in the regression coefficients (Hughes 1993; Frost and Thompson 2000). Such bias may not pose an issue when prediction is made for the same predictands via the same regression model using new observations from the same observing systems. For example, if one desires the predictands to be subject to the same uncertainty characteristics as the observed outcomes of the predictands used in the regression analysis itself, the attenuation bias would not be a concern. In QPE, however, the objective is to estimate true precipitation amounts as accurately as possible so that they may be used as observed initial or boundary conditions in a wide range of applications. Because precipitation observations almost always have significant uncertainties, particularly at high spatiotemporal resolutions, the presence of CB in QPE is the norm rather than the exception. Therefore, addressing CB is an essential topic in multisensor QPE.

One may, in general, differentiate the CB into Types I and II. The Type-I CB, defined as $E[X|\hat{X}=\hat{x}]-\hat{x}$, where X, $\hat{X}$ and $\hat{x}$ denote the unknown truth, the estimate, and the realization of $\hat{X}$, respectively (Joliffe and Stephenson 2003), is associated with false alarm. The Type-II CB, defined as $E[\hat{X}|X=x]-x$, where x denotes the realization of X, is associated with failure to detect an event. Whereas the Type-I CB may be reduced by calibration, the Type-II CB cannot (Wilks 2006; Seo et al. 2018a,b). Ciach et al. (2000) found that minimizing the MSE in radar rainfall estimates increases the CB and that, when estimating extremes is of interest, there is a trade-off to consider between minimizing the MSE and reducing the CB. To address the detrimental effects of the Type-II CB on estimating extremes, Seo (2013) introduced a new optimal linear estimation method that minimizes the weighted sum of error variance and expectation of the CB squared. When cast in the form of kriging, the method yields CB-penalized kriging (CBPK), which has been shown to broadly outperform conventional kriging in the prediction of high flows and estimation of heavy to extreme rainfall (Brown and Seo 2013; Seo 2013; Seo et al. 2014; Kim et al. 2018). When cast in the form of the Kalman filter (KF), the method yields the CB-penalized KF (Seo et al. 2018a,b; Shen et al. 2019) and, in ensemble form, the CB-penalized ensemble KF (Lee et al. 2019). Though the CB-penalized estimation techniques significantly improve performance over the tails, they do not minimize the MSE in the unconditional sense. This deterioration in unconditional performance may be reduced by prescribing the weight to the CB penalty adaptively based, e.g., on the best available estimate of the unknown true state (Kim et al. 2018; Shen et al. 2019); if the best estimate is near the median or in the tails of the distribution, one may reduce the weight close to zero or to a large value, respectively. When there exist large observational uncertainties, or the precipitation field has limited predictability, however, the above approach may not be able to identify the state of the system with sufficient accuracy and consistency to be effective.

In this chapter, we introduce adaptive CB-penalized optimal estimation for merging rain gauge data and radar-only or radar-satellite-fused QPE, which explicitly optimizes the weight for the CB penalty in real-time. The specific optimal linear estimation technique considered is the CB-penalized cokriging (CBPCK, Kim et al. 2018). The resulting technique is referred to as adaptive CBPCK. The outcome sought is improved estimation over the tails of the distribution of precipitation while performing comparably to OCK in the unconditional sense. We then comparatively evaluate adaptive CBPCK with OCK used in the National Weather Service's (NWS) Multisensor Precipitation Estimator (MPE, Habib et al. 2013; Kitzmiller et al. 2013; Nelson et al. 2016; Seo et al. 2010). As part of the evaluation, we also assess the incremental value, in reference to the radar-only QPE, of rain gauge-based mean field bias (MFB) correction of radar-only QPE, merging of rain gauge data and MFB-corrected radar QPE, the fusion of MFB-corrected

SCaMPR QPE with MFB-corrected radar QPE, and merging of rain gauge data and radar-satellite-fused QPE. The main contributions of this work are the development and comparative evaluation of adaptive CBPCK, assessment of the incremental value of MFB correction, fusion, and merging, and advances in the understanding of the CB in multisensor QPE and its correction. The rest of this chapter is organized as follows. Section 2.2 describes the data used. Section 2.3 describes the methods used. Section 2.4 presents the results. Section 2.5 provides the conclusions and future research recommendations.

2.2 Data used

Two analysis periods are used: Oct 7-9, 2016, and Sep 13-30, 2015. Figs 2.1 and 2.2 show the radar-only precipitation totals. The first period includes Hurricane Matthew on the East Coast, weakly organized convective storms in the central US, and a coastal storm in the Pacific Northwest. The second period generally includes convective events in the Pacific Northwest, Midwest, Northeast, and Southeast of the US. The rain gauge data used are the hourly observations collected through the Hydrometeorological Automated Data System (HADS, Kim et al. 2009) operated by the NWS. Fig 2.3 shows the gauge locations, which total over 21,000. The HADS is a real-time data acquisition, processing, and distribution system supporting the NWS's Flood and Flash Flood Warning programs. The system acquires raw hydrologic and meteorological observations throughout the US from the Geostationary Operational Environmental Satellites (GOES) Data Collection Platforms, mostly owned and/or operated by various federal, state, and local agencies. The hourly radar QPEs used are from the Multi-Radar Multi-Sensor System (MRMS; Zhang et al. 2011, 2016) at 1 km resolution. MRMS is a system of automated algorithms that integrate data from multiple radars, surface and upper-air observations, lightning detection systems, and satellite and numerical weather forecast models.

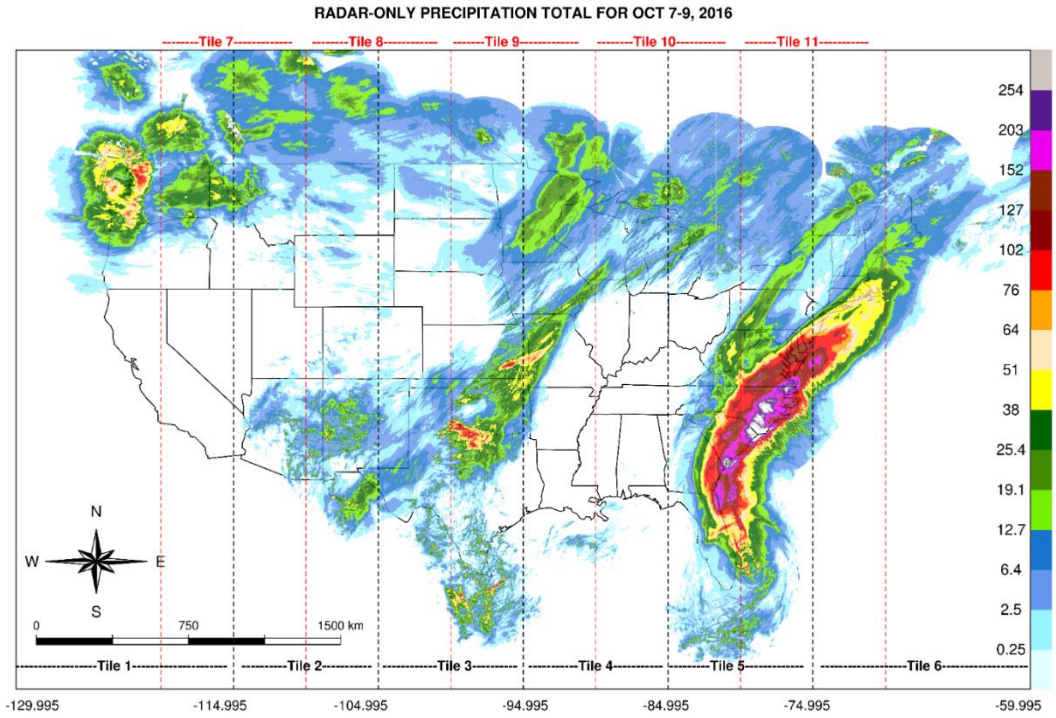


Figure 2.1: Radar-only precipitation map for Oct 7-9, 2016. The 11 geographic tiles shown are used in the MRMS for parallel processing. The first six tiles cover the full CONUS, and the last 5 straddle the first 6 to reduce the edge effects. The map is obtained by averaging all overlapping estimates over the CONUS

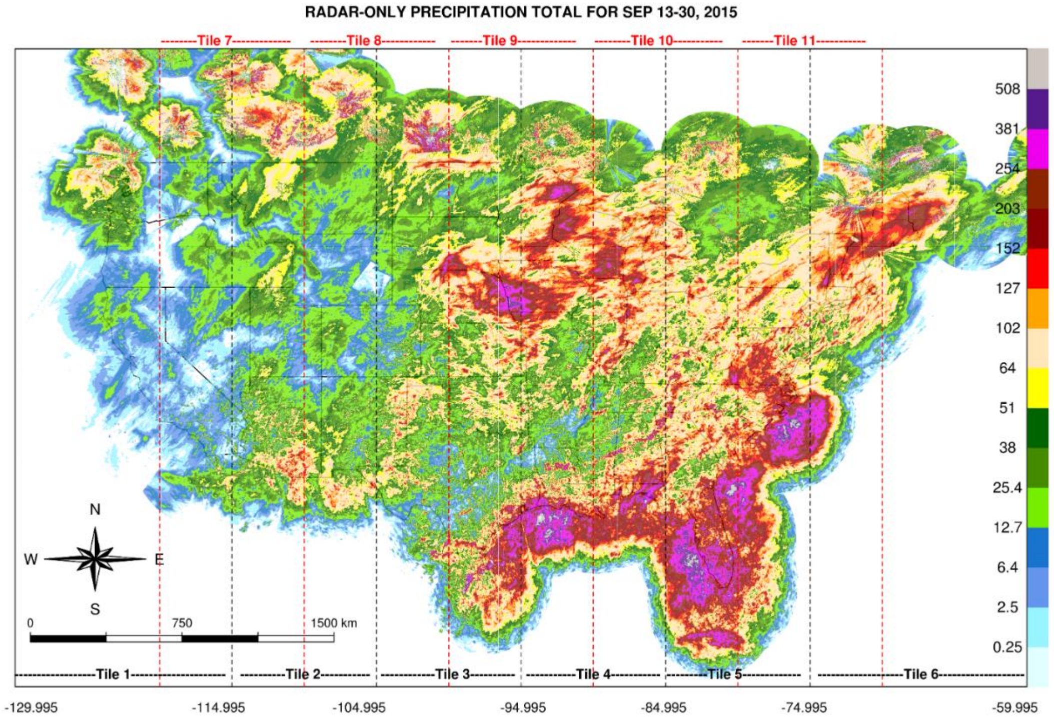


Figure 2.2: Same as Fig 2.1 but for Sep 13-30, 2015.

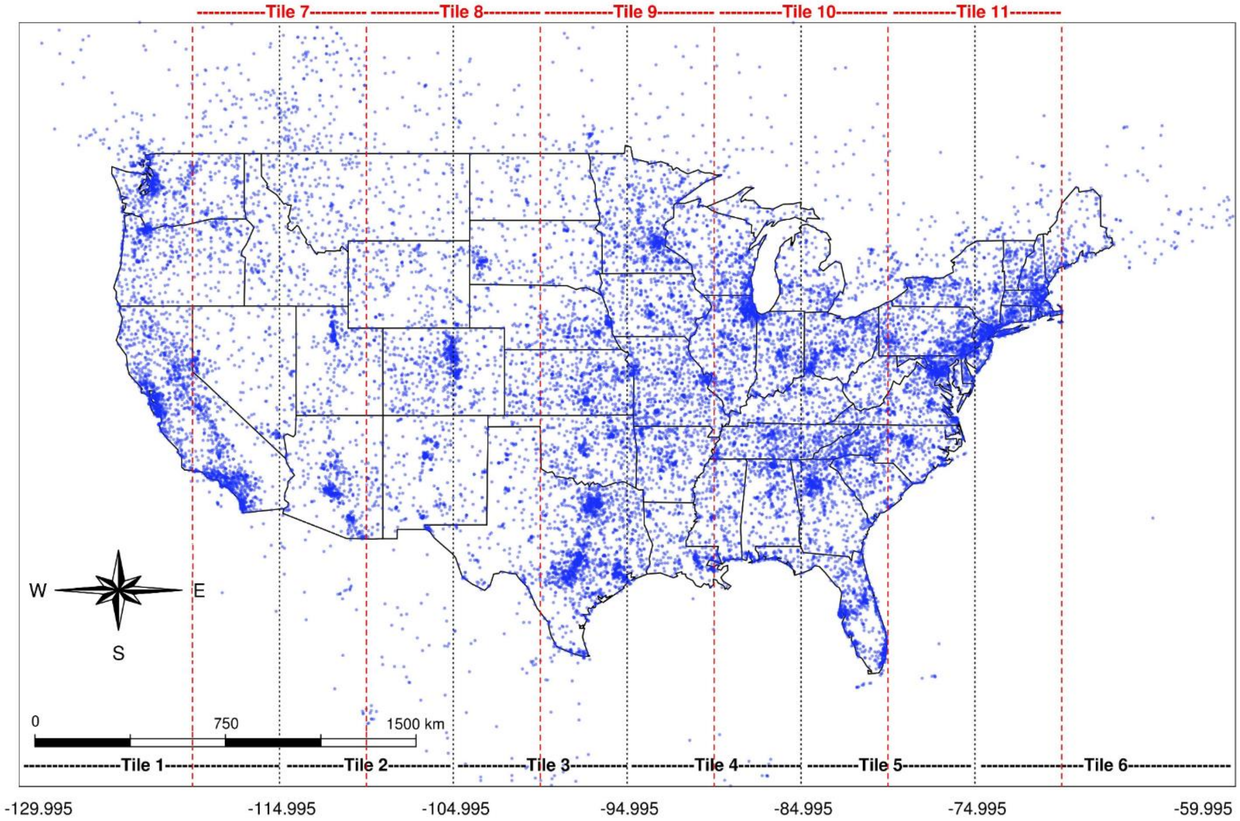


Figure 2.3: Rain gauge location map.

The system generates a suite of 2D multisensor products for monitoring and short-term prediction of hail, wind, tornado, QPE, convection, icing, and turbulence. The hourly satellite QPE used in the study was from the Self-Calibrating Multivariate Precipitation Retrieval (SCaMPR; Kuligowski 2002, 2010; Kuligowski et al. 2013) at 4 km resolution. SCaMPR uses GOES infrared data for predictor information and calibrates them against microwave-based rain rates. The algorithm performs rain/no rain discrimination using discriminant analysis, precipitation rate calibration using regression, nonlinear transformations of the predictors to optimize the regression fits, and accounts for evaporation of hydrometeors below the cloud base in dry environments. In this work, all multisensor QPE operations are carried out for each of the 11 geographic tiles used in the MRMS. The first six tiles cover the full CONUS, and the last 5 straddle the first 6 to reduce

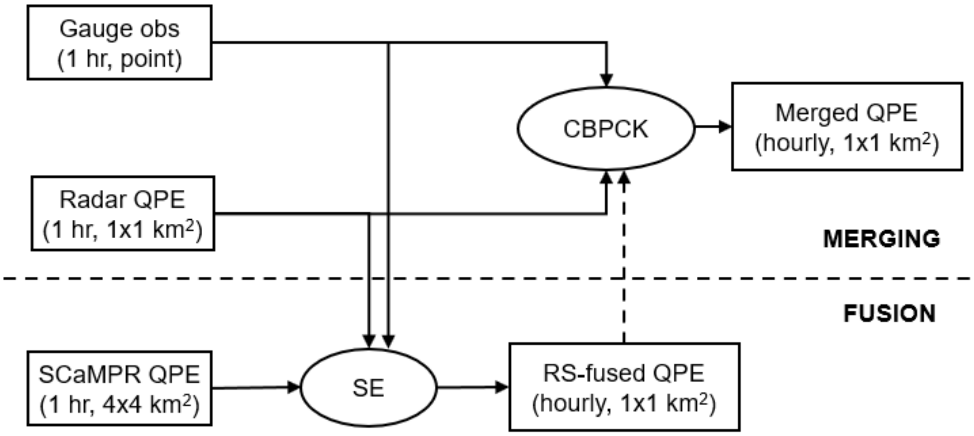


Figure 2.4: Schematic of the merging and fusion processes used.

the edge effects (see Fig 2.1 or 2.2). The final analysis is obtained by averaging all overlapping estimates over the CONUS.

2.3 Methods used

This section describes the multisensor QPE methods used in this work. Fig 2.4 shows the schematic of the estimation process.

2.3.1 Bias correction of gridded QPE

The MRMS radar-only and SCaMPR QPEs (Sep 2015 only for the latter) are first MFB-corrected for each tile using the respective gridded QPEs and all available gauge data within the tile. The procedure used is the MFB correction algorithm of Seo et al. (1999), which has been in operation in the MPE since the mid-2000s. In the MPE, the algorithm operates for each radar and updates the radar umbrella-wide biases at multiple temporal scales of aggregation ranging typically from hourly to multi-annual. The bias estimated is the multiplicative correction factor, $\beta_k = \int_{A_c} G_k(u)du \,/ \int_{A_c} R_k(u)du$, to be applied to the radar-only QPE spatially-uniformly where β_k denotes the MFB in the gridded QPE at the k-th hour, A_c denotes the precipitation area, and $G_k(u)$

and $R_k(u)$ denote the gauge and radar precipitation at location u, respectively. To estimate β_k, the procedure uses all available (including posting-delayed) collocated and synchronous pairs of positive gauge and positive radar precipitation observations, and updates β_k in real-time at multiple time scales via exponential smoothing (Schweppe 1973). Conceptually, the smoothing operation amounts to recursively estimating age-weighted moving averages of gauge and radar precipitation observations over time windows of different lengths. The updated bias associated with the smallest time scale is then chosen as the final estimate among those with an effective sample size greater than the user-set minimum. In this way, a shorter- and longer-term bias is used in gauge-rich and –poor areas, respectively. For further details, the reader is referred to Seo et al. (1999). Unlike in the MPE, here we estimate β_k for an entire tile which is much larger than the typical effective coverage of a single S-band radar such as the Weather Surveillance Radar – 1988 Doppler version. The above choice is made necessary by the fact that the MRMS radar-only QPE is already a mosaic of data from multiple radars (Zhang et al. 2016). MFB correction for the SCaMPR QPE is completely analogous. Given the large latitudinal dimension of the tiles, multiple storms with disparate biases may exist in a single tile. In such cases, the effectiveness of MFB correction, as implemented in this work, is likely to be reduced. Performance assessment under different tile sizes, however, was beyond the scope of this work and is left as a future endeavor.

2.3.2 Merging of rain gauge data and gridded QPE

We use OCK and CBPCK to estimate the true precipitation amount at an arbitrary location, u_0, using rain gauge observations at $u_i, i = 1, \dots, n_g$, and remotely-sensed QPE at $u_i, i = n_g + 1, \dots, n_g + n_r$, where n_g and n_r denote the number of rain gauges and remotely-sensed QPE, respectively. Though referred to as cokriging, the formulation described here applies to kriging as well, except that in the latter, all observations are from a single observing system. Dropping the time index for brevity, we write the linear estimator for gauge precipitation at an ungauged location, u_0, as:

$$Z_0^* = \sum_{i=1}^{n} \lambda_i \frac{m_0}{m_i} Z_i \tag{2.1}$$

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{2.2}$$

In the above, Z_0^* denotes the precipitation estimate at u_0, n denotes the total number of data used in the estimation, $n = n_g + n_r$; Z_i , $\mathrm{i} = 1, \dots, n_g$, denote the nearest gauge observations; Z_i , $\mathrm{i} = n_g + 1, \dots, n_g + n_r$, denote the nearest remotely-sensed precipitation estimates; $\lambda_i, i = 1, \dots, n$, denote the weights for Z_i, i=1,…,n; and m_0 and m_i denote the climatological mean precipitation at locations u_0 and u_i, respectively, for which we use the monthly PRISM climatology (Daly et al. 1994). The constraint, Eq. (2.2), renders the estimate Z_0^* in Eq. (2.1) unbiased in the mean under the assumption that the gridded QPE is unbiased relative to the gauge observations. Implicit in Eq. (2.1) is the assumption that the MFB-corrected radar QPE is climatologically unbiased (Seo et al. 2010).

OCK and CBPCK differ mainly in the calculation of the weights, $\lambda_i, i = 1, \dots, n$ in Eq. (2.1). Whereas OCK minimizes the error variance of Z_0^*, $J_{EV} = E[\{Z_0^* - Z_0\}^2]$, CBPCK minimizes the weighted sum of J_{EV} and the expectation of the CB squared, $J_{CB} = E[\{\mathrm{E}[Z_0^* | Z_0] - Z_0\}^2]$, i.e., $J = J_{EV} + \alpha J_{cB}$ where α denotes the weight given to the CB penalty. If $\alpha = 0$, CBPCK reduces to OCK. Accordingly, one may consider CBPCK as generalized OCK. The CBPCK system is given by (Kim et al. 2018):

$$\sum_{j=1}^{\mathrm{n}} \lambda_j \left(\rho_{ij} + \alpha \cdot \rho_{i0} \cdot \rho_{j0}\right)\sigma_i \cdot \sigma_j = (1 + \alpha)\rho_{i0} \cdot \sigma_i \cdot \sigma_0 \quad i = 1, \dots, \mathrm{n} \tag{2.3}$$

$$\sum_{i=1}^{\mathrm{n}} \lambda_i = 1 \tag{2.4}$$

where ρ_{ij} denotes the (cross-) correlation between the two variables at u_i and u_j; and σ_i and σ_j denote the standard deviation of the two variables at ui and uj, respectively. The objective function, J, is associated with the CBPCK solution from Eqs. (2.3) and (2.4) is given by (Kim et al. 2018):

$$J = (1+\alpha)(\sigma_0^2 - \Sigma_{i=1}^{n} \lambda_i \rho_{i0} \sigma_i \sigma_0) - \mu \tag{2.5}$$

where μ denotes the Lagrange multiplier. Because J reflects not only the error variance but also the CB penalty, it is larger than J_{EV} for $\alpha > 0$. To evaluate J_{EV}, we decompose J into J_{EV} and αJ_{cB} to obtain:

$$J_{EV} = \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} \lambda_i \lambda_j \rho_{ij} \sigma_i \sigma_j - 2\Sigma_{i=1}^{n} \lambda_i \rho_{i0} \sigma_i \sigma_0 + \sigma_0^2 \tag{2.6}$$

Because CBPCK does not minimize (unconditional) error variance, Eq. (2.6) is necessarily larger than the OCK error variance for $\alpha > 0$. The optimal weights, λ_i, i=1,…,n, are functions of the covariance among the gauge observations, covariance among the remotely-sensed QPE, and cross-covariance between the gauge observations and remotely-sensed QPE. For example, the cross-covariance between the gauge observation at u_i, Z_i , $\mathrm{i} = 1, \ldots, n_g$, and the remotely-sensed QPE at u_j, Z_j , $\mathrm{j} = n_g + 1, \ldots, n_g + n_r$, may be written as (seo 1998 a,b):

$$\begin{aligned} Cov[Z_i, Z_j] = {} & \sigma_g \sigma_r \left[m_{Ig}\left(1 - m_{Ig}\right)\right]^{1/2} [m_{Ir}(1 - m_{Ir})]^{1/2} \rho_c\left(\left|u_i - u_j\right|\right) \rho_{Ic}\left(\left|u_i - u_j\right|\right) \\ & + m_g m_r [m_{Ig}\left(1 - m_{Ig}\right)]^{1/2} [m_{Ir}(1 - m_{Ir})]^{1/2} \rho_{Ic}\left(\left|u_i - u_j\right|\right) + \\ & \sigma_g \sigma_r m_{Ig} m_{Ir} \rho_c\left(\left|u_i - u_j\right|\right) \end{aligned} \tag{2.7}$$

where σ_g and σ_r denote the standard deviation of gauge and remotely-sensed precipitation, respectively; m_{Ig} and m_{Ir} denote the mean fractional coverage by gauge and remotely-sensed precipitation; m_g and m_r denote the unconditional mean of gauge and remotely-sensed precipitation; and $\rho_c(|u_i - u_j|)$ and $\rho_{Ic}(|u_i - u_j|)$ denote the conditional and indicator cross-correlation at a separation distance of $|u_i - u_j|$ which model the inner variability (i.e., variability of positive precipitation) and intermittency of precipitation (i.e., variability of precipitation vs. no precipitation), respectively. Subsection 2.3.4 describes how the statistical parameters in Eq. (2.7) estimated in this work. Due to the generally large skewness in hourly precipitation and larger errors near median than OCK, CBPCK produces negative estimates more frequently than OCK in areas of very light precipitation (see Seo 2013; Seo et al. 2014). To address this, we apply the correction procedure of Kim et al. (2018) in which the negative and positive CBPCK estimates are set to zero and adjusted by a scaling factor, respectively. For details, the reader is referred to Kim et al. (2018) and Seo et al. (2014).

The CBPCK solution depends on α. If there is little or no CB, we have $\alpha \approx 0$, and OCK suffices. If there is a large CB due to small predictability in the precipitation field, sparsity in the rain gauge network or large uncertainty in the remotely sensed precipitation data (Seo 2013), one may expect CBPCK with a large α to produce more accurate estimates and error variances than OCK. Because there are multiple sources of CB, prescribing α a priori is a difficult proposition. In adaptive CBPCK, the weight α is optimized in real-time in a data-driven manner such that the MSE, or any other performance measure or measures of choice, are optimized. To that end, we add the following adaptive steps to CBPCK: 1) discretize the possible range of α, 2) given a value of α, use CBPCK to estimate precipitation at all gauge locations within the tile in a cross-validation mode, 3) repeat Step 2 for all values of α, 4) identify the MSE-minimizing α, and 5) perform CBPCK analysis over the entire tile using the “optimal” α. Because the sample size from cross-validation may vary greatly from hour to hour within the same tile and from tile to tile for the same hour, the resulting α is subject to potentially large sampling uncertainties. The time series plots of α identified in this way often show unrealistically large fluctuations compared to the characteristic time scales of CB that may be expected from the predictability conditions. For the above reason, we employ exponential smoothing (Schweppe 1973) of the error statistics at multiple time scales in a manner analogous to that used in MFB correction. Though effective in reducing sampling uncertainties, such smoothing may potentially over-smooth the error statistics in fast-changing conditions, in which case the quality of the resulting α would be lower.

Because changing α has similar effects to changing the covariance structure (see Eq. (2.3)), it is possible that optimizing α may not only correct for the CB but also compensate for possibly incorrect covariance structure. Differentiation of the two, however, is not readily possible because the true covariance structure is not known in the real world. Synthetic experiments to address the above are beyond the scope of this work and are left as a future endeavor. In practice, the lack of differentiation may not pose a significant issue because the uncertainty in the covariance structure is likely to be a contributing factor to the CB as well. In this work, we exponentially smooth the second-order statistics used in CBPCK in a manner analogous to that used in MFB correction. In this way, all statistical parameters used are subject to similar levels of sampling uncertainty. Subsection 2.3.4 provides details on how the statistical parameters are estimated.

2.3.3 Radar-satellite fusion

The MRMS radar-only and the SCaMPR QPEs are first MFB-corrected individually, as described in Subsection 3.1. The resulting gridded QPEs are fused via Simple Estimation (SE; Rafieeinasab et al., 2015). In this procedure, the 1-km hourly MRMS radar-only estimates are aggregated to the spatial scale of the 4-km hourly SCaMPR estimates. The two estimates are then fused at each SCaMPR grid box according to Eq. (2.8), where the time index has been dropped for brevity:

$$X_U^* = wR_U + (1-w)S_U \tag{2.8}$$

In the above, R_U denotes the upscaled MFB-corrected radar QPE; S_U denotes the MFB-corrected SCaMPR QPE; w denotes the optimal weight to be determined; and X_U^* denotes the hourly fused estimate at a 4-km scale. The weight, w, is obtained via adaptive CB-penalized optimal linear estimation, which minimizes the linearly weighted sum of the error variance and the Type-II CB squared:

$$J = E[(X_U^* - X_U)^2] + \alpha_f E[(E[X_U^*|X_U] - X_U)^2] \tag{2.9}$$

where α_f denotes the weight given to the CB penalty term for fusion. Seo (2013) arrives at the following Fisher-like solution for the CB-penalized optimal linear estimate, X_U^*, and the error variance, Σ:

$$\Sigma = B\left[\widehat{U}^T \Lambda^{-1} U\right]^{-1} \tag{2.10}$$

$$X_U^* = \left[\widehat{U}^T \Lambda^{-1} U\right]^{-1} \widehat{U}^T \Lambda^{-1} Z = [w \;\; (1-w)] \begin{bmatrix} R_U \\ S_U \end{bmatrix} \tag{2.11}$$

In the above, the (2×1) modified unit vector $\widehat{U}$, the modified observation error covariance matrix Λ, and the scaling matrix, B, are given by:

$$\widehat{U}^T = (1+\alpha_f)U^T \tag{2.12}$$

$$\Lambda = \mathcal{R} - \alpha_f\left(\alpha_f + 1\right)U\sigma_{X_U}^2 U^T \tag{2.13}$$

$$B = \alpha_f \sigma_{X_U}^2 \widehat{U}^T \Lambda^{-1} \widehat{U} + (1+\alpha_f)I \tag{2.14}$$

In the above, $\sigma_{X_U}^2$ denotes the variance of X_U, and the observation error covariance matrix, $\mathcal{R}$, is given by:

$$\mathcal{R} = \begin{bmatrix} Var[R_U - X_U] & 0 \\ 0 & Var[S_U - X_U] \end{bmatrix} \tag{2.15}$$

If $\alpha_f = 0$, Eqs. (2.12) and (2.13) reduce to the classical Fisher (i.e., maximum likelihood) solution (Schweppe 1973). The diagonality of $\mathcal{R}$ reflects the very reasonable assumption that the observation errors in R_U and S_U are independent. To estimate $Var[R_U - X_U]$ and $Var[S_U - X_U]$ in Eq. (2.15), we update the error statistics of the radar and SCaMPR QPEs vs. gauge precipitation in real-time via exponential smoothing. The weight, α_f, is optimized in real-time analogously to adaptive CBPCK. To obtain the hourly fused estimate at a 1-km scale, X_U^* is disaggregated under the assumption that the 1-km hourly radar QPE perfectly captures the spatial variability of precipitation within each grid box of the SCaMPR QPE:

$$X_{i,j}^* = \frac{X_U^*}{R_U} R_{i,j}, R_U > 0;\ i = 1, \dots, 4;\ j = 1, \dots 4 \tag{2.16}$$

In the above, $R_{i,j}$ and $X_{i,j}^*$ denote the radar and fused estimates at the ij-th pixel, respectively. By replacing X_U^* in Eq. (2.16) with Eq. (2.8), we may rewrite $X_{i,j}^*$ as:

$$X_{i,j}^* = \left(w + (1-w)\frac{S_U}{R_U}\right) R_{ij} = wR_{i,j} + (1-w)S_{i,j}^*, R_U > 0; i = 1, \dots, 4;\ j = 1, \dots 4 \tag{2.17}$$

where

$$S_{i,j}^* = \frac{R_{i,j}}{R_U} S_U, R_U > 0; i = 1, \dots, 4;\ j = 1, \dots 4 \tag{2.18}$$

Eq. (2.17) states that one may interpret SE as a multiplicative correction of radar-only QPE where the correction factor is given by $w + (1-w)\frac{S_U}{R_U}$. Alternatively, SE may be interpreted as a combination of naïve downscaling via Eq. (2.18) of the SCaMPR QPE and constrained linear regression via Eq. (2.17) of the radar QPE and the downscaled SCaMPR QPE. Unlike in merging, the weight w does not vary in space owing to the gridded nature of the input QPEs, which renders the fusion algorithm extremely simple. It is important to note that, whereas rain gauge data are used to estimate the observation error statistics, they are not used in the fusion itself unlike in OCK or CBPCK. As such, the timeliness of the rain gauge data is not as critical to fusion as it is to merging.

2.3.4 Estimation of statistical parameters

As implemented in this work, both OCK and CBPCK require modeling spatial covariance structures of intermittency and inner variability of precipitation. Under the assumption of local homogeneity, we equate the mean fractional coverage of precipitation with the probability of precipitation. As in Seo (1998b), we assume $\sigma_g = \sigma_r$; $m_{Ig} = m_{Ir}$; $m_g = m_r$; $\rho_g(|u_i - u_j|) = \rho_c(|u_i - u_j|) = \rho_r(|u_i - u_j|)$; and $\rho_{Ig}(|u_i - u_j|) = \rho_{Ic}(|u_i - u_j|) = \rho_{Ir}(|u_i - u_j|)$. The correlograms are estimated using the hourly radar-only QPE (Seo 1998b). It is impractical to model fully spatiotemporally-varying covariance structures in real-time due to insufficient data, modeling complexity, and large computing requirements. In the current implementation of OCK in the MPE, the correlation structures are not estimated in real-time but are based on climatological estimates (Seo 1998b, Seo and Breidenbach 2000). In this work, we assume the exponential model (Journel and Huijbregts 1978) for both intermittency and inner variability and estimate the parameters for the correlation model for each hour for each tile. To reduce computing time, we estimate the correlation coefficients only for the first few lags, from which the nugget effect and the range are estimated (Journel and Huijbregts 1978). Though the exponential model very often provides the best fit for hourly radar and rain gauge precipitation among widely used correlogram models (Seo and Breidenbach 2002; Seo et al. 2014), it may not be reasonable for stochastically highly-regular (i.e., mean-square differentiable) precipitation fields (Vanmarcke 1983) or orographic precipitation (Chua and Bras 1982). In addition, the correlation structure may not be homogeneous within a tile given the large latitudinal dimension (see Fig 2.1). As such, the covariance models are subject to potentially significant uncertainties, which may also contribute to the CB. Additional research is needed to improve real-time modeling of the spatiotemporally-varying covariance structure of precipitation over large areas.

In the western US, significant radar coverage gaps and areas of significant beam blockage or occultation exist (Breidenbach et al. 1999, 2001; Maddox et al. 2002). In these areas, the lag-0 cross-correlation coefficients, $\rho_{Ic}(|0|)$ and $\rho_c(|0|)$, are generally significantly lower, particularly when beam overshooting occurs or radar observes frozen hydrometeors (Seo et al. 2000, 2010). When and where the correlation coefficients are very small, OCK and CBPCK are effectively reduced to OK and CBPK using rain gauge data alone, respectively. In the other extreme, when and where the correlation coefficients are very high, the OCK and CBPCK analyses largely mimic the MFB-corrected radar QPE except in the immediate vicinity of the rain gauge locations. Because we are using ordinary, rather than simple, cokriging (Journel and Huijbregts 1978), specifying the radius of influence requires additional care. In this work, we set the default radius of influence for locating the neighboring gauge observations to be 2.5 times larger than the larger of the indicator and conditional correlation scales. In this way, weakly correlated rain gauge observations may also be included in the estimation process. It is well known that the predictability of precipitation depends strongly on the magnitude of precipitation. In general, the predictability peaks around the median and decreases toward the tail ends of the distribution of precipitation amount (Seo 1996). Because high-resolution positive precipitation is highly skewed, the predictability tends to decrease rather quickly as the precipitation amount increases. The above dependence may be modeled explicitly using nonlinear estimation techniques such as indicator cokriging (Brown and Seo 2013). They are, however, computationally very expensive and require large amounts of data for parameter estimation. In this work, we partially affect the above dependence by parameterizing the radius of influence with the radar precipitation amount at u_0 as follows:

$$ROI = ROI_{def} \exp\left(-\frac{Z_R(u_0)}{L}\right) \tag{2.19}$$

where ROI denotes the radius of influence, ROI_{def} denotes the default radius of influence, $Z_R(u_0)$ denotes the radar precipitation at u_0, and L denotes the characteristic precipitation amount (mm). If radar precipitation does not exist at u_0, the default radius of influence is used. The maximum number of neighboring gauge observations used in the estimation process is 30 to limit the amount of computation. The actual number of rain gauge observations used (i.e., those within ROI), however, is generally much smaller, particularly in gauge-sparse areas. The fractional coverage of precipitation is estimated by dividing the number of positive observations by the total number of observations within ROI_{def}.

2.3.5 Evaluation

For comparative evaluation of the different QPEs considered, we carried out true validation. For Sep 13-30, 2015, we randomly selected 4 to 5% of all available gauge observations within each tile for each hour, withheld them for validation, and used the rest for parameter estimation. The gauge network density varies greatly from one tile to another and within each tile (see Fig 2.3). In gauge-poor areas, setting aside more gauges would keep the neighboring gauges too distant to provide a predictive skill that may actually exist. The choice of 4 to 5% represents a compromise between retaining as much of the actual gauge network as possible and increasing the sample size for validation. For large-sample validation, the above experiment was repeated five times; each time withholding previously unwithheld gauges. The total number of data points obtained in this way for validation is over 20,000 for the 11 tiles. For Oct 7-9, 2016, 3 to 4% of the gauges were withheld. The experiment was then repeated ten times using previously unwithheld gauges. The total number of data points obtained for validation is over 19,000 for this analysis period. Comparative evaluation for the Oct 2016 case is focused on addressing the following questions: How much does the MFB-corrected radar QPE improve over the radar-only?; How much do the OCK estimates from merging gauge data and MFB-corrected radar QPE improve over the latter alone?; and How much does adaptive CBPCK improve over OCK? Comparative evaluation for the Sep 2015 case is focused on addressing: How much do the SE estimates from fusing the MFB-corrected SCaMPR and radar QPEs improve over the latter alone?; How much do the OCK estimates from merging gauge data and MFB-corrected gridded QPE improve over the latter alone?; and How much does adaptive CBPCK improve over OCK? For both cases, we also address how the above comparative performance may vary according to the magnitude of precipitation being estimated.

For evaluation metrics, we use the root mean square error (RMSE), its decomposition, and percent reduction in RMSE relative to reference QPE. The RMSE collectively measures biases in the mean and standard deviation and strength of correlation. Both OCK and CBPCK are unbiased estimators, and their performance with respect to the three attributes above has also been reported in Seo et al. (2010), Seo et al. (2014), and Kim et al. (2018). The RMSE is given by:

$$RMSE = \sqrt{\sum_{k=1}^{K} \sum_{i=1}^{n_{g,k}} [Z_{k,i}^{*} - G_{k,i}]^{2} \,/\, \sum_{k=1}^{K} n_{g,k}} \tag{2.20}$$

where K denotes the total number of hours in the analysis period; $n_{g,k}$ denotes the total number of gauges used in the true validation at hour k; $Z_{k,i}^{*}$ denotes the estimated precipitation at location

u_i at hour k; and $G_{k,i}$ denotes the verifying gauge observation at location u_i. The percent reduction in RMSE, or $PRiRMSE(QPE_{eval})$, is defined as:

$$PRiRMSE(QPE_{eval}) = \frac{RMSE(QPE_{ref}) - RMSE(QPE_{eval})}{RMSE(QPE_{ref})} \times 100 \tag{2.21}$$

where $RMSE(QPE_{eval})$ and $RMSE(QPE_{ref})$ denote the RMSEs of the QPE under evaluation and the reference QPE, respectively. MSE decomposition (Murphy and Winkler 1987; Nelson et al. 2010) is given by:

$$MSE = (m_e - m_o)^2 + (\sigma_e - \sigma_o)^2 + 2\sigma_e\sigma_o(1 - \rho_{e,o}) \tag{2.22}$$

where m_e and m_o denote the mean of the estimate and verifying observation, respectively, σ_e and σ_o denote the standard deviation of the estimate and verifying observation, respectively, and $\rho_{e,o}$ denotes the correlation between the estimate and verifying observation. MFB correction and merging address the first- (i.e., systematic) and second-order (i.e., random) errors, respectively. One may consider the CB a 1.5th-order error in that it is systematic but exists only over the tails of the distribution. First-order errors impact the accuracy across the board. As such, one may expect MFB correction to have the largest impact if significant first-order errors exist. The focus of this work is on addressing the CB. We are hence interested in assessing the accuracy of the merged QPE relative to that of the MFB-corrected radar, in addition to that of the radar-only QPE. For this reason, we present the results of both MFB correction and merging relative to radar-only QPE so that the improvement due solely to each may be ascertained.

2.4 Results

This section presents the merging results for Oct 7-9, 2016, and Sep 13-30, 2015, and the fusion and merging results for Sep 13-30, 2015.

2.4.1 Radar-gauge merging

The performance of the procedures described above depends on the predictability of the precipitation field, the skill in the remotely sensed QPE, and the gauge network density. To assess the first two factors between the two periods, we first examine the indicator and conditional spatial correlation scales of radar QPE, and the (spatial) lag-0 indicator and conditional cross correlation between gauge observation and radar QPE as estimated in the simulated real-time mode. Fig 2.5 shows the histograms of the above four parameters for the entire analysis periods over the CONUS.

Because these estimates reflect different types of precipitation events in different phases of development and dissipation under a wide range of conditions for radar observation of precipitation, the sample statistics necessarily vary greatly. Nonetheless, it is readily seen that all four parameters for Sep 2015, particularly the conditional correlation and conditional correlation scale, are significantly smaller than those for Oct 2016, an indication that the precipitation fields in the first period are characterized by significantly smaller predictability. One may hence expect merging to be more potent for Oct 2016, and the CB to be larger for Sep 2015.

Oct 7-9, 2016

This period includes precipitation from storms in largely three different regions: extreme amounts from Hurricane Matthew along the Atlantic Coast; significant amounts from a relatively weakly organized convective storm in the central US; and significant amounts from a coastal storm in the Pacific Northwest (see Fig 2.1). Fig 2.6a shows the RMSE of the radar-only, MFB-corrected radar, OCK, and adaptive CBPCK estimates over the CONUS conditional on the verifying observed hourly precipitation exceeding the amount shown on the x-axis. The values on the y-axis at x=0 represent the RMSE conditional on the verifying observation being nonzero, which is very close to the unconditional RMSE.

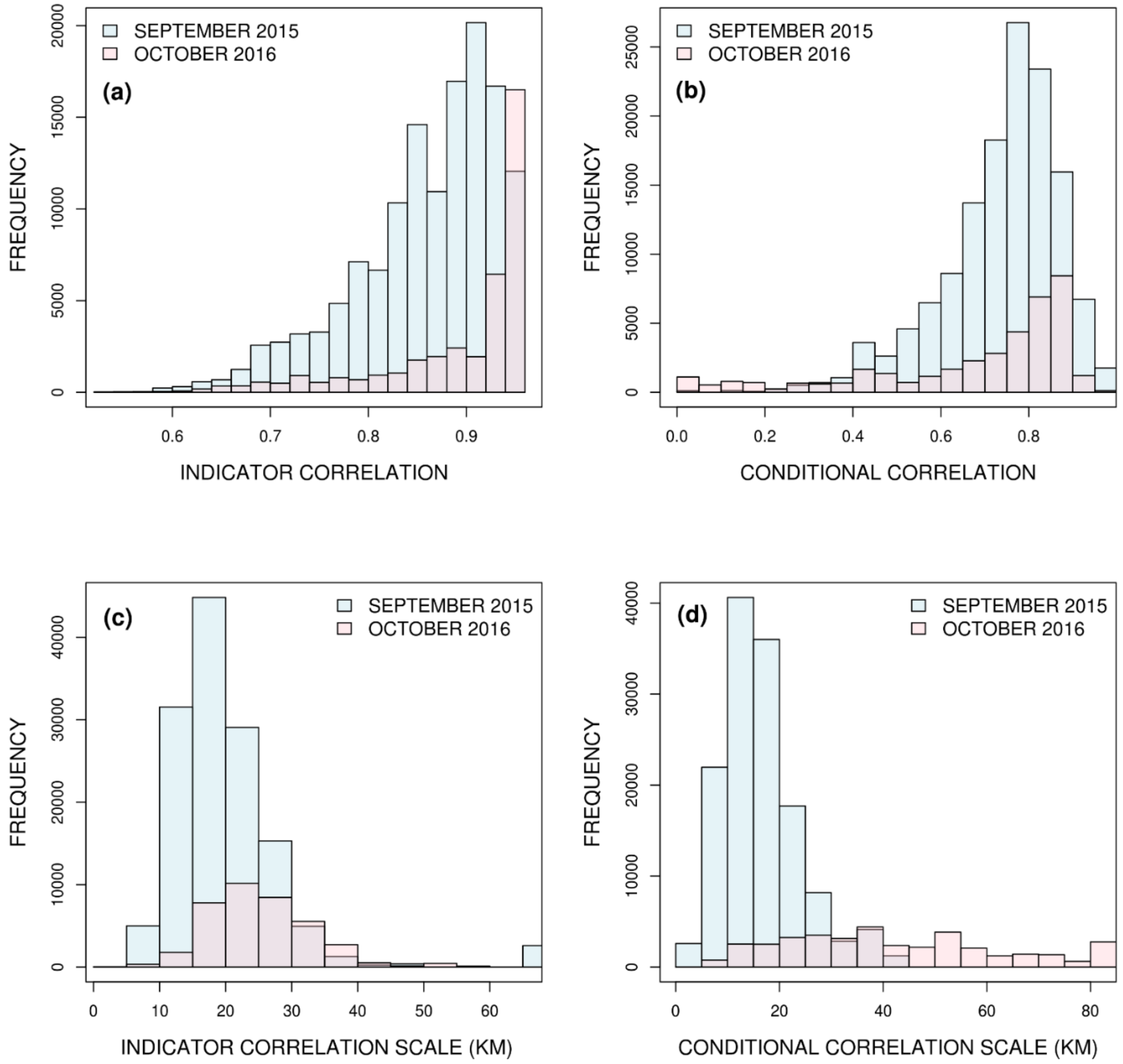


Figure 2.5: Histograms of (a) lag-0 indicator correlation, (b) lag-0 conditional correlation, (c) indicator correlation scale, and (d) conditional correlation scale for Sep 13-30, 2015, and Oct 709, 2016.

As such, we refer to the results for x=0 and x > 0 as unconditional and conditional performances, respectively. Also shown in the figure is the sample size, whose axis is shown at the right end of the plot on a logarithmic scale. Fig 2.6b shows the percent reduction in RMSE of the MFB-corrected radar, OCK, and adaptive CBPCK estimates relative to the radar-only QPE. In both figures, the conditioning threshold is cut off at about 34 mm, above which the results are very noisy due to the small sample size.

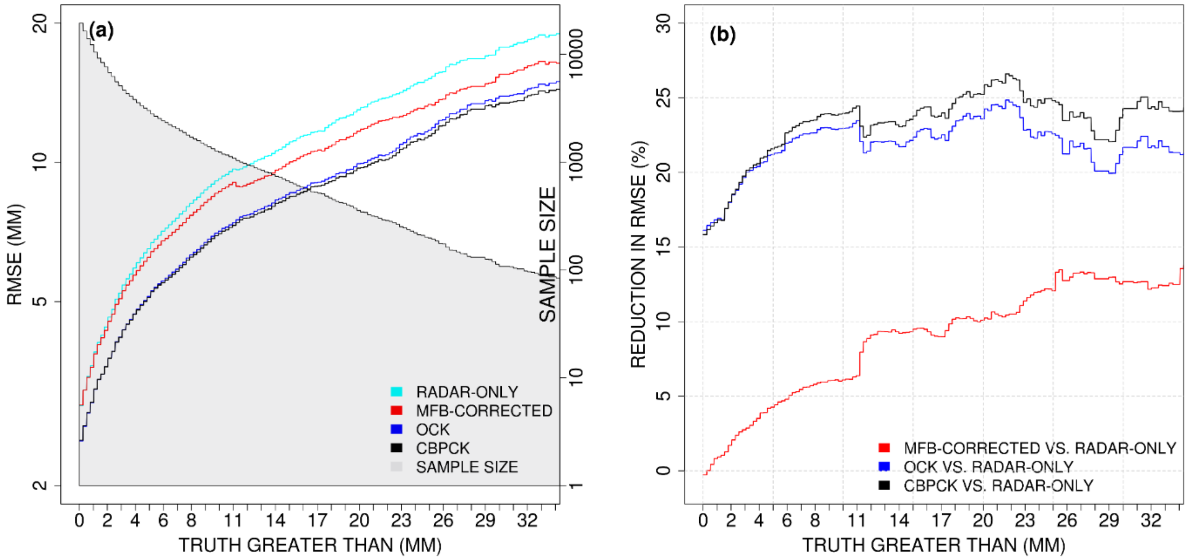


Figure 2.6: (a) RMSE of radar-only, MFB-corrected radar, OCK, and adaptive CBPCK estimates for Oct 7-9, 2016, conditional on truth exceeding the amount on the x-axis, (b) same as (a) but for percent reduction in RMSE of radar-only QPE by MFB-corrected radar, OCK, and adaptive CBPCK estimates.

Fig 2.6b may be summarized as follows. Both OCK and adaptive CBPCK reduce the unconditional RMSE of radar-only QPE by about 16%, whereas MFB correction provides no reduction. Adaptive CBPCK performs comparably to OCK unconditionally, an indication that adaptively optimizing α does produce the desired effect of improving the unconditional performance of CBPCK. For precipitation amounts exceeding 1.5 mm, adaptive CBPCK improves over OCK. At the conditioning amount of 24.5 mm, MFB correction, OCK and adaptive CBPCK reduce the conditional RMSE of radar-only QPE by about 13, 22 and 25%, respectively. The relatively modest improvement by adaptive CBPCK over OCK is a reflection that the CB is generally not very large in this period owing to the large spatial predictability in the well-developed precipitation systems in the east.

To show how the different estimates may compare, we show in Fig 2.7 the scatter plots of radar-only (upper-left), MFB-corrected radar (upper-right), OCK (lower-left), and adaptive CBPCK (lower-right) estimates vs. the verifying gauge precipitation.

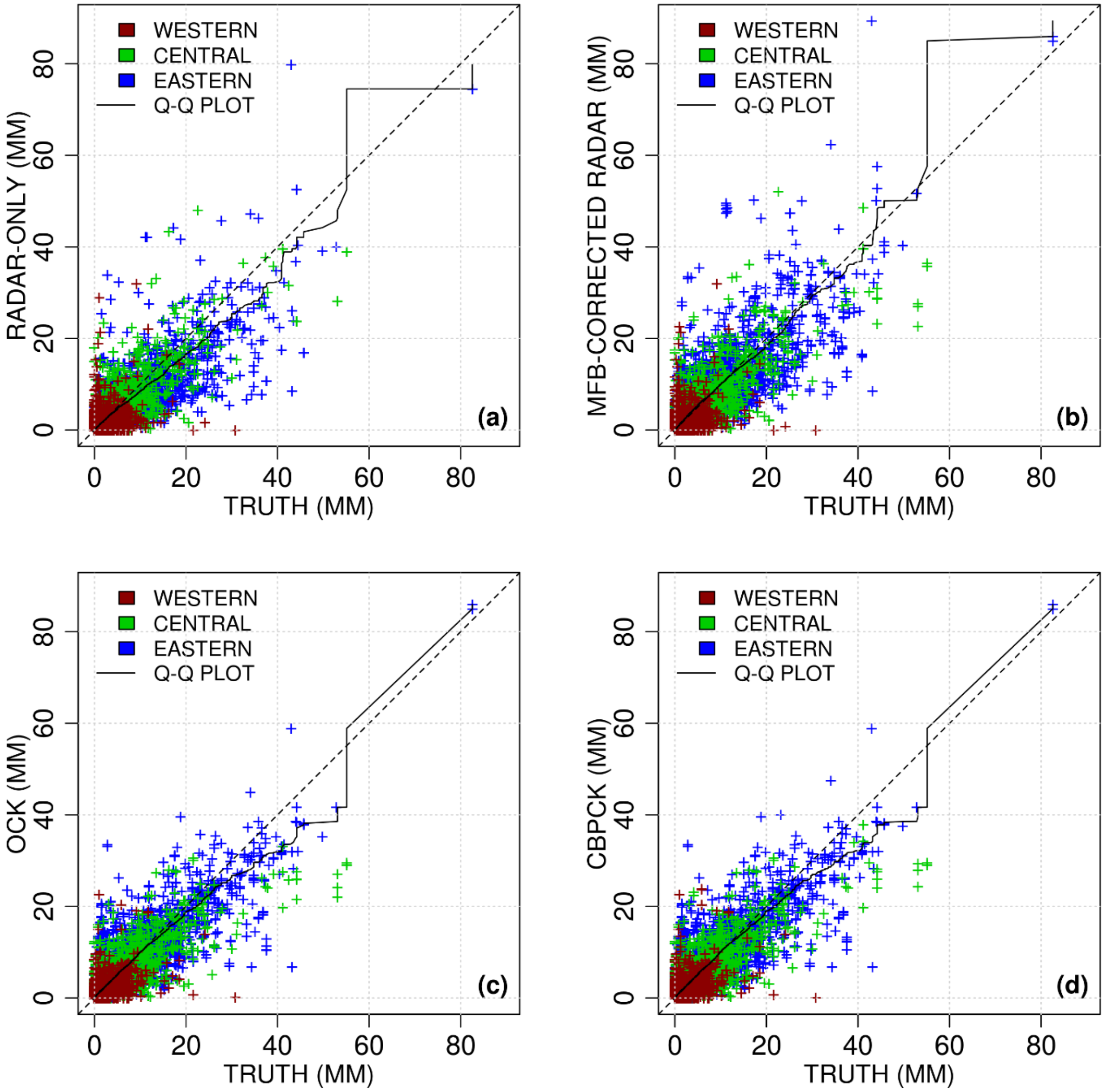


Figure 2.7: Scatter plots of (a) radar-only, (b) MFB-corrected radar, (c) OCK, and (d) adaptive CBPCK estimates for Oct 7-9, 2016, vs. truth.

The data points are color-coded by region so that the performance for different storms within the CONUS may be examined. The generally positive impact of MFB correction of radar QPE is readily seen in the figure as reflected by a nearly diagonal quantile-quantile (QQ) plot, but at the expense of inflating large, overestimated radar-only precipitation. Such mis-corrections occur because biases in radar-only QPE may be spatially nonuniform (Seo and Breidenbach 2000) or nonlinear, i.e., precipitation magnitude-dependent, which cannot be addressed by MFB correction alone. The lower panel shows that both OCK and adaptive CBPCK greatly reduce the scatter around the diagonal. Overall, the OCK and adaptive CBPCK estimates are very similar, as Fig 2.6 would suggest. For the eastern region, which encompasses Hurricane Matthew, the two estimates show little difference, an indication that there is little CB present owing to the large predictability, dense gauge networks, and generally favorable conditions for radar observation of precipitation. Though small, noticeable differences are seen in the plotting areas of 40 < truth < 60 (mm) and 20 < estimate < 40 (mm) for the Central Region, and 0 < truth < 20 (mm) and 0 < estimate < 20 (mm) for the Western Region, where a number of adaptive CBPCK estimates are closer to the diagonal than the OCK estimates.

Sep 13-30, 2015

This analysis period includes multiple mostly convective events of relatively low predictability in the Pacific Northwest, Midwest, Northeast, and Southeast of the US (see Figure 2.2). Fig 2.8 is the same as Fig 2.6 but for Sep 2015.

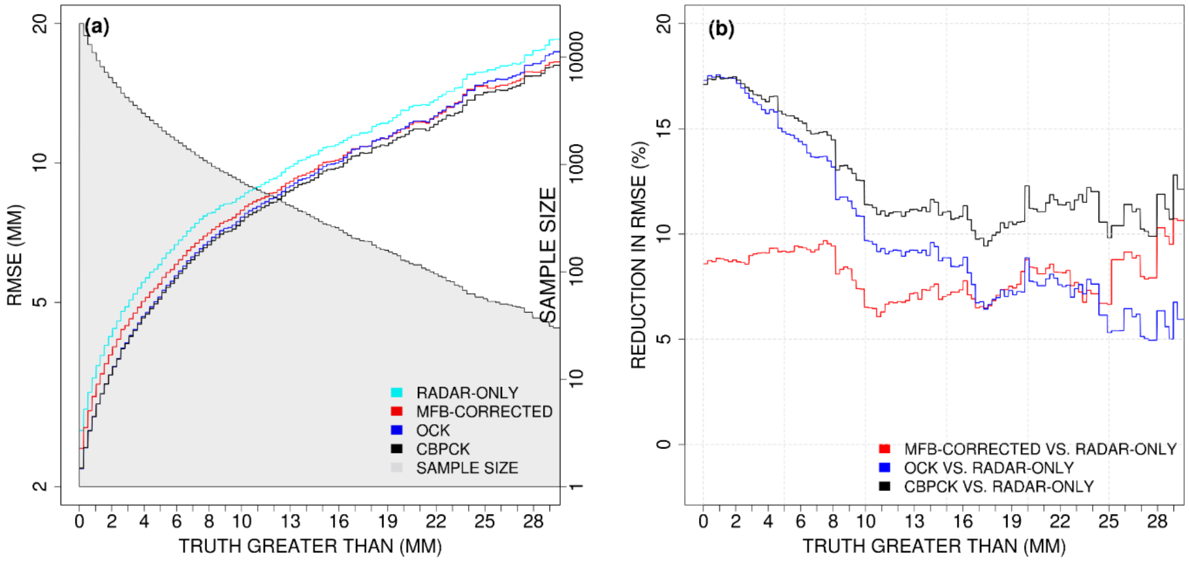


Figure 2.8: Same as figure 2.6 but for Sep 13-30, 2015

Figs 2.8ab may be summarized as follows. MFB correction reduces the unconditional RMSE of radar-only QPE by about 9%. Both OCK and adaptive CBPCK are able to increase the margin of reduction to about 17%. For these lower predictability events, however, the effectiveness of merging is reduced as the conditioning amount increases. For truth exceeding about 18 mm, the OCK estimates are no longer more accurate than the MFB-corrected radar in the mean squared error sense. The adaptive CBPCK estimates, on the other hand, perform better than the better of the OCK and the MFB-corrected radar estimates for all conditioning amounts. The larger margin of improvement by adaptive CBPCK over OCK reflects the presence of larger CB due to smaller spatial predictability in this period. About 65% of the true-validated estimates in Fig 2.8 are associated with $\alpha=0$, i.e., the adaptive CBPCK estimates are the same as the OCK estimates.

To assess the quality of the adaptive CBPCK estimates exclusively, it is necessary to consider only those estimates associated with non-zero α. To that end, we plot in Fig 2.9 the RMSEs of the radar-only, MFB-corrected radar, OCK, and adaptive CBPCK estimates (upper panels), and the percent reduction in RMSE of the radar-only QPE by the MFB-corrected radar, OCK, and adaptive CBPCK estimates (lower panels) for different ranges of positive α and conditioning amounts of truth. Also shown in the upper panels are the 90% confidence intervals for the OCK and adaptive CBPCK estimates obtained via bootstrapping. Fig 2.9 indicates that, when the CB is present (i.e., $\alpha > 0$), the adaptive CBPCK estimates are superior to the OCK estimates for truth exceeding 25.4 mm at a significance level of 0.10, but that, when all amounts of truth are considered, the improvement is not statistically significant. The latter is not at all surprising given the fact that smaller amounts of precipitation, for which little CB exists, far outnumber large amounts.

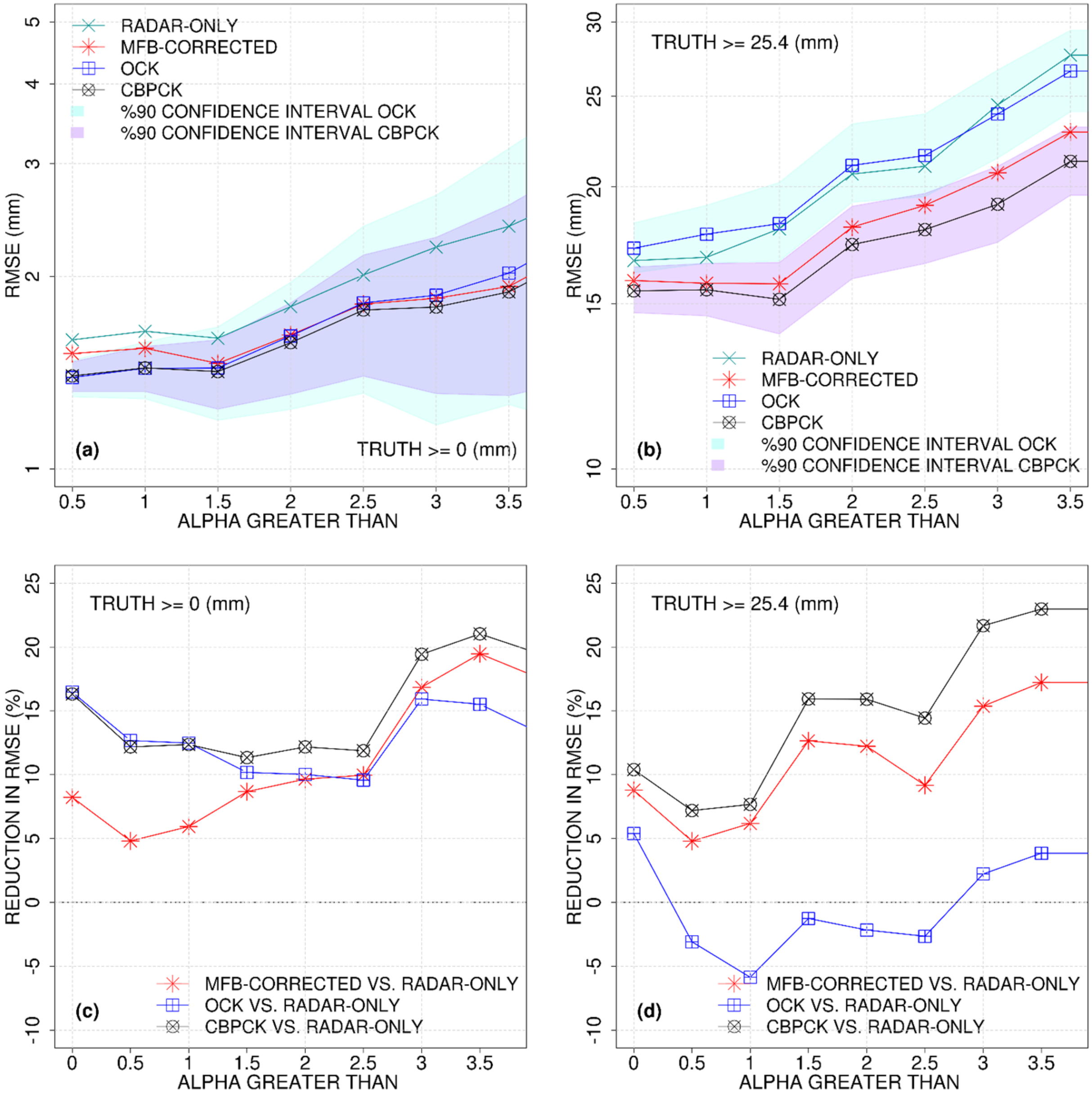


Figure 2.9: (a) RMSE of radar-only, MFB-corrected radar, OCK, and adaptive CBPCK estimates and 90% confidence interval of the OCK and adaptive CBPCK estimates, for Sep 13-30, 2016, conditional on α exceeding the value on the x-axis for truth exceeding zero, (b) same as (a) but for truth exceeding 25.4 mm, (c) same as (a) but for percent reduction in RMSE of radar-only QPE by MFB-corrected radar, OCK, and adaptive CBPCK estimates, (d) same as (c) but for truth exceeding 25.4 mm.

Fig 2.10 shows the scatter plots of the estimates vs. the observed. Reduction of bias and scatter due to bias correction and merging, respectively, is readily seen. For this period, larger differences are seen between the OCK and adaptive CBPCK estimates than in Oct 2016, particularly in the

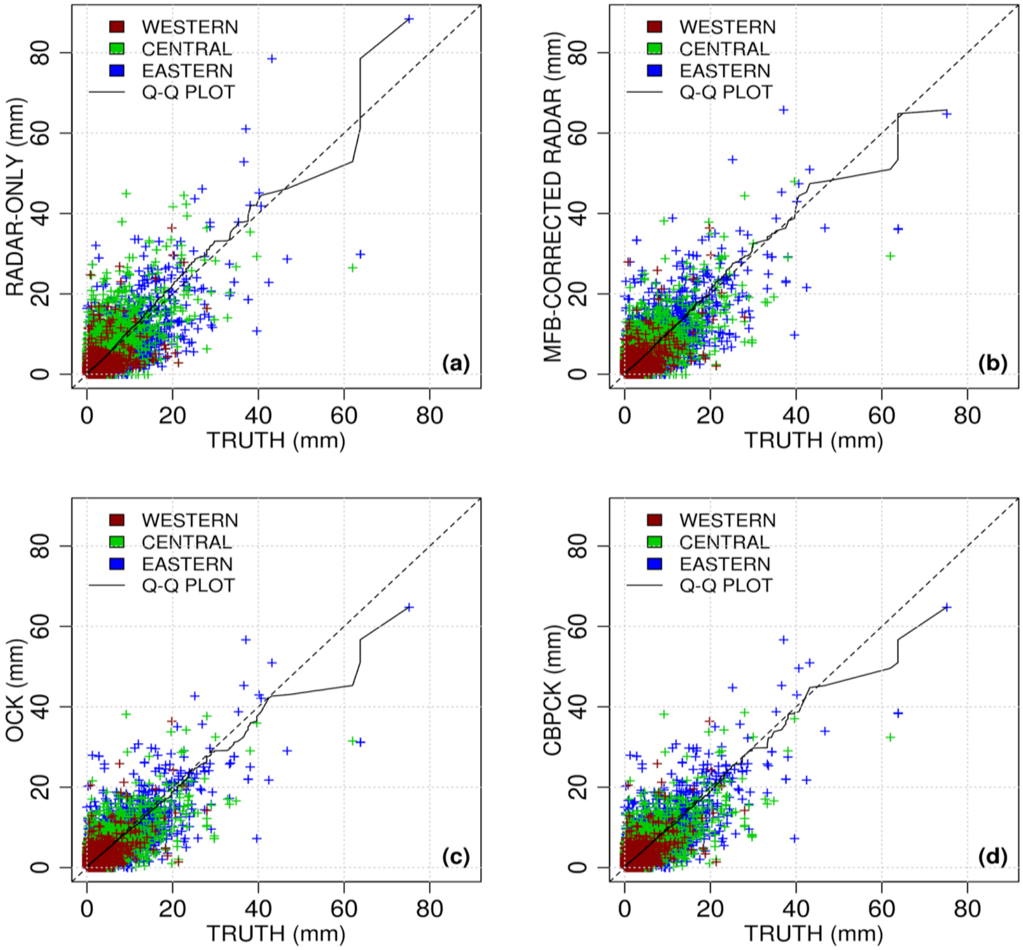


Figure 2.10: Same as figure 7 but for Sep 13-30, 2015.

eastern region. The general direction of change is that the CBPCK estimates tend to decrease and increase very small and larger OCK estimates, respectively.

Figs 2.11a, 2.11b, and 2.11c show the $(m_e - m_o)^2$, $(\sigma_e - \sigma_o)^2$, and $\rho_{e,o}$ terms in the MSE decomposition of Eq. (2.22), respectively. They are for the radar-only, MFB-corrected, OCK, and adaptive CBPCK estimates conditioned on the truth exceeding 0, 6.4, 12.7, and 25.4 mm. In each figure, the sample statistics for the Sep 2015 and Oct 2016 periods are plotted on the x- and y-axes, respectively. For each conditioning threshold, the sample statistics for the four different estimates are connected with a dashed line. The most desirable estimates would place the sample statistics of $(m_e - m_o)^2$ and $(\sigma_e - \sigma_o)^2$ closest to the lower-left corner in Figs 2.11a and 2.11b, respectively, and that of $\rho_{e,o}$ closest to the upper-right corner in Fig 2.11c. Fig 2.11 indicates that the MFB-corrected estimates are least biased in the mean and standard deviation, followed by the CBPCK estimates, and that the CBPCK estimates are most strongly correlated with the truth, followed by the OCK estimates. The first result is not very surprising in that the sole purpose of MFB correction is to reduce bias in the mean. Also, being conditional expectation operators, OCK and adaptive CBPCK necessarily reduce variability due to averaging. One may avoid such smoothing by performing conditional simulation (Deutsch and Journel 1992, Seo et al. 2000) using OCK or adaptive CBPCK in an ensemble framework. Such an approach, however, is computationally too expensive to be practical for real-time QPE. Fig 2.11 indicates that OCK and adaptive CBPCK reduce MSE over MFB-corrected radar by a combination of smoothing and improved correlation with the truth (i.e., reduced σ_e and increased $\rho_{e,o}$in Eq. (2.22), respectively), and that the adaptive CBPCK estimates are superior to the OCK estimates in all categories except in unconditional bias in the mean and conditional bias in the standard deviation for the 25.4-mm threshold.

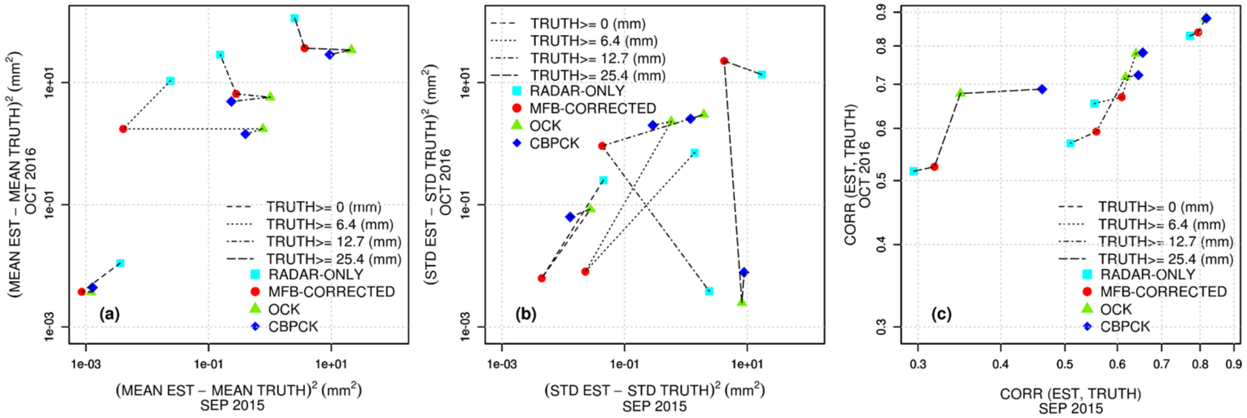


Figure 2.11: Decomposition of MSE of radar-only, MFB-corrected, OCK, and adaptive CBPCK estimates for Sep 13-30, 2015 (x-axis) and Oct 7-9, 2016 (y-axis) into (a) $(m_e - m_o)^2$, (b) $(\sigma_e - \sigma_o)^2$, and (c) $\rho_{e,o}$ conditioned on truth exceeding 0, 6.4, 12.7, and 25.4 mm.

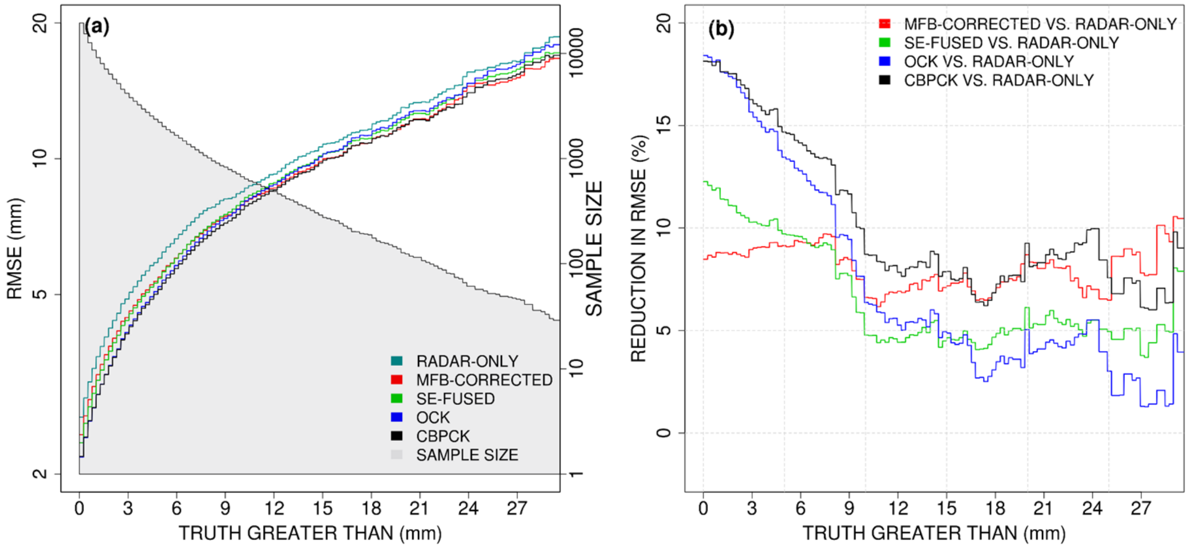


Figure 2.12: Same as figure. 2.8 but for radar-only, MFB-corrected radar, SE-fused, OCK, and adaptive CBPCK estimates. SE fuses the MFB-corrected radar and MFB-corrected SCaMPR. OCK and adaptive CBPCK merge the SE-fused estimates and rain gauge data.

2.4.2 Fusion of SCaMPR QPE

Fig 2.12 is the same as Fig 2.8, but for the radar-only, MFB-corrected radar, SE-fused, OCK, and adaptive CBPCK estimates for Sep 2015 over the CONUS. The SE results are based on fusing the MFB-corrected radar, and MFB-corrected SCaMPR estimates as described in Subsection 2.3.3 (see also Fig 2.4).

The OCK and adaptive CBPCK results are based on merging the gauge data with the fused QPE. Fig 2.12 indicates that the MFB-corrected radar and adaptive CBPCK provide significantly better conditional performance than the others. Fig 2.12b shows that the percent reduction in unconditional RMSE by the MFB-corrected radar, SE-fused, OCK, and adaptive CBPCK over the radar-only QPE is about 8, 12, 18 and 18%, respectively. As the conditioning amount increases, however, the accuracy of the OCK estimates deteriorates and, at the conditioning amount of about 15 mm, it falls below that of the fused estimates. The adaptive CBPCK estimates, on the other hand, remain better than the MFB-corrected radar for conditioning amounts of up to 25 mm. The above results demonstrate the adaptive CBPCK's ability to improve conditional performance over OCK while performing comparably to OCK in the unconditional sense.

We now assess the impact of the MFB-corrected SCaMPR QPE in radar-satellite fusion by comparatively evaluating the SE-fused vs. the MFB-corrected radar estimates. Note that, in this comparison, we are not assessing the relative value of the radar- and SCaMPR-only estimates, but that of the MFB-corrected radar and MFB-corrected SCaMPR estimates. Accordingly, additional factors such as the rain gauge network density, spatiotemporal variability of precipitation, the skill in radar-only QPE, and the efficacy of MFB correction come into play. Fig 2.12b shows that MFB correction reduces the RMSE of the radar-only QPE by about 8% over the CONUS, and the fused QPE additionally reduces the RMSE by about 4%. The tile-specific results indicate that, for the western half of the CONUS (i.e., Tiles 1 through 3), the additional reduction by the fused QPE is over 6% whereas for the eastern half it is only about 3%. Compared to the western half of the CONUS, the eastern half has significantly denser rain gauge networks (see Fig 2.4), and is generally more favorable for radar QPE. The latter point may be seen in the tile-specific correlation coefficient of radar-only QPE with rain gauge observations; for Sep 2015, the correlation is 0.64, 0.44, 0.66, 0.77, 0.77 and 0.77 for Tiles 1, 2, 3, 4, 5, and 6, respectively. It is surmised that, for the eastern half, the combination of the skillful radar QPE and the dense rain gauge networks is able to produce significantly more accurate MFB-corrected radar QPE, and that the comparative skill of the MFB-corrected SCaMPR estimates is too small to improve on the former significantly. The western region, on the other hand, has substantially lower density of rain gauges and are not very favorable for radar QPE. As such, the MFB-corrected SCaMPR is able to provide larger improvement. Co-examination of Figs 2.8 and 2.12 indicates that, whereas the addition of the SCaMPR QPE does improve the accuracy of the final merged QPE in the unconditional sense, in particular in the western half of the CONUS, it deteriorates the conditional performance. Figs 2.8b and 2.12b show that the added value of the SCaMPR estimates is lost when the precipitation amount exceeds about 1.5 mm, above which merging rain gauge data and MFB-corrected radar QPE via adaptive CBPCK is superior. The above findings suggest that the SCaMPR product should be used selectively in the multisensor QPE framework of this work.

Computationally, adaptive CBPCK requires solving an $(n_g + n_r)$-dimensional linear system multiple times whereas OCK requires solving a comparable system only once. In the above, n_g and n_r are usually on the orders of 10 and 1, respectively, and α may range from 0 to 4 incremented by 1. With the naïve optimization of α used in this work, adaptive CBPCK is hence several times more expensive than OCK. To improve understanding of the dependence of the CBPCK solution on α, and to develop a more effective and computationally efficient approach for its optimization, additional research is needed.

2.5 Conclusions and future research recommendations

The principal conclusion of this work is that, to produce multisensor estimates that are more accurate than the ingredient QPEs for all precipitation amounts, it is necessary to address the CB and that adaptive CBPCK described in this chapter improves the estimation of significant amounts of precipitation by explicitly considering the CB. It is shown that, beyond the reduction in RMSE due to MFB, both OCK and adaptive CBPCK additionally reduce the unconditional RMSE of radar-only QPE by 16 and 9 % over the CONUS for the more and less predictable Oct 7-9, 2016, and Sep 13-30, 2015, events, respectively, and that adaptive CBPCK improves over OCK for estimation of hourly precipitation exceeding about 1 mm. Jointly, MFB correction and adaptive CBPCK reduce the RMSE of the radar-only QPE by about 16 to 26% for the more predictable Oct

7-9, 2016, events and by about 10 to 17% for the less predictable Sep 13-30, 2015, events for all ranges of precipitation amounts.

It is shown that for the Sep 2015 events, fusing the MFB-corrected radar QPE with the MFB-corrected SCaMPR QPE reduces the unconditional RMSE of radar-only QPE by about 12% over the CONUS, whereas the reduction by MFB-corrected radar QPE alone over radar-only QPE is about 8%. For the western half of the CONUS, where the rain gauge network is sparser, and the radar QPE is less skillful, the margin of reduction increases to 6% from the above 4%. The conditional performance of the fused QPE, however, falls below that of the MFB-corrected radar QPE as the conditioning amount exceeds about 7 mm of hourly precipitation. The above suggests that the SCaMPR product should be used selectively in the multisensor QPE framework of this study.

The error variance estimates from adaptive CBPCK were not used in this work. Co-utilizing both the estimate and the error variance is likely to improve optimization of α. Additional evaluation is needed to assess the performance of adaptive CBPCK further, and to optimize its parameters. We note here that the algorithm described in this work is being comparatively evaluated with Stage IV (Nelson et al. 2010) in real time for possible operational implementation in MRMS (Tang et al. 2019), and the results will be reported in the near future.

Chapter 3: CBP-MLR and CompMLR: Theory and Application to Multimodel Streamflow Prediction

3.1 Introduction

Streamflow is arguably the most important predictand in operational hydrology and water management. With the fast-increasing availability of multiple streamflow forecasts from different sources in many parts of the world (Muhammad et al., 2018), objectively combining forecasts in ways that will yield consistently superior forecast products is of great wide interest. For the above purpose, multiple linear regression (MLR) is an extremely attractive technique because of its simplicity and interpretability. MLR also forms the basis for the standard Bayesian normal conjugate linear model if an extension to Bayesian modeling, such as Bayesian Model Averaging (BMA, Madigan and Raftery 1994; Hoeting et al. 1999; Raftery et al. 2005, 2015; Duan et al. 2007), is desired. MLR is arguably one of the most popular modeling and prediction techniques used in a wide range of areas. Khan et al. (2013) used MLR for nonlinear fuzzy set-based uncertainty propagation for improved prediction of dissolved oxygen. Khanmohammadi et al. (2018) used MLR for evapotranspiration trend calculation. Carl and Kühn (2008) used linear regression and wavelet analysis to analyze spatial ecological data. Diks and Vrugt (2010) compared point forecast accuracy of model averaging methods in hydrologic applications, which included MLR. It is well known in statistics and econometrics that observation error in the predictors may introduce significant low and high biases in the least-squares solution over the upper and lower tails of the predictand, respectively (Fuller, 1987; Hausman, 2001; Seber and Wild, 1989). In regression, the above effect is referred to as regression dilution, which results in attenuation, or conditional, bias in the regression coefficients (Frost and Thompson, 2000; Hughes, 1993). Theil (1961) assessed the impact of measurement error on ordinary least squares (OLS) estimation in regressions with two predictors measured with error. He showed that the multivariate attenuation factor, which applies to all coefficients and generalizes the standard attenuation factor in univariate regressions, increases with increasing correlation between the two explanatory variables (Abel, 2017). Levi (1973) examined the case with two or more predictors where one variable is observed with measurement error and the rest are observed without error. He showed that the OLS estimate of the coefficient for the variable measured with error is attenuated toward zero (Abel, 2017).

To address the detrimental effects of conditional bias (CB) on estimation of extremes, Seo (2013) introduced a linear estimation method that minimizes the weighted sum of error variance and expectation of the Type-II error squared. Recall that the Type-I and II errors are associated with false alarm and failure to detect, respectively. Whereas the Type-I CB may be reduced by calibration, the Type-II CB cannot (Seo et al., 2018a, 2018b; Wilks, 2005). As such, reducing the Type-II CB is of particular interest in estimation and prediction of extremes. When cast in the form of kriging, CB-penalized linear estimation yields CB-penalized kriging (CBPK), which has been shown to broadly outperform conventional kriging in the prediction of high flows and estimation of heavy to extreme rainfall (Brown and Seo, 2013; Jozaghi et al., 2019; Kim et al., 2018; Seo, 2013; Seo et al., 2014). When cast in the form of the Kalman filter (KF), the method yields the CB-penalized KF (CBPKF, Seo et al. 2018a, b; Shen et al. 2019) and, in ensemble form, the CB-penalized ensemble KF (CBEnKF, Lee et al. 2019). CBPKF and CBEnKF have been shown to outperform KF and EnKF for the prediction of extremes and floods, respectively (Lee et al., 2019; Shen et al., 2019). In this work, we cast conditional bias-penalized linear estimation in the form of MLR and apply the resulting conditional bias-penalized multiple linear regression (CBP-MLR) to multi-model streamflow prediction for multiple forecast groups in the NWS Middle Atlantic River

Forecast Center's service area. A forecast group is an NWS term referring to a collection of forecast locations generally making up some portion of the same river basin. Whereas CBP-MLR improves prediction over tails, it yields degraded performance near the median in the mean squared error (MSE) sense. To retain MLR-like performance near the median while harnessing the skills of CBP-MLR over tails, we develop composite MLR (CompMLR), which linearly weight-averages the MLR and CBP-MLR estimates according to the standard-normalized MLR estimate.

In multi-model streamflow prediction, the predictors are not observations but forecasts of streamflow valid at some future times. Accordingly, forecast errors in multi-model streamflow prediction act as observation errors in MLR. Streamflow forecasts are subject to forcing, hydrologic and anthropogenic uncertainties modulated by hydroclimatological, hydrometeorological, hydrologic and hydraulic processes, as well as human control of storage and movement of water (Krzysztofowicz, 1999; Seo et al., 2006). These uncertainties result in errors in streamflow forecasts that are both systematic and random in nature and vary in space and time, in magnitude, and with lead time (Alizadeh et al., 2020a). The wide range of possible errors in streamflow forecasts means that regression-based multi-model streamflow prediction is subject to CB of varying magnitude, and is likely to benefit from CBP- and CompMLR. The significant new contributions of this paper are the development of CBP- and CompMLR, assessment of predictive skill of operationally produced streamflow forecasts in the MARFC's service area, comparative evaluation of multiple streamflow forecasts and MLR-based multi-model prediction, and advances in understanding of the multi-objective nature of multi-model streamflow prediction. This paper is organized as follows. In Section 2, we describe the CBP- and CompMLR methods. Section 3 describes how evaluation is carried out. Section 4 presents the results. Section 5 offers discussion.

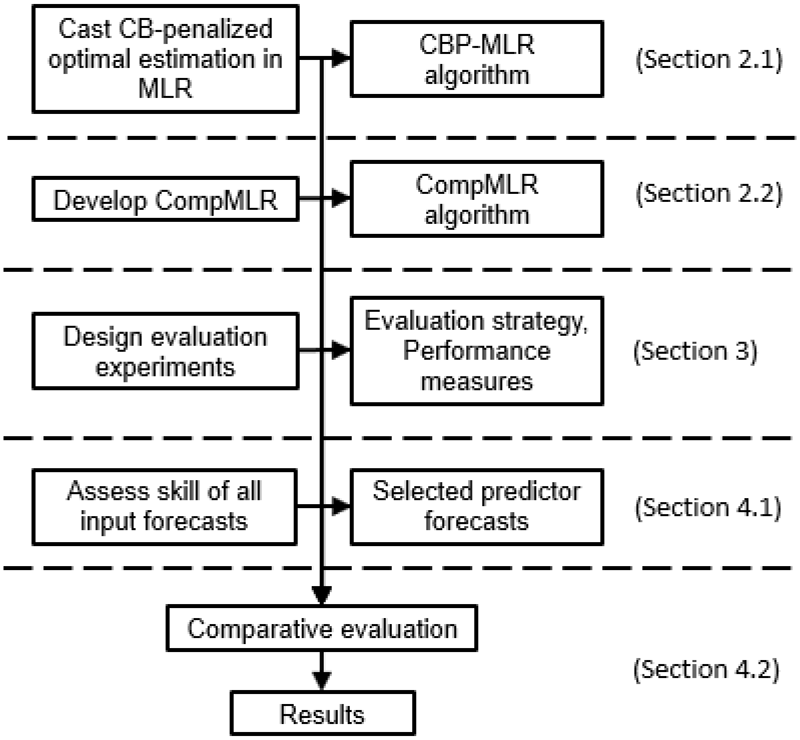


Figure 3.1: Schematic of CompMLR, a linear weighted average of MLR and CBP-MLR estimates

Section 6 provides the conclusions and future research recommendations. Fig 3.1 shows the flowchart of this study.

3.2 Methods

In this section, we describe the formulations of CBP- and CompMLR. The development of CBP-MLR is analogous to that of other conditional bias-penalized estimation techniques (Lee et al., 2019; Seo, 2013; Shen et al., 2019). The main difference is that CBP-MLR rests solely on least-squares minimization whereas the others build on Fisher and Bayesian estimation (Schweppe, 1973). The identity between the two estimates follows from the fact that the Fisher solution, or the maximum likelihood solution under normality, is identical to the weighted least squares solution (Schweppe, 1973).

3.2.1 Conditional bias-penalized multiple linear regression

The MLR model is given by:

$$Y = Xb + e \tag{3.1}$$

where Y is the $(n \times 1)$ vector of the predictand or the response variable, X is the $(n \times m)$ matrix of the predictors or the explanatory variables, b is the $(m \times 1)$ coefficient vector and e is the $(n \times 1)$ vector of independent and identically distributed zero-mean normal errors, $e \sim N(0, \sigma^2 I)$. The OLS estimate for b in Eq. (3.1), $\hat{b}$, is obtained by solving $\min_{b \in \mathbb{R}^m} \|Y - Xb\|^2$ where $\|\cdot\|$ denotes the Frobenius norm as**:**

$$\hat{b} = (X^T X)^{-1} X^T Y \tag{3.2}$$

The least squares-fitted predictands, $\hat{Y}$, are given by:

$$\hat{Y} = X\hat{b} = X(X^T X)^{-1} X^T Y \tag{3.3}$$

Variance of $\hat{b}$ is given by:

$$\mathrm{E}[(\hat{b} - b)\left(\hat{b} - b\right)^T] = \sigma^2 (X^T X)^{-1} \tag{3.4}$$

To derive CBP-MLR, we first define the Type-II error of $\hat{Y}$, $E\left[\hat{Y}\middle|Y = y\right] - y$, where Y, $\hat{Y}$ and y denote the unknown truth, its estimate, and its realization, respectively (Jolliffe and Stephenson, 2011). The Type-I error, on the other hand, is defined as $E\left[Y\middle|\hat{Y} = \hat{y}\right] - \hat{y}$, where $\hat{y}$ denotes a realization of $\hat{Y}$. CBP-MLR jointly reduces systematic Type-II error, or Type-II CB, and the sum of errors squared. For the CB penalty, we use the expectation of the Type-II error squared as in Seo (2013), Lee et al. (2019) and Shen et al. (2019):

$$\Sigma_{CB} = E[(Y - E[\hat{Y}|Y])^T (Y - E[\hat{Y}|Y])] \tag{3.5}$$

Using Eq. (3.3), we may write the Type-II CB in Eq. (3.5) as:

$$Y - E[\hat{Y}|Y] = Y - E[X|Y]\hat{b} \tag{3.6}$$

Note that, because X is not a random variable, the expectation operation of $E[X|Y]$ in Eq. (3.6) is meaningless. The expression is used here only to indicate that it is an estimate of X and depends on Y. To model $E[X|Y]$, we use MLR (i.e., reverse MLR) just as we used MLR to approximate $E[Y|X]$ in Eq. (3.3):

$$\hat{E}[X|Y] = Y(Y^TY)^{-1}Y^TX \tag{3.7}$$

where $\hat{E}[\,]$ signifies that it is only an estimate. We may then write the Type-II error of $\hat{Y}$ as:

$$Y - E[\hat{Y}|Y] = Y[I - (Y^TY)^{-1}Y^TX\hat{b}] \tag{3.8}$$

Whereas MLR minimizes only the sum of errors squared, $\Sigma_{EV} = (Y - \hat{Y})^T(Y - \hat{Y})$, CBP-MLR minimizes a linearly weighted sum of Σ_{EV} and the sum of the Type-II error squared:

$$J = \Sigma_{EV} + \alpha\Sigma_{CB} \tag{3.9a}$$

$$= \left(Y - \hat{Y}\right)^T\left(Y - \hat{Y}\right) + \alpha\left(Y - E_{\hat{Y}}\left[\hat{Y}\middle|Y\right]\right)^T\left(Y - E_{\hat{Y}}\left[\hat{Y}\middle|Y\right]\right) \tag{3.9b}$$

$$= \left(Y - X\hat{b}_{cb}\right)^T\left(Y - X\hat{b}_{cb}\right) + \alpha\left(I - C\hat{b}_{cb}\right)^T Y^TY\left(I - C\hat{b}_{cb}\right) \tag{3.9c}$$

where α denotes the weight given to the CB penalty and $C = (Y^TY)^{-1}Y^TX$. Minimizing J with respect to $\hat{b}_{cb}$, we have the following least-squares solution for the CBP-MLR coefficient:

$$\hat{b}_{cb} = (X^TX + \alpha C^TY^TYC)^{-1}(X^TY + \alpha C^TY^TY) \tag{3.10a}$$

$$= (1 + \alpha)[X^TX + \alpha X^TY(Y^TY)^{-1}Y^TX]^{-1}X^TY \tag{3.10b}$$

Eq. (3.10b) follows from Eq. (3.10a) via the matrix inversion lemma (Schweppe, 1973). Eq. (3.10b) is somewhat similar in appearance to Tikhonov regularization (Tikhonov et al., 1995; Tikhonov and Arsenin, 1977) or ridge regression (Hoerl, 1962; Hoerl and Kennard, 1970), but the nature of the penalty is very much different. Whereas ridge regression or its L_1 version, the Least

Absolute Selection and Shrinkage Operator (LASSO, Tibshirani 1996, 1997), seeks to shrink the coefficients, CBP-MLR looks to stretch them as shown in the Results Section.

In Eq. (3.10), α may be seen as a tuning parameter that controls the strength of the penalty as λ does in ridge regression or LASSO. If there is no penalty for CB (i.e., $\alpha = 0$), $\hat{b}_{cb}$ is reduced to the conventional linear regression coefficient in Eq. (3.2). If in the other extreme, the objective function is made solely of the CB penalty (i.e., $\alpha \to \infty$), we have for $\hat{b}_{cb}$:

$$\hat{b}_{cb} = [X^T Y (Y^T Y)^{-1} Y^T X]^{-1} X^T Y \tag{3.11}$$

Note that, with Eq. (3.11), the CB penalty term in Eq. (3.9) vanishes. It is instructive to consider the above solution in the context of simple linear regression. With $m = 1$, Eq. (3.9) with $\alpha = 0$ and $\alpha \to \infty$ yields $\hat{b}_{cb} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$ and $\hat{b}_{cb} = \frac{1}{\rho_{XY}} \frac{\sigma_Y}{\sigma_X}$, respectively, when expressed in terms of statistical parameters where ρ_{XY} denotes the correlation between X and Y, and σ_X and σ_X denote standard deviation of X and Y, respectively. Note that the second expression for $\hat{b}_{cb}$ above is the slope in linear reverse regression (Conway and Roberts, 1983; Goldberger, 1984; Green and Ferber, 1984). The CBP-MLR solution, therefore, lies between the (forward) regression slope and the reverse regression slope, and its slope depends on the relative importance between the MSE and Type-II error squared penalties. Unlike MLR, the variance of $\hat{b}_{cb}$ does not have an analytical solution. One may, however, obtain an approximation by expressing Eq. (3.9) analogously to Eq. (3.3):

$$\hat{b}_{cb} = (A^T A)^{-1} A^T Y \tag{3.12}$$

where

$$A = (XX^T)^{-1} X (1 + \alpha)^{-1} [X^T X + \alpha X^T Y (Y^T Y)^{-1} Y^T X] \tag{3.13}$$

With the above, one may approximate variance of $\hat{b}_{cb}$ as:

$$\mathrm{E}[(\hat{b}_{cb} - b)\big(\hat{b}_{cb} - b\big)^T] \approx \sigma^2(A^T A)^{-1} \tag{3.14}$$

The choice of α depends on whether the state of the system that manifests the predictor-predictand relationship being observed is in a normal range or in an extreme range. If the system is at or near its median state, α should be set to near zero so that CBP-MLR is effectively reduced to MLR and produce the least-squares solution. If the system is in an extreme state, α should be set to a large value so that CB is reduced as much as possible. Because the true state of the system is not known, prescribing α dynamically is a challenge. In this work, we optimize α in the training period for maximum conditional performance, i.e., performance over the tails, and linearly weight-average the MLR and CBP-MLR estimates according to the exceedance probability of the MLR prediction. For example, if the sample size for the training data set is n, one may obtain the n MLR estimates first, sort them and assign exceedance probabilities using, e.g., the Weibull plotting position $r/(n+1)$ where r denotes the rank from largest to smallest (Makkonen, 2006). If the exceedance probability of the MLR estimate is near the median (i.e., 0.5), CB is likely to be small and we weigh more heavily the MLR estimate. If the exceedance probability of the MLR estimate is very small, CB is likely to be large and we weigh more heavily the CBP-MLR estimate. In the real-time mode, one may estimate the exceedance probability of a real-time MLR estimate via a lookup table or interpolation of the training results. The above approach, CompMLR, amounts to performing two regressions, MLR and CBP-MPR, and weight-averaging the two in the transition zone. Fig 3.2 illustrates the concept in which the diagonal line (dashed) is also shown for reference. The details of CompMLR follow below.

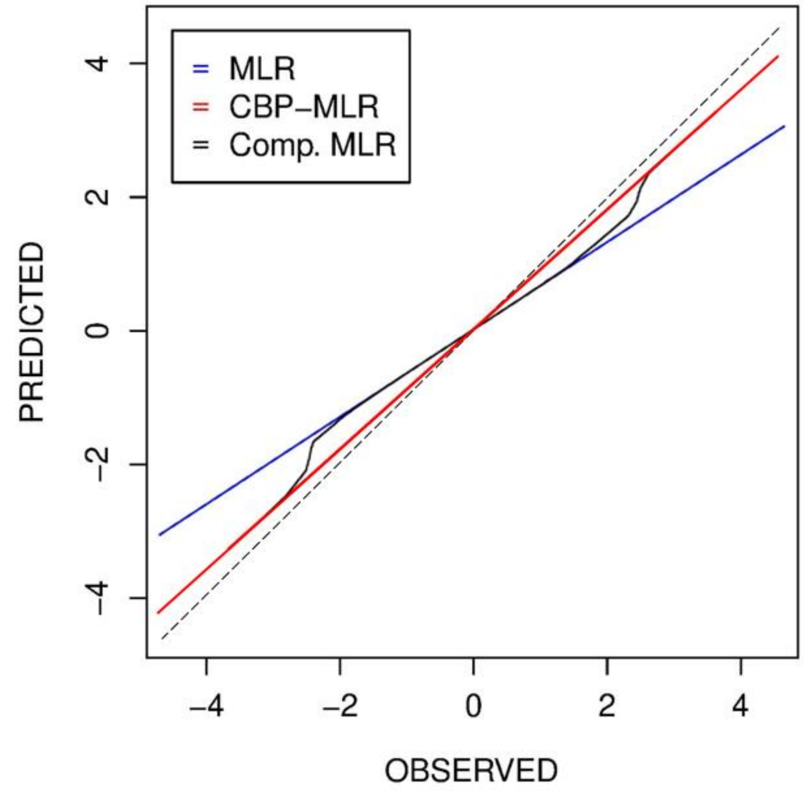


Figure 3.2: Illustration of CompMLR (black) as a composite of MLR (blue) and CBP-MLR (red) regressions.

3.2.2 Composite MLR

CompMLR linearly weight-averages the MLR and CBP-MLR estimates according to:

$$\hat{y}_{co,i} = \varpi_i \hat{y}_i + (1 - \varpi_i)\hat{y}_{cb,i} \tag{3.15}$$

where $\hat{y}_i$ and $\hat{y}_{cb,i}$ denote the MLR and CBP-MLR estimates for the i-th prediction, ϖ_i denotes the weight for the i-th MLR estimate, and $\hat{y}_{co,i}$ denotes the i-th CompMLR estimate. There are a number of different ways to prescribe ϖ_i (see Kim et al. 2018; Shen et al. 2019 for examples in conditional bias-penalized cokriging and CBPKF, respectively). We experimented with several for multi-model streamflow prediction and found the following to be effective:

$$\varpi_i = e^{-|\beta z_i|^{\gamma}}, \gamma > 0 \tag{3.16}$$

In the above, z_i denotes the standard normal deviate of the i-th MLR estimate $\hat{y}_i$, β and γ denote the width and smoothing parameters to be determined, respectively. If the i-th MLR estimate is at the median, we have $z_i = 0$ and hence $\varpi_i = 1$. If it is in a tail end, we have a very large $|z_i|$ and hence a very small ϖ_i. The width parameter β controls the width of the sub-range of z_i centered around the median (zero in the standard normal space) over which MLR is heavily weighted in Eq. (3.15). The smoothing parameter γ controls the shape of the curve connecting the two slopes (see Fig 3.2). A small γ renders ϖ_i to approach a finite-width pulse centered at $z_i = 0$. A large γ renders ϖ_i to vary very smoothly from one regression line to the other.

Before we describe how the parameters in Eq.(3.16) are estimated, it is useful to describe the nature of the regression problem posed by CompMLR. Because the MLR estimate, and hence z_i in Eq. (3.16), are never perfect, ϖ_i as described above can only be prescribed suboptimally. The goal of CompMLR is, therefore, to perform as closely to MLR as possible when the (unknown) true state of the system is predicted to be near the median and as closely to CBP-MLR as possible when it is predicted to be in the tails. Because ϖ_i is meant to reflect only the relative magnitude of the verifying truth, the predictive skill sought in z_i is rank correlation rather than product-moment correlation. Given the above picture, one may view CompMLR as a nonlinear regression comprising two linear regressions with a transition zone in between (see Fig 3.2). A useful way to assess the upper bound of the performance of CompMLR is to prescribe z_i in Eq. (3.16) perfectly by using the verifying observation in place of the MLR estimate (see, e.g., Shen et al. 2019 for an example using CBPKF). With the uncertainty in prescribing z_i eliminated, such "idealized" CompMLR should perform very closely to MLR near the median and CBP-MLR over the tails, and further guide the modeling of ϖ_i for improved performance.

As described above, there are three parameters to be optimized in CompMLR: α, β and γ. Given that MLR and CBP-MLR are least-squares solutions, it behooves to minimize the MSE of the predictions for parameter optimization. The MSE reflects correlation and biases in mean and standard deviation (Murphy and Winkler, 1987; Nelson et al., 2010), all of which are important performance attributes for streamflow forecasting. Minimizing only the unconditional MSE, however, is too limiting a criterion for CBP-MLR as elaborated below. In the unconditional MSE sense, CBP-MLR is inferior to MLR. As such, CBP-MLR reduces unconditional MSE over MLR only if significant CB exists. One may hence expect MLR to frequently outperform CBP-MLR in the unconditional MSE sense, resulting in α of zero. In CompMLR, we allow the unconditional MSE of CBP-MLR to exceed that of MLR by some margin, referred to herein as the maximum acceptable increase in percent MSE, δ (%), and optimize α given δ by minimizing the MSE conditional on the verifying observation exceeding a user-specified threshold (e.g., action stage for flood warning). Generally speaking, the larger the value of δ is, the larger the conditional MSE-minimizing α is. If the primary interest is with modeling and prediction of extremes in the upper tail, one may choose a large δ for improved conditional performance but at the expense of reduced unconditional performance. If conditional performance is not of concern, one may set δ to zero so that CBP-MLR with non-zero α takes effect only when the training data indicates a presence of significant CB. Pictorially (see Fig 3.2), optimizing the four parameters, δ, α, β and γ, amounts to adjusting the slope of the CBP-MLR solution and the shape of the curve that connects the two regression lines such that the resulting nonlinear regression jointly maximizes conditional and unconditional performances.

Because δ and α pertain only to the conditional performance of CBP-MLR, they may be estimated first before β and γ. Also, most predictands have much larger probability densities near median than over tails. As such, one may estimate β with γ fixed at some large value so that ϖ_i in Eq. (3.16) is effectively a rectangular pulse with a peak of unity and a base of zero. The width parameter, β, may be optimized by minimizing the unconditional MSE of CompMLR which already reflects the optimized δ and α from minimization of conditional MSE. Lastly, the smoothing parameter, γ, may be optimized by further minimizing the unconditional MSE of CompMLR given δ, α and β. In this work, the parameters are optimized by defining their feasible regions, discretizing them and carrying out a 2-dimensional search for δ and α, and two 1-dimensional searches: first for β given δ, α and a very large γ, and then for γ given δ, α and β. Admittedly, the parameter estimation procedure described above is ad hoc and may potentially be improved significantly. More objective and rigorous multi-objective parameter optimization using, e.g., Pareto optimality (Goodarzi et al., 2014; Jahan et al., 2016) is left as a future endeavor.

3.3 Evaluation

In this section, we describe the input forecasts, i.e., the predictors, used, how comparative evaluation is carried out, and the performance measures used.

3.3.1 Input forecasts used

The input forecasts considered are the MARFC single-valued forecast, the Hydrologic Ensemble Forecast System (HEFS, Demargne et al. 2014) ensemble forecast, the National Water Model (NWM, Graziano et al. 2017) medium-range single-valued forecast, and the Meteorological Model-based Ensemble Forecast System (MMEFS, Adams 2015) ensemble forecasts forced by the Global Ensemble Forecast System (GEFS, Toth and Kalnay 1997; Cui et al. 2012), the North American Ensemble Forecast System (NAEFS, Zhu, Y., Toth 2008) and the Short-Range Ensemble Forecast System (SREF, Du et al. 2004). The above streamflow forecasts are referred to as RFC, HEFS, NWM, GEFS, NAEFS and SREF for brevity. All flow forecasts are for instantaneous discharges. For ensemble streamflow forecasts (i.e., HEFS, GEFS, NAEFS and

SREF), only the ensemble mean is used. The period of record was from Jan 1, 2017, to Oct 29, 2018. The first year was moderately drier than normal, whereas the second year was significantly wetter than normal. The short period of record reflects the availability of routinely archived forecasts at MARFC and hence provides a realistic constraint on the amount of data that may be readily available for parameter estimation in an operational setting.

The pairs of the RFC and SREF, the GEFS and NAEFS, and the HEFS and NWM streamflow forecasts have maximum lead times of 4.25, 7, and 10 days, respectively. In this work, we cap the lead time at seven days so that there are at least four different forecasts available at all lead times. This provides six candidate forecasts for Days 1 through 4, and 4 candidate forecasts for Days 5 through 7. For all forecasts except NWM, the hydrologic models used are the continuous Antecedent Precipitation Index model (API-CONT, Sittner et al. 1969; Fedora and Beschta 1989) for soil moisture accounting, unit hydrograph (Chow et al., 1988) for surface runoff routing, SNOW-17 (Anderson, 2006, 1973) for snow ablation and Lag/K for routing (National Weather Service, 2006). A number of forecast points are influenced by reservoirs and flow regulations. Reservoirs and their operations are simulated via the RES-SNGL regulation models (NWS 2008ab; Adams 2016) in the Community Hydrologic Prediction System (CHPS, Roe et al. 2010), the primary operational forecasting system at the RFCs. The HEFS, GEFS, NAEFS and SREF streamflow forecasts are initialized with the same model states that MARFC maintains for their deterministic runs. As such, any differences among the above four forecasts are due solely to the differences in forcings.

The ensemble quantitative precipitation forecast (QPF) and quantitative temperature forecast (QTF) for HEFS are generated with the Meteorological Ensemble Forecast Processor (MEFP, (Schaake et al. 2007; Wu et al. 2011; NWS 2016). The MEFP uses ensemble mean QPF and QTF from GEFS using statistical parameters obtained from the GEFS reforecast dataset (Hamill et al., 2013). The NAEFS forcing forecast combines the ensemble forecasts from the Meteorological Service of Canada and the NWS (Zhu and Toth, 2008). A degree of collinearity may exist among the HEFS, GEFS and NAEFS streamflow forecasts given that they share a common forecast system, i.e., the GEFS, for their forcing forecasts and the same hydrologic models and ICs. The NWM, which uses WRF-Hydro (Gochis et al., 2018), uses the control run of the GEFS forcing forecast for its medium-range streamflow forecast. The hydrologic models used in NWM are, however, very much different from those used in the others. As such, collinearity is not likely to be an issue.

3.3.2 Evaluation strategy and performance measures used

For comparative evaluation of the MLR-based predictions, we performed 10-fold cross-validation for each forecast point. The period of record was divided into ten subperiods of equal length, each of which was used for validation while the rest were used for parameter estimation. Initially, the four parameters, δ, α, γ and β, were optimized in each training period as described above, and the resulting parameter values were used in validation. Early results indicated, however, that δ, β and γ do not vary much among different training periods or forecast points, and that $\delta = 15$ (%), $\beta = 1/2$ and $\gamma = 0.1$ generally work well. As such, only α is estimated for each training period and for each forecast point. Instead of training the procedure for each forecast point, one may consider regionalization and pool multiple forecast points into a group. Such an approach was not pursued in work so that we may examine the variations in contributions of different input forecasts across different forecast points. One may also consider seasonal stratification in parameter estimation similarly to the MEFP (Schaake et al. 2007; Wu et al. 2011; NWS 2016) or the ensemble streamflow post-processor in HEFS (Seo et al. 2006; NWS 2015). Due to the limited amount of training data, however, such stratification was not considered in this work.

Evaluation is focused on prediction of high flows for flood forecasting and for all ranges of flows for water supply and environmental flow forecasting. We use conditional and unconditional RMSE, mean error (ME), and CORR to measure the performance for the above. The conditioning is on verifying observed flow exceeding the 95th percentile flow for each forecast point. The 95th percentile is not in general associated with flooding conditions. The use of a higher threshold, however, would reduce the sample size and, more importantly, predominantly sample under-forecasts due to the limited predictive skill for large flows, particularly at longer lead times. Such biased sampling would favor for conditional performance in a forecast system that consistently over-forecasts and hence has a large high overall bias. The use of the 95th percentile threshold is to avoid the above situation by including both over- and under-forecasts in the diffuse upper tail for balanced sampling. For performance evaluation over all ranges of flow, we use unconditional RMSE, ME and CORR.

3.4 Results

In this section, we first assess the comparative predictive skill of all available forecasts for the selection of the predictors. We then present the evaluation results.

3.4.1 Assessment of skill in input forecasts

If a subset of the input forecasts is substantially more skillful than the rest, one may not expect merging all input forecasts to outperform consistently the best input forecast (Konstantine P Georgakakos et al., 2004). As such, we first assess the predictive skill of all available input forecasts and their spatio-temporal variations to screen the predictors. Fig 3.4 shows the unconditional CORR for headwater forecast points in the North Branch of the Susquehanna and James forecast groups. All unconditional performance or skill measures are based on including all ranges of verifying observed flow. The first represents the northernmost forecast group and the second represents one of the southernmost forecast groups in the MARFC's service area (see Fig 3.3). We also examined similar plots for correlation conditional on the verifying observed flow exceeding the 95th percentile, referred to as conditional CORR, and for downstream forecast points for the 13 forecast groups. Collectively, Fig 3.4 and similar figures for all other forecast groups may be summarized as follows. In general, the NWM streamflow forecast is consistently less skilful than all other forecasts for most of the forecast horizon. The above suggests that uniformly including NWM is not likely to benefit multi-model merging. The forecasts are significantly more skilful for the northern forecast groups than the southern and for the cool season than the warm, in reflection of the relative skill in precipitation forecast (Brown et al., 2014a, 2014b; Siddique et al., 2015). For the southernmost forecast groups, the skill is sharply lower due to smaller skill in QPF and larger hydrologic uncertainty. Contributing factors to the latter include less accurate calibrations for some locations due to sparser rain gauge networks and storages that are not well accounted for. In addition, the southern forecast groups exhibit varying degrees of collinearity (i.e., very high correlation among different input forecasts) among the HEFS, GEFS, NAEFS and SREF streamflow forecasts in very dry conditions during inter-storm periods. The above is attributable to the fact that the 4 forecasts share the same hydrologic models, ICs and mean areal potential evapotranspiration. The above observations suggest that merging all input forecasts (i.e., without NWM) some of which are likely to be collinear is not likely to benefit the southern forecast groups. Based on the above assessment, we use all input forecasts except NWM for all headwater forecast points in the Chemung, Delaware, North Branch of the Susquehanna and Passaic forecast groups, and for all downstream forecast points in the Lower Main Stem of the Susquehanna Forecast Group in addition to the above 4. There are 10, 31, 11, 14 and 12 forecast points in the Chemung, Delaware, Lower Main Stem of the Susquehanna, North Branch of the Susquehanna and Passaic forecast groups, respectively. All results presented below are hence based on RFC, HEFS, GEFS, NAEFS and SREF for lead times of 0.25 to 4.25 days, and HEFS, GEFS and NAEFS for lead times of 4.5 to 7 days.

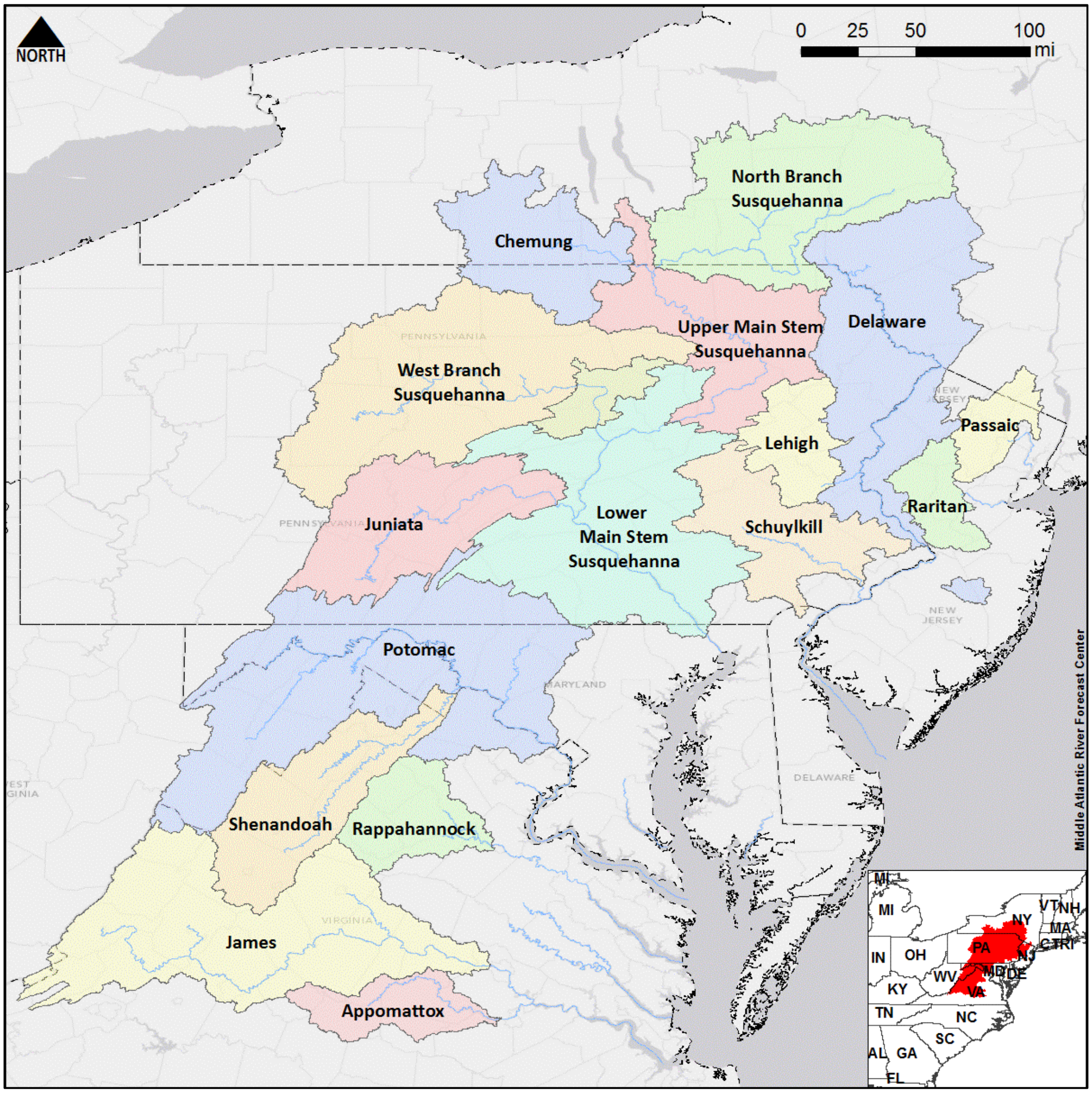


Figure 3.3: Map of forecast groups in the MARFC's service area.

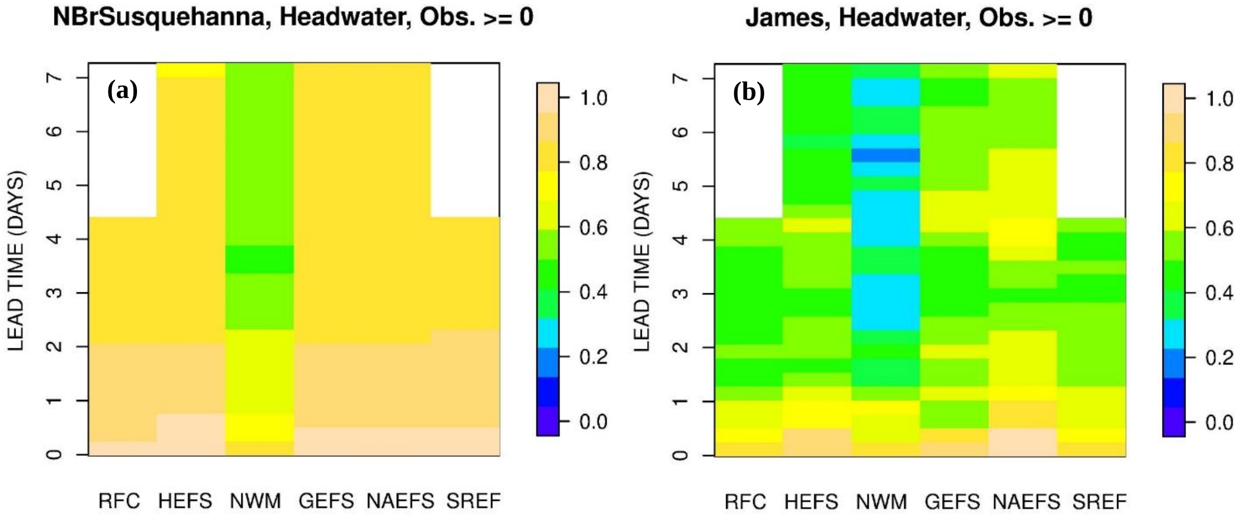


Figure 3.4: Unconditional correlation of forecasts with verifying observed flow vs. lead time for headwater forecast points in the (a) North Branch of the Susquehanna and (b) James Forecast Groups.

3.4.2 Comparative evaluation results

Figs 3.5ab show the conditional RMSE of all predictions for all headwater and downstream forecast points in the evaluation domain, respectively. The sample size ranges from 453 to 517 for headwater forecast points and 751 to 857 for downstream points across all lead times. As expected, the downstream points are much more predictable for which the conditional RMSE plateaus at about 4 days into the future. The headwater results show peak RMSE values recurring every 24 hrs due to the diurnal cycle in the skill of precipitation forecast (Brown et al., 2014a, 2014b; Siddique et al., 2015). Fig 3.5a shows that CBP-MLR is superior to the best input forecast in the MSE sense for the headwater points at all lead times, and that CompMLR is slightly inferior to CBP-MLR as one may expect from Eqs. (3.15) and (3.16). MLR, on the other hand, is increasingly inferior to CBP-MLR as the lead time increases, a reflection of increasing CB for less predictable headwater points as the lead time increases. Fig 3.5b shows that CBP- and CompMLR are superior to the best input forecast in the MSE sense for the downstream points except at the longest lead times. For downstream points, the MLR prediction stays superior to the best input forecast up to about 4 days into the future owing to the larger predictability. Fig 3.5b indicates that, among all input forecasts, NAEFS has the best overall conditional performance though HEFS is somewhat better than NAEFS for lead times of about 2 to 3.5 days for downstream points, and that CBP- and CompMLR outperform NAEFS for both headwater and downstream points except at the longest forecast lead times where the predictive skill in input forecasts is too small for merging to be effective.

Figs 3.5cd are the same as Figs 3.5ab, respectively, but for unconditional RMSE. The sample size ranges from 8,750 to 9,750 for headwater points and 14,150 to 16,000 for downstream points across all lead times. With all ranges of flow included, the skillful lead time is increased to about 3 and 5 days or longer for headwater and downstream points, respectively. Figs 3.5cd show that the relative unconditional performance among the input forecasts is very much different from the relative conditional performance seen in Figs 3.5ab. With all ranges of observed flow considered, HEFS is by far the best performing input forecast in the MSE sense, and the MMEFS forecasts, i.e., GEFS, NAEFS and SREF, have significantly larger unconditional RMSE particularly for downstream points. Among the MLR-based predictions, MLR provides the best unconditional performance as expected, whereas CBP-MLR is inferior. CompMLR out- and underperforms CBP-MLR and MLR, respectively, as may be expected from Eq. (3.16), and generally outperforms the best-performing input forecast. Collectively, Fig 3.5 illustrates the challenge in multi-model streamflow prediction of improving over the best input forecast both conditionally and unconditionally, and for both less predictable headwater forecast points and more predictable downstream points. Fig 3.5 shows that only CompMLR is generally able to meet this objective.

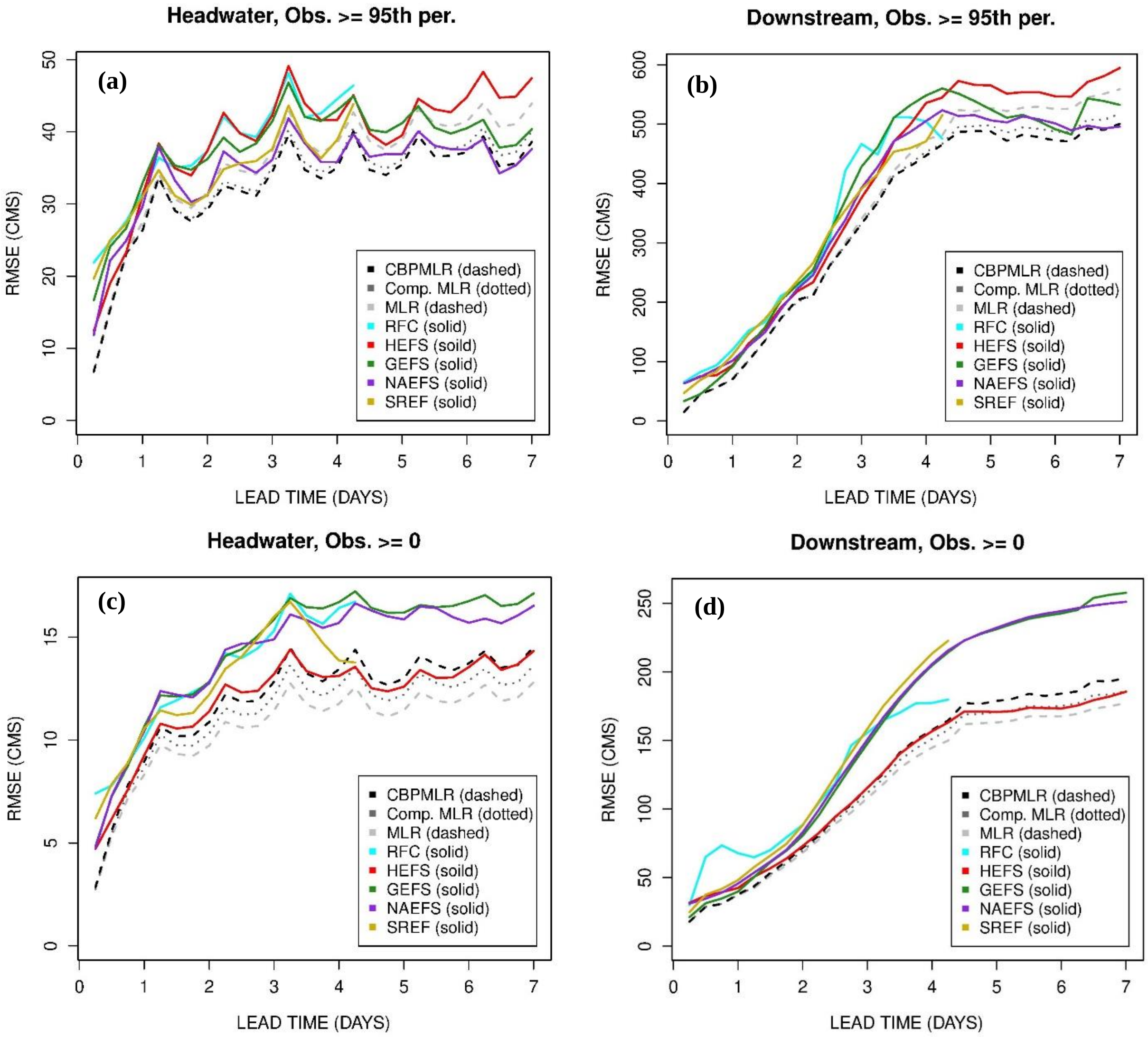


Figure 3.5: Conditional RMSE vs. lead time of CBP-MLR, CompMLR, MLR, HEFS, GEFS, NAEFS, and SREF for the verifying observed flow exceeding the 95th percentile for (a) headwater, and (b) downstream forecast points, (c) same as (a) but unconditional RMSE, and (d) same as (b) but unconditional RMSE.

Figs 3.6ab show the conditional and unconditional ME, respectively, for downstream forecast points. Those for headwater points are qualitatively similar. Figs 3.6a shows that the MMEFS forecasts have significantly smaller conditional ME than all other input forecasts and all MLR-based predictions. HEFS, on the other hand, has very large negative CB (i.e., consistently under-forecasts large flows) for both headwater and downstream points. Figs 3.6b shows that MLR and HEFS are unconditionally unbiased, but that the MMEFS forecasts are unconditionally biased high (i.e., over-forecast in the mean) and the high unconditional bias increases as the lead time increases. Fig 3.6 serves to illustrate the nature of the multi-model streamflow prediction challenge in that the smaller bias in the MMEFS forecasts for high flows was achieved at the expense of large high bias in flows of all magnitudes, and the unconditional unbiasedness in HEFS was achieved at the expense of large low bias in high flows. It is interesting to observe in Fig 3.6 that CBP- and CompMLR place their conditional and unconditional MEs between the respective two bounds associated with the HEFS and MMEFS forecasts. The above result is a reflection of the balancing act of reducing biases in mean and standard deviation, and increasing the strength of linear association under the MSE minimization (Murphy and Winkler, 1987; Nelson et al., 2010).

To help visualize the key aspects summarized in Figs 3.5 and 3.6, Fig 3.7 shows an example set of scatter plots of forecast vs. verifying observed flow for all ranges of flow for all downstream points in the Chemung forecast group for lead time of 6.75 days. Recall that, at this lead time, only HEFS, GEFS and NAEFS are available as input forecasts. In each panel, the solid red line shows the quantile-quantile (QQ) plot. If the empirical distribution or, equivalently, all statistical moments of the prediction match those of the observed, the QQ plot would lie on the diagonal line. Three observations are noteworthy. The GEFS and NAEFS streamflow forecasts are unconditionally biased high (Figs 3.7ef) which contributes to superior conditional performance among the input forecasts. The HEFS streamflow forecast is conditionally severely biased (Fig 3.7d) even though it is unconditionally unbiased. MLR is unconditionally unbiased but is significantly conditionally biased (Fig 3.7a) which CBP-MLR addresses (Fig 3.7b). CompMLR appears very similar to CBP-MLR due to the fact that differences are predominantly near median (Fig 3.7c). Figs 3.8ab show the unconditional CORR for headwater and downstream points, respectively. The conditional CORR results are similar. It is readily seen in Fig 3.8 that predictive skill diminishes much more quickly for headwater forecast points than for downstream points. Fig 3.8 indicates that, albeit modestly, multi-model merging generally improves both conditional and unconditional CORR over all input forecasts for both headwater and downstream points, and this improved CORR contributes to the reduction in RMSE seen in Fig 3.5.

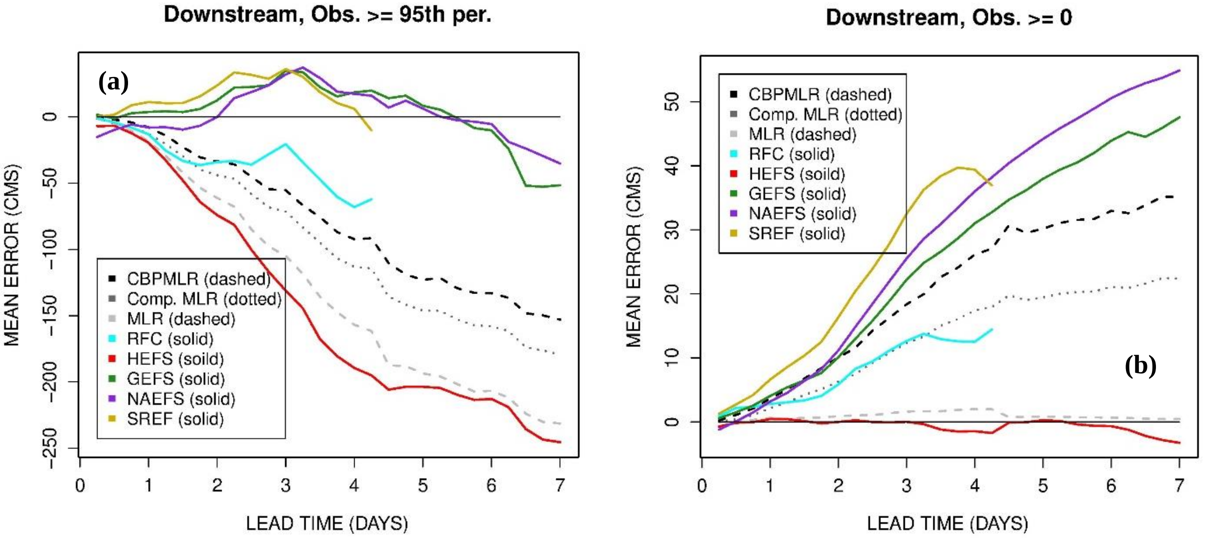


Figure 3.6:(a) Conditional and (b) unconditional ME vs. lead time of CBP-MLR, CompMLR, MLR, HEFS, GEFS, NAEFS, and SREF for downstream forecast points.

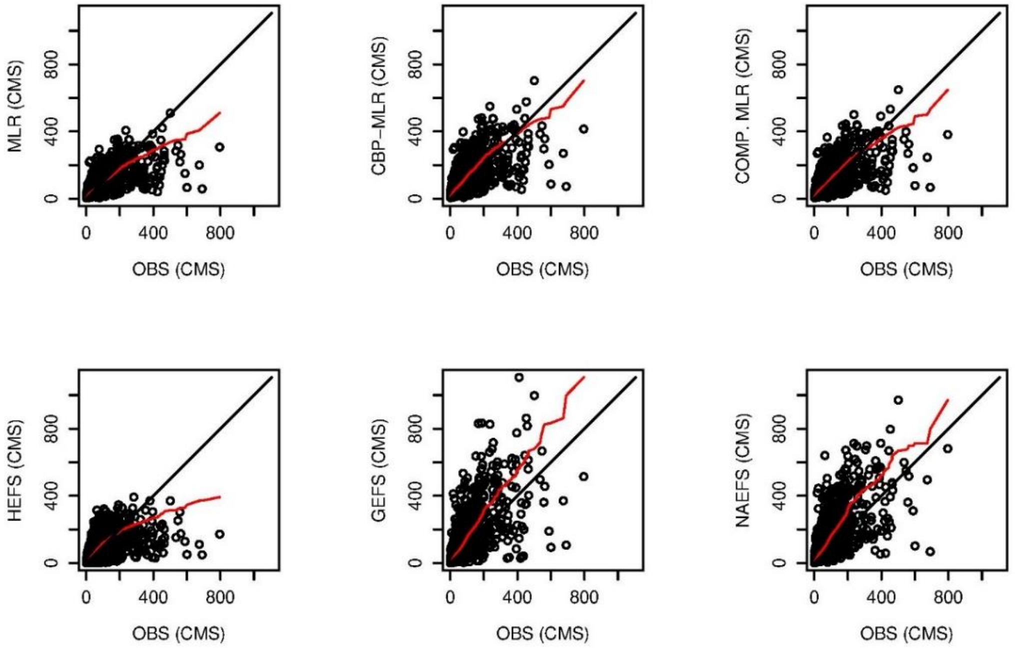


Figure 3.7:Scatter plots of forecast vs. verifying observed flow for all ranges of flow for all downstream forecast points in the Chemung Forecast Group for lead time of 6.75 days.

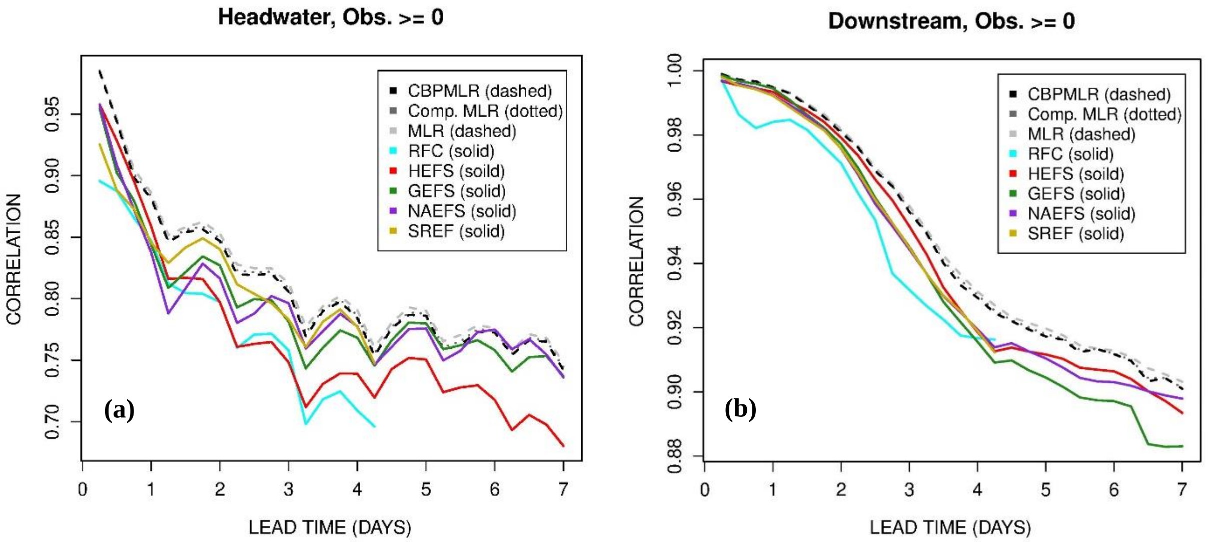


Figure 3.8:Unconditional correlation of streamflow forecasts vs. lead time for (a) headwater and (b) downstream forecast points.

The very contrasting conditional vs. unconditional performance of the HEFS and MMEFS forecasts seen above indicates that the MMEFS forecasts may be better-suited for flood forecasting than water supply or environmental flow forecasting whereas the opposite may be true for HEFS. The HEFS streamflow forecast performs very well unconditionally because MEFP is designed to generate ensemble forecasts that are unconditionally unbiased (Schaake et al. 2007; Wu et al. 2011; NWS 2016). The current version of MEFP, however, is CB-unaware and hence the MEFP forecasts are significantly conditionally biased resulting in rather poor conditional performance as shown in the literature (Ghazvinian et al., 2020; Zhang et al., 2017). The distinguishing value of CB-aware MLR is that it is generally superior to the best input forecast both in the conditional sense and in the unconditional sense under wide-ranging conditions of predictability and predictive skill.

Figs 3.9a and 3.9b show the box-and-whisker plots of the MLR and CBP-MLR weights for headwater and downstream forecast points for lead time of 2 and 4 days, respectively. It is readily seen that the CBP-MLR weights have larger ranges of variation than the MLR weights. This

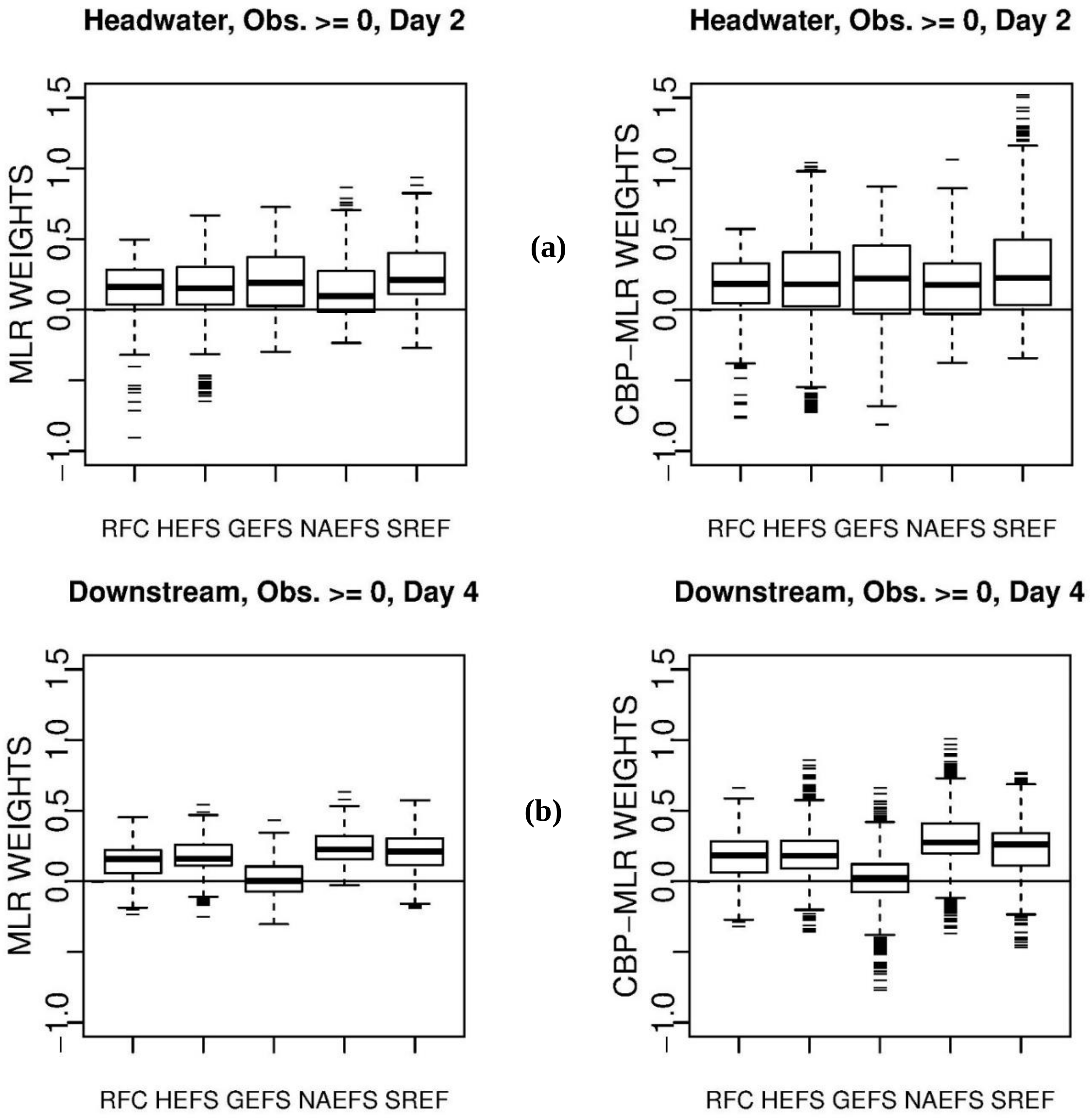


Figure 3.9:Box-and-whisker plots of the MLR and CBP-MLR weights for (a) headwater forecast points for lead time of 2 and (b) downstream forecast points for lead time of 4 days

‘stretching’ of the boxes and whiskers by CBP-MLR is often quite large even at short lead times, an indication that MLR predictions are subject to significant CB. For a number of forecast points, the weights are significantly negative, an indication that there might exist multi-collinearity among certain input forecasts. To examine this possibility, we performed additional experiments in which HEFS and GEFS were excluded one at a time as well as together while NAEFS was kept in all cases as the most skillful forecast for headwater points. Similarly, we excluded GEFS and NAEFS one at a time as well as together while HEFS was kept in all cases as the most skillful for downstream points. It was found that the above exclusions do not significantly change the weight patterns of the remaining input forecasts, an indication that multi-collinearity is not a significant issue. Fig 3.9 indicates that large variations in the weights across different forecast points makes selecting a single “best” forecast for the day or for the event at hand may be impractical in real-time operation.

Figs 3.10ab show the mean of the CBP-MLR weights for each input forecast and for each lead time for all headwater and downstream forecast points, respectively. Those of the MLR weights are very similar as may be inferred from Fig 3.9. It is seen that all forecasts included in merging contribute significantly. Fig 3.10 indicates that NAEFS contributes the most to CBP-MLR for headwater forecast points whereas HEFS contributes the most for downstream points in agreement with Fig 3.5. GEFS shows the largest variations across forecast lead time, for attribution of which additional investigation is needed. We also examined similar plots for cool vs. warm seasons. They indicate that the contributions of different input forecasts are most evenly spread for headwater

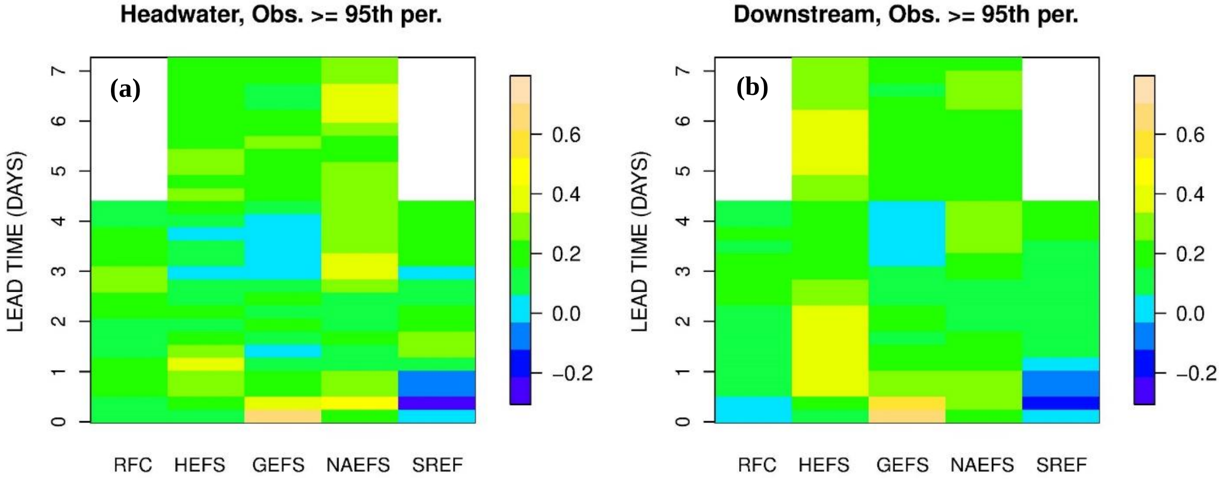


Figure 3.10: Mean of the CBP-MLR weights for each input forecast and for each lead time, conditional on verifying observed flow exceeding the 95th percentile, for all (a) headwater, and (b) downstream forecast points.

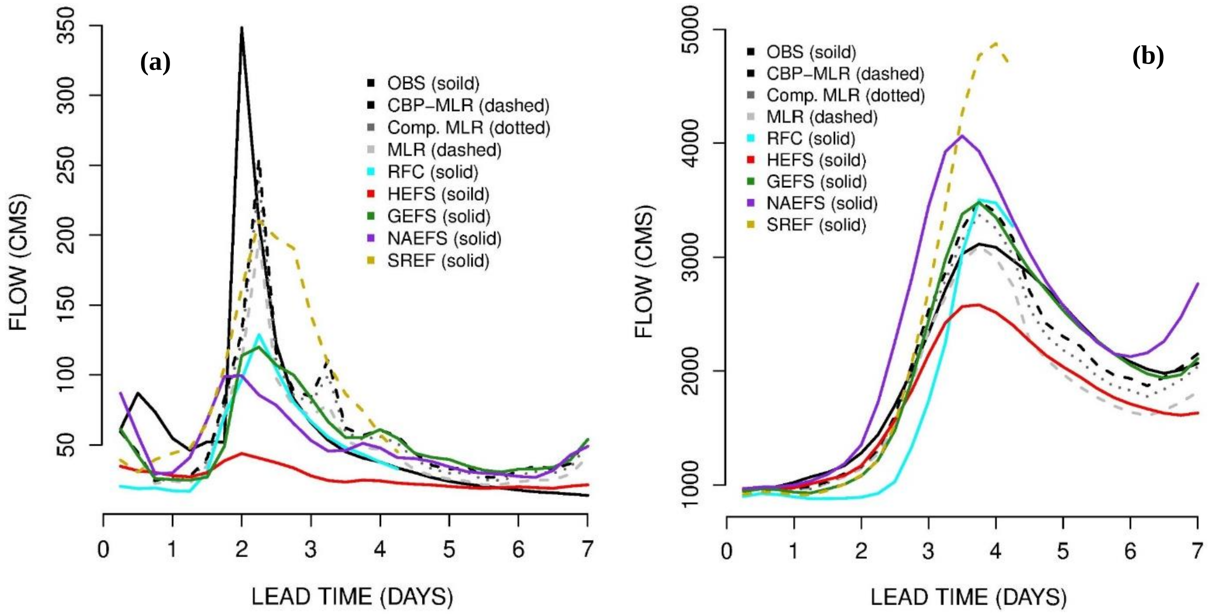


Figure 3.11: Example hydrographs of CBP-MLR, CompMLR, MLR, HEFS, GEFS, NAEFS, SREF and verifying observed flow for large events for (a) headwater and (b) downstream forecast points.

forecast points in the cool season. This is due to the more uniform predictive skill among different forcing forecasts in the cool season, and the fact that biases in the forcing forecasts do not grow as large for headwater points, in terms of streamflow volume, as for downstream points due to much smaller contributing areas.

Figs 3.11ab show examples of the hydrograph representation of MLR-based predictions for large events for headwater and downstream forecast points, respectively. For comparison, the verifying observed hydrographs and all input forecasts are also shown. These examples are chosen to illustrate qualitatively the key characteristics of different forecasts and predictions rather than to indicate representative performance. Fig 3.11a shows that, albeit still under-forecasts, the CBP- and CompMLR rise above MLR, and all input forecasts to provide significantly more realistic peak flow predictions. Without being CB-aware, such predictions are not possible. Fig 3.11a also illustrates the detrimental impact of CB in HEFS as it is barely able to indicate a rise in flow. Whereas Fig 3.11a provides an example of CBP- and CompMLR acting as extrapolators, Fig 3.11b provides one as interpolators. In Fig 3.11b, owing to the larger predictability and predictive skill at this downstream location, the hydrograph representations of the MLR-based predictions are very realistic. Also seen in Fig 3.11b is the high bias in NAEFS and SREF amplified through integration over a large contribution area.

A potential criticism of the proposed approach is that MLR-based predictions are forecast lead time-specific and do not consider temporal dependence, instead relying solely on the model dynamics of the input forecasts for temporal consistency. When presented as hydrographs for practical applications, MLR-based predictions may hence lack realism. Indeed, the secondary peak associated with the MLR-based predictions in Fig 3.11a at lead time of 3.25 days is, in all likelihood, not real but an artefact of the lead time-specific weights lacking temporal consistency due to sampling uncertainty. In theory, it is possible to include input forecasts valid at multiple lead times to account for temporal dependence albeit at the expense of significantly larger dimensionality of the estimation problem. Experience with regression-based statistical post-processing for streamflow prediction (Seo et al. 2006; Regonda et al. 2013; NWS 2015; Alizadeh et al. 2020) suggests that adding temporally-dependent predictors provides only marginal improvement, and is more likely to produce unstable solutions due to under-determinedness inherent in large-dimensional regression and multi-collinearity. Further research is needed, however, to assess potential gains in the context of CBP- and CompMLR.

3.5 Discussion

The premise of multi-model streamflow forecasting is that one may optimally combine multiple streamflow forecasts from different sources to produce a new forecast which is more accurate in the MSE sense than the best-performing individual forecasts at all times at all locations. The above premise is realized in practice, but in the unconditional MSE sense only, via the Gauss-Markov theorem (Deutsch 1965) on which MLR stands. Due to regression dilution, however, MLR-based multi-model forecast is conditionally biased, resulting in under- and overprediction of large and small flows, respectively. For this reason, a multi-model forecast for large flows based on MLR is inferior to the best-performing individual forecast even though the former is superior to the latter when all ranges of flow are considered. For example, in Figs 3.5a and 3.5b in the revised version, MLR-based multi-model forecast is inferior to NAEFS or SREF beyond Day 2 for headwater locations, and to NAEFS or GEFS beyond Day 4.5 for downstream locations, respectively. Accurate prediction of floods, in particular, extreme floods, is arguably the most important service in operational streamflow forecasting as they greatly impact society. As illustrated in Fig 3.2 (see also Seo 2013, Jozaghi et al. 2019, Lee et al. 2019, Shen et al. 2019), MLR is adversely impacted by CB with increasing severity as the predictand increases in magnitude. Accordingly, the larger, and hence more impactful, the flood flow is, the less effective MLR-based multi-model prediction becomes relative to the best-performing individual forecasts.

Given the above picture, one might instead consider compositing the best-performing individual forecasts according to the flow conditions, location and lead time. Such an approach, however, is impractically difficult because the best forecast varies with the magnitude and predictability of the (unknown) flow being predicted. In addition, such compositing is bound to produce very frequently forecast hydrographs that are physically unrealistic over the forecast horizon and across the range of variations in flow magnitude. Relative to MLR, CompMLR reduces RMSE by about 10 and 4 percent or more beyond lead times of 2 and 4.5 days for flows above the 95th percentile for headwater and downstream locations, respectively (see Figs 3.5ab). Whereas these reductions may seem only modest, they critically render CompMLR the best performing forecast for both headwater and downstream locations, for both high flows and all ranges of flows, and across all lead times except for the longest lead times where the collective predictive skill of individual forecasts is too small for multi-model prediction to benefit from. The singular advantage of the proposed method over MLR hence is that it provides the best performing forecast in the MSE sense over the entire effective forecast horizon regardless of the flow conditions and locations with only a modest increase in complexity and without additional data requirements. Needless to say, such performance characteristics are a prerequisite for machine learning-based automatic multi-model streamflow forecasting.

Due to the additional parameters involved, parameter estimation or training of CompMLR is more involved than that of MLR. As explained in Section 3, however, three of the 4 additional parameters vary little from one training data set to another and only α requires active optimization. Pictorially, the above means that the general shape of the curve in Fig 3.2 connecting the two regression lines does not change much from one training data set to another, but that the slope of the 2nd regression does. If the dimensionality of the predictand is high, solving for the 2nd slope can be computationally significantly more expensive than solving for the 1st slope only (i.e., as in MLR) due to the additional matrix inversion involved (see Eq.(3.10b)). If the predictand is 1-dimensional as in this work, however, the additional inversion, i.e., $(Y^T Y)^{-1}$ where Y is the predictand, is not computationally significant. Note also that any additional computation for the proposed method relative to MLR is limited only to training. For real-time operation, the additional computation for CompMLR is negligible.

Finally, as alluded to in the Introduction Section, another possible point of comparison for the proposed method is BMA (Madigan and Raftery 1994; Hoeting et al. 1999; Raftery et al. 2005, 2015; Duan et al. 2007, Jozaghi et al. 2020). If the formulation is based on the standard Bayesian normal conjugate linear model (Raiffa and Schlaifer 1961), BMA amounts to Bayesian MLR weight-averaged via the total probability law according to the goodness of different model combinations. As such, BMA is subject to CB just as MLR is. It is noted here that we are also evaluating BMA for multi-model streamflow prediction. To improve performance for large flow, however, we have implemented CBP-MLR in the BMA package, BMS (Zeugner 2011). We are currently carrying out a comparative evaluation of BMA based on CBP-MLR (Jozaghi et al. 2021a) and the results will be reported in the near future.

3.6 Conclusions and future research recommendations

For modeling and prediction of extremes such as floods using multiple linear regression (MLR), attenuation bias is a very serious issue as it results in systematic under- and over-prediction in the upper and lower tails of the predictand, respectively. In this work, we introduce conditional bias-penalized multiple linear regression (CBP-MLR) which reduces attenuation bias, or conditional bias (CB), by jointly minimizing mean squared error (MSE) and Type-II error squared. Whereas CBP-MLR improves prediction over tails, it degrades performance near the median. To effect MLR-like performance near median while performing closely to CBP-MLR over tails, we employ CompMLR, which linearly weight-averages the MLR and CBP-MLR estimates. We then apply MLR, CBP- and CompMLR to multi-model streamflow prediction for multiple forecast groups in the US National Weather Service (NWS) Middle Atlantic River Forecast Center's (MARFC) service area. The input forecasts considered are six different streamflow forecasts produced by the NWS MARFC and the National Water Model. The maximum lead time considered is seven days. The main conclusions are as follows.

The relative performance among different input forecasts varies most significantly with the range of the verifying observed flow. Generally speaking, the streamflow forecasts forced by the North American Ensemble Forecast System, or NAEFS, and the Short-Range Ensemble Forecast, or SREF, are better-suited for flood forecasting whereas those generated with the Hydrologic Ensemble Forecast Service, or HEFS, are better-suited for water supply and environmental flow forecasting. It is shown that CompMLR generally outperforms the best individual forecasts in the MSE sense for observed flow exceeding the 95th percentile, for all ranges of observed flow, for less predictable headwater forecast points, and for more predictable downstream points for lead times of up to about six days. If forecasting high flows is of little importance, CB is not likely to be an issue and MLR may suffice. Large variations are seen in the MLR and CBP-MLR weights associated with different input forecasts across the 7-day forecast horizon and across different forecast points. The above suggests that selecting a single “best” forecast system for the forecast point of interest or a single “best” forecast for the day or event at hand is likely a difficult proposition. The distinguishing value of CB-aware MLR is that it produces predictions that are generally superior to the best input forecast both in the conditional sense and in the unconditional sense under wide-ranging conditions of predictability and predictive skill. In this work, objective screening of non-contributing forecasts via systematic variable selection was not considered. We are currently implementing CBP-MLR in Bayesian Model Averaging (BMA) using the standard Bayesian normal conjugate linear model (Jozaghi et al. 2020, 2021b). Possible gains from variable selection will be evaluated in the context of BMA through model combinations, and the results will be reported in the near future. Finally, the results presented in this work are based on a period of record of less than two years. Systematic assessment of data requirements vs. performance, particularly for forecasting extreme floods, is necessary along with potential benefits of regionalization for space-for-time trade-off. With limited data available for each station and forecast horizon, unified machine learning techniques, in particular ANN’s, similar to the one proposed by Ghazvinian et al. (2021) can be possibly used. Such a network not only can digest and project spatio-temporal binary predictors (e.g., forecast lead time and station ID’s, embeddings) to a large vector of inputs but also performs well in characterizing predictor interactions translating to superior predictive skills.

Chapter 4: CBP-BMA: Theory and Application to Multimodel Streamflow Prediction

4.1 Introduction

Accurate streamflow prediction over a range of lead times is essential to various water and emergency management activities such as flood warning and control (Danandeh Mehr et al., 2015; Liu et al., 2018; Luo et al., 2019), water supply, irrigation (Jozaghi et al., 2018; Jozaghi and Shamsai, 2017), navigation, recreation, water quality and sediment transport (Adnan et al., 2019). The accuracy of hydrological forecasts is affected by various sources of uncertainty such as forcing inputs, and model parameters, initial conditions, and structures (Chen et al., 2008; Tiwari and Adamowski, 2013; Xu et al., 2019). To assess the uncertainty in hydrological forecasting, various techniques have been developed over the years (Kasiviswanathan and Sudheer, 2017; Liu et al., 2017). Among them, the ensemble streamflow prediction approaches are the most popular (Madadgar and Moradkhani, 2014; Michaels, 2015; Seo, 2006). Whereas those techniques using multiple inputs and parameter sets can capture significant portions of the predictive uncertainty, they cannot address the uncertainty within an individual hydrologic model structure (Konstantine P. Georgakakos et al., 2004; Vrugt and Robinson, 2007). Recently, multi-model techniques, which merge forecasts from multiple sources, have been used to obtain more skillful forecasts by addressing the model structural uncertainty.

The increasing availability of streamflow forecasts from multiple sources (Muhammad et al., 2018) provides an opportunity to improve forecast accuracy by optimally merging multiple forecasts (Beven, 2001; Doblas-Reyes et al., 2000; Konstantine P. Georgakakos et al., 2004; Madigan and Raftery, 1994; Regonda et al., 2006). Toward that end, a number of techniques for multi-model streamflow prediction have been developed (Ajami et al., 2006; Doblas-Reyes et al., 2005; Duan et al., 2007a; Rajagopalan et al., 2002; Yun et al., 2005). Among them, statistical techniques and regression analysis are the most popular (Policelli et al., 2019; Rahman et al., 2020; Sahoo et al., 2009; Shih and Shih, 1979; Supriya et al., 2015; Um et al., 2011) of which the simplest are multiple linear regression (MLR), least absolute shrinkage and selection operator (LASSO; Tibshirani, 1997, 1996), ridge regression (Hoerl and Kennard, 1970), simple averaging (DelSole and Tippett, 2007), Granger–Ramanathan averaging (Granger and Ramanathan, 1984). These statistical methods, however, do not consider structural uncertainty in the statistical models used in that they typically choose a model which is then to the data. Such approaches disregard the uncertainty in model selection, resulting in overconfident inferences (Jennifer A. Hoeting et al., 1999; Madigan and Raftery, 1994; Raftery et al., 2005). To address the above, more complex approaches have been developed such as Artificial Neural Network (ANN; Ghazvinian et al., 2021; Hassanzadeh et al., 2020; Lashkarara et al., 2021) and Bayesian Model Averaging (BMA; Hoeting et al., 1999; Raftery, 1995, 1993; Raftery et al., 2005, 1997) which produce more reliable probabilistic forecasts by quantifying predictive uncertainty better. BMA is an extension of the Bayesian perspective approaches to the model selection problems in which one models uncertainty in model posteriors and posterior parameters via Bayes' theorem in addition to modeling parametric uncertainty through the prior distribution. Numerous efforts have been made to demonstrate the ability of the BMA approach in generating more skillful predictions (Arsenault et al., 2015; Baharvand et al., 2020; Cees G.H. Diks and Vrugt, 2010; Genell et al., 2010; Jozaghi et al., 2021b, 2020; Liu et al., 2016; Viallefont et al., 2001; Wang et al., 2004). Most BMA implementations employ the standard Bayesian normal-conjugate linear model which uses MLR as the base model. Because MLR is based on least squares minimization, its solution is subject to low and high biases, or conditional bias (CB), over the upper and lower tails of the predictands, respectively (Fuller, 1987; Hausman, 2001; Seber and Wild, 1989) with detrimental effects for prediction of extremes, as elaborated below.

It is well known in statistics and econometrics that error variance (i.e., least squares) minimization is subject to potentially significant negative and positive biases over the upper and lower tails of the predictand, respectively, in the presence of observation error in the predictors (Fuller, 1987; Hausman, 2001; Seber and Wild, 1989). This effect in regression is referred to in the statistical literature as regression dilution, which causes conditional bias in the regression coefficients (Frost and Thompson, 2000; Hughes, 1993). In multi-model streamflow prediction using regression techniques, the regressors are the streamflow forecasts from different sources. As such, forecast errors, which are very often very significant, act as observation errors in multi-model streamflow prediction. Only a few reports may be found in the statistical literature that attempt to address CB (Jozaghi et al., 2019; Kim et al., 2018; Lee et al., 2019; Seo et al., 2014; Shen et al., 2019). Ciach et al. (2000) described the CB problem as the error-in-variable or attenuation effects. Seo (2013) proposed a linear estimation technique that minimizes the weighted sum of error variance and expectation of the Type-II error squared to address the harmful effects of conditional bias (CB) on estimating extremes.

This study develops and evaluates Conditional Bias-Penalized Bayesian Model Averaging (CBP-BMA) which replaces MLR in BMA with CBP-MLR for improved performance for large flows. CBP-BMA is then applied to multi-model streamflow prediction for 13 forecast groups in the NWS Middle Atlantic River Forecast Center’s (MARFC) service area. The new contributions of this work are the development of CBP-BMA, evaluation of predictive skill of individual forecast, comparative assessment of CBP-BMA, assessment of the performance of BMA and CBP-BMA in relation to hydroclimatology and hydrologic conditions, assessment of the impact of using non-diffuse priors, and advances in understanding of the multi-objective nature of multi-model streamflow prediction. This chapter is organized is as follows. Subsection 2 describes the study area and data used. Subsection 3 describes the methods used. Subsection 4 describes the evaluation metrics used. Subsection 5 presents the results and Subsection 6 provides the conclusions and future research recommendations.

4.2 Study area and data used

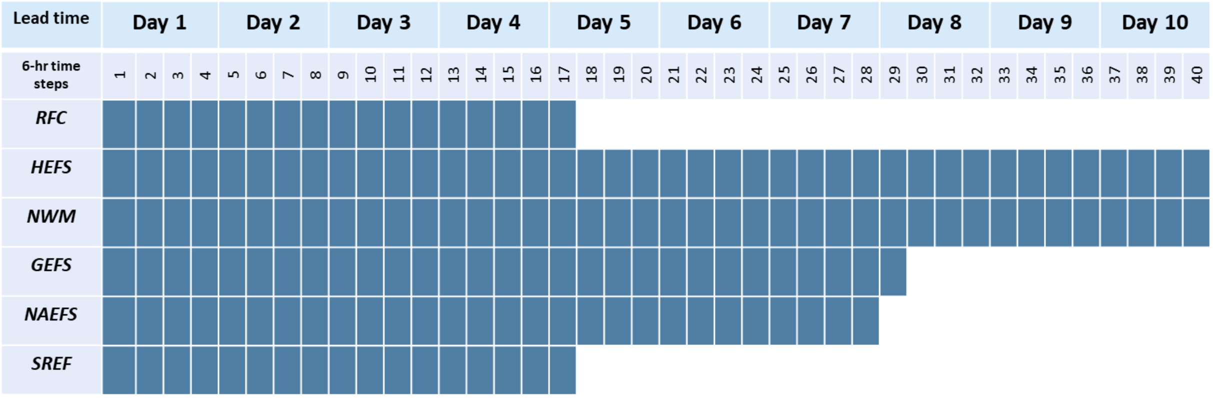
The input forecasts used in the model are from the MARFC. The MARFC service area includes the Chemung, North Branch Susquehanna, Delaware, Upper Main Stem Susquehanna, West Branch Susquehanna, Passaic, Raritan, Schuylkill, Lower Main Stem Susquehanna, Juniata, Potomac, Rappahannock, James, Lehigh, and Appomattox Forecast Groups (see Fig 3.2). The last two groups are not considered in this study due to their small size. The MARFC produces multiple hydrologic forecasts. In addition to flood forecasts which are the most well-known product, MARFC provides river and water information used for reservoir operations, recreation, navigation, and water supply plans. In this study, we consider six streamflow forecasts: the MARFC single-valued forecast, the National Water Model (NWM) medium-range single-valued forecast (Graziano et al., 2017), the Hydrologic Ensemble Forecast System (HEFS) ensemble forecast (Demargne et al., 2014), and the Meteorological Model-based Ensemble Forecast System (MMEFS) ensemble forecasts (Adams, 2015). The MMEFS ingests the numerical weather prediction ensemble grids from the NOAA/NWS National Centers for Environmental Prediction, including the Global Ensemble Forecast System (GEFS; Cui et al., 2012; Toth and Kalnay, 1997), North American Ensemble Forecast System (NAEFS; Zhu and Toth, 2008), and Short Range

Ensemble Forecast System (SREF; Du et al., 2004). The resulting streamflow forecasts are referred to herein as RFC, NWM, HEFS, GEFS, NAEFS, and SREF, respectively. In this study, we use the ensemble mean for ensemble forecasts HEFS (48 ensemble members), GEFS (21 ensemble members), NAEFS (42 ensemble members), and SREF (26 ensemble members). All streamflow forecasts are instantaneous discharges. The analysis period is from Jan 1, 2017, to Oct 29, 2018. Although the period of record is short, it reflects the availability of operational forecasts at MARFC. It hence provides a realistic constraint on the amount of data that can be used for forecasting. Two forecasts RFC and SREF, have maximum lead times of 102 hours, GEFS and NAEFS 168 hours, and HEFS and NWM 240 hours (see Fig 4.2 forecast horizon). We evaluate the proposed model for lead times up to 168 hours; therefore, we will have all six forecasts up to 102 hours lead times and four forecasts NAEFS, GEFS, HEFS, and NWM for all lead times.

The total uncertainty in streamflow forecast in HEFS is considered as a combination of the meteorological and hydrologic uncertainties (Seo et al., 2010). The Meteorological Ensemble Forecast Processor (MEFP) is used to model forcing uncertainties. The MEFP uses ensemble mean quantitative precipitation forecast (QPF) and quantative temperature forecast (QTF) from GEFS to produce QPF and QTF conditional on raw ensemble mean forecast (Schaake et al., 2007; Wu et al., 2011). The total uncertainty in streamflow forecast is addressed in two steps. The forcing ensemble forecasts are first generated with MEFP which are then input to the hydrologic models. The resulting raw streamflow ensemble forecasts may be post-processed with the Ensemble Postprocessor (EnsPost; Seo, 2006) to reduce hydrologic uncertainty. For more information on HEFS, the reader is referred to Brown et al. (2014b, 2014a). To produce more skillful forecast, the NAEFS blends the GEFS of the NWS and the Canadian Meteorological Centre Ensemble (CMCE) of the Meteorological Service of Canada (MSC; Candille, 2009). The NAEFS combines 21 ensemble members from the Canadian Meteorological Centre operational Global Environmental Multiscale (GEM) model and 21 ensemble members from the NOAA/NWS GEFS. Except for NWM, all input forecasts to multimodel streamflow prediction utilize the Continuous Antecedent Precipitation Index model (API-CONT) for soil moisture accounting (Sittner et al., 1969), unit hydrograph for surface runoff routing (Chow et al., 1988), SNOW-17 for snow ablation (Anderson, 2006) and Lag/K for channel routing (National Weather Service, 2006). The initial conditions used are the same for HEFS, GEFS, NAEFS, and SREF streamflow forecasts. Consequently, the differences in these forecasts are due to the differences in forcings. Due to the same hydrologic models used, the above streamflow forecasts may contain some degree of collinearity (Jozaghi et al., 2021a). The NWM, on the other hand, uses the control run of the GEFS forcing forecast through WRF-Hydro for its medium-range streamflow forecast (Gochis et al., 2018). As such, collinearity is not an issue.

4.3 Methods used

Multimodel streamflow prediction is conducted using BMA and CBP-BMA. This section describes the formulation of the proposed model. The multimodel streamflow prediction process is depicted in Fig 4.2. Before describing the CBP-BMA procedure, we first assess the predictive skill of all available forecasts according to Shannon's information theory (Shannon, 1948).



Lead time	Day 1				Day 2				Day 3				Day 4				Day 5				Day 6				Day 7				Day 8				Day 9				Day 10			
6-hr time steps	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
RFC	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																							
HEFS	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
NWM	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
GEFS	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■											
NAEFS	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■												
SREF	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																							

Figure 4.1: Forecast horizon for each of input forecasts

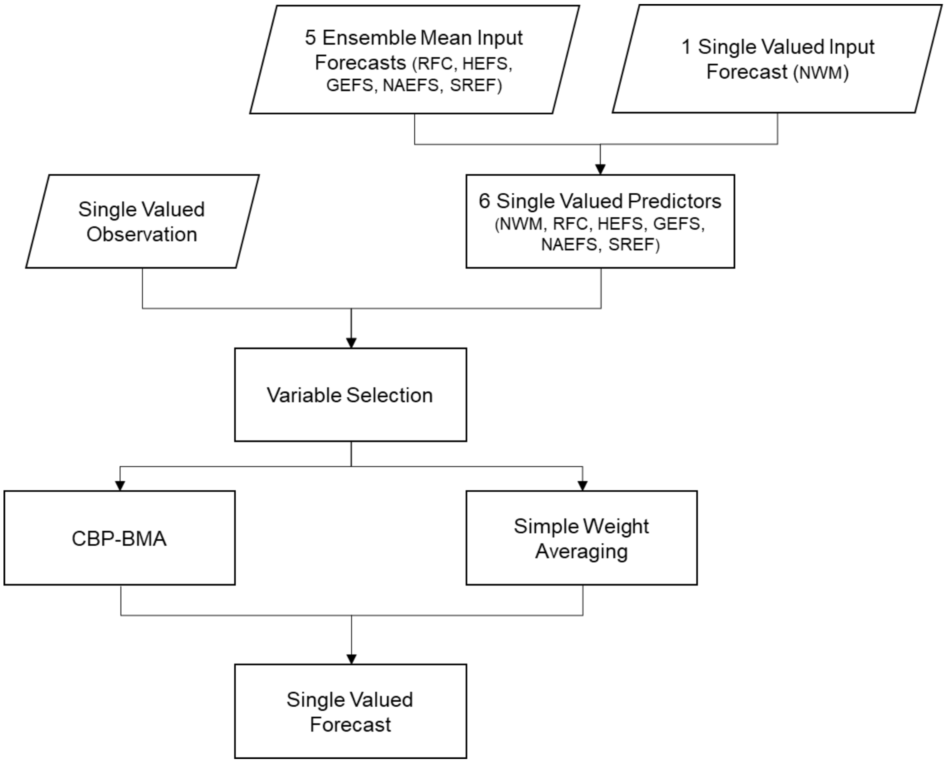


Figure 4.2: Schematic of the merging input forecasts (see section 3 for details)

4.3.1 Feature selection by Shannon's Entropy

The predictive skill of all available input forecasts is first evaluated for each lead time using Shannon entropy (1948). In the context of information theory, entropy is described as a measure of the amount of information required to describe a random variable (Keum and Coulibaly, 2017). As such, entropy may be considered as a measure of uncertainty in a random variable. The basis of the information theory is that the information obtained from an event with the probability of occurrence p is log $(1/\mathrm{p})$. The formal definition of entropy of an input variable X with pdf $p(x)$ is:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log\big(p(x_i)\big) = E\left[log\left(\frac{1}{p(x)}\right)\right] \tag{4.1}$$

where $H(X)$ is the marginal entropy of input forecast X in bits (the base of the logarithm is assumed to be equal to 2), $E[\,]$ denotes the expectation operation and n denotes the number of observations. For streamflow, which is a continuous variables, a finite number of discrete intervals are chosen. A similar procedure is used to calculate the information content for individual forecasts and observed flow. Dropping the i index for brevity, we write the joint entropy between each of the input forecasts of RFC, HEFS, NWM, GEFS, NAEFS and SREF and the verifying observed flow as:

$$H(X,Y) = -\sum_{x,y} p(x,y) log(p(x,y)) \tag{4.2}$$

In the above, $p(x,y)$ denotes the joint probability distribution for the input forecast, X and the verifying observed flow, Y; $H(X,Y)$ is the joint entropy between X and Y. The joint entropy measures the uncertainty in X and Y taken together. If the variables are independent, the joint entropy is equal to the summation of their marginal entropies. Eq. (4.3) shows the relation between the marginal and joint entropies:

$$H(X,Y) = H(X) + H(Y) - I(X,Y) \tag{4.3}$$

where $I(X,Y)$ is the mutual information (MI) between the input forecast X and observed flow Y. The MI is zero if X and Y are statistically independent. The MI is given by: (Shannon and Weaver, 1949)

$$I(X,Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{4.4}$$

where $H(X|Y)$ denotes the conditional entropy of X given Y, and H(Y|X) is the conditional entropy Y given X. The conditional entropy of X given Y is given by:

$$H(X|Y) = -E\left[log\big(p(x|y)\big)\right] = -\sum_{x,y} p(x,y) log(p(x|y)) \tag{4.5}$$

The conditional entropy shows how much uncertainty remains about the random variable X given the value of Y. In contrast to correlation coefficients that only address the linear relationship between two variables, mutual information contains information about both linear and nonlinear dependencies (Smith, 2015).

4.3.2 CBP-BMA

We implemented CBP-BMA in the R BMS package (Zeugner 2011) by replacing MLR with CBP-MLR. In this section, we describe MLR, CBP-MLR and BMA.

4.3.2.1 MLR

Given an $(n \times 1)$ dependent variable vector, Y and an $(n \times m)$ matrix of the independent variables where n is the number of observations and m is the number of independent variables, the MLR model is given by:

$$Y = X\beta_\gamma + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I) \tag{4.6}$$

where β_γ is the $(m \times 1)$ vector of regression coefficients to be estimated, and ε denotes an $(n \times 1)$ vector of independent and identically distributed zero-mean normal errors. The ordinary least squares (OLS) estimate for β_γ in Eq. (4.6), $\hat{\beta}_\gamma$, is given by minimizing $\|Y - Xb\|^2$, where $\|\cdot\|$ denotes the Frobenius norm, as:

$$\hat{\beta}_\gamma = (X^T X)^{-1} X^T Y \tag{4.7}$$

Variance of $\hat{\beta}_\gamma$ is given by:

$$\mathrm{E}[(\hat{\beta}_\gamma - \beta_\gamma)\left(\hat{\beta}_\gamma - \beta_\gamma\right)^T] = \sigma^2 (X^T X)^{-1} \tag{4.8}$$

4.3.2.2 CBP-MLR

To address the detrimental effects of CB in MLR, Jozaghi et al. (2021a) proposed CBP-MLR, which reduces Type-II CB by jointly minimizing the sum of errors squared, $\Sigma_{EV} = (Y - \hat{Y})^T(Y - \hat{Y})$, and the sum of the Type-II error squared:

$$J = \Sigma_{EV} + \alpha\Sigma_{CB} \tag{4.9}$$

In the above, α denotes the weight given to the CB penalty and Σ_{CB} is the CB penalty derived from the expectation of the Type-II error squared as in Seo (2013):

$$\Sigma_{CB} = E[(Y - E[\hat{Y}|Y])^T(Y - E[\hat{Y}|Y])] \tag{4.10}$$

The least-squares solution for the coefficients is given by Jozaghi et al. (2021a):

$$\hat{\beta}_{cb_\gamma} = (1 + \alpha)[X^TX + \alpha X^TY(Y^TY)^{-1}Y^TX]^{-1}X^TY \tag{4.11}$$

$$\mathrm{E}[(\hat{\beta}_{cb_\gamma} - \beta_{cb_\gamma})\left(\hat{\beta}_{cb_\gamma} - \beta_{cb_\gamma}\right)^T] = \sigma^2(A^TA)^{-1} \tag{4.12}$$

where $A = (XX^T)^{-1}X(1 + \alpha)^{-1}[X^TX + \alpha X^TY(Y^TY)^{-1}Y^TX]$. In Eq. (4.11), if we consider no penalty for CB (i.e., $\alpha = 0$), $\hat{\beta}_{cb_\gamma}$ is reduced to the least square solution for the MLR model of Eq. (4.7).

4.3.2.3 BMA

To address the model structural uncertainty, BMA employs all 2^m MLR or CBP-MLR models, where m is the number of input forecasts, and weight-averages all of 2^m models via the total probability law below. The model-weighted posterior distribution for coefficients, β , is given by (Jennifer A. Hoeting et al., 1999; Raftery, 1995):

$$p(\beta|y, X) = \sum_{\gamma=1}^{2^m} p\left(\beta\middle|M_\gamma, y, X\right). p(M_\gamma|y, X) \tag{4.13}$$

where M_γ are the candidate MLR models, $p\big(\beta\big|M_\gamma, y, X\big)$ denotes the conditional probability of the model coefficients, $p(M_\gamma|y, X)$ is the posterior model probability (PMP) of the model M_γ. The PMP of model M_γ may be obtained via Bayes' theorem as:

$$p\big(M_\gamma|y, X\big) = \frac{p\big(y|M_\gamma\ , X\big)p\big(M_\gamma\ \big)}{\sum_{\mathrm{i}=1}^{2m} p(y|M_i\ , X)p(M_i\ )} \tag{4.14}$$

where $p\big(Y|M_\gamma\ , X\big) = \int p(Y|\theta_i, M_i, X). p(\ \theta_i|M_i)\ d\theta_i$ is the integrated likelihood of the model M_i, $p(Y|\theta_i, M_i, X)$ is the likelihood function of the model M_i, θ_i denotes the vector of parameters (β_i, σ) in model M_i, and $p(M_i\ )$ denotes the prior probability that model M_i is the true model.

BMS uses Zellner's g-prior, i.e., $\beta|\sigma^2 \sim N(0, g\sigma^2(X^TX)^{-1})$ where g denotes the scalar hyperparameter (Goel and Zellner, 1986). Under the normal conjugate linear model, the posterior distribution of coefficients Eq. (4.6), $p(\beta_i|M_i, y, X)$ has a t-distribution with conditional mean and covariance of:

$$E[\beta_i|M_i, y, X, g] = \tfrac{g}{g+1}\hat{\beta}_i \tag{4.15}$$

$$Cov(\beta_i|y, X, g, M_i) = \frac{(y-\bar{y})'(y-\bar{y})}{N-3}\frac{g}{1+g}\left(1-\frac{g}{1+g}R_i^2\right)(X_i'X_i)^{-1} \tag{4.16}$$

In the above, $\hat{\beta}_i$ is the standard OLS estimator for i^{th} MLR or CBP-MLR model, N is the number of observations, and R_i^2 is the coefficient of determination given by:

$$R^2 = 1 - \frac{\big(y-X\hat{\beta}\big)'\big(y-X\hat{\beta}\big)}{(y-\bar{y})'(y-\bar{y})} \tag{4.17}$$

The hyperparameter g assesses how certain one is about the coefficients indeed being zero. A large g means large variance for the prior coefficients. It hence implies that one is very uncertain that the coefficients are zero. In contrast, a small g means that one is quite certain that the coefficients are indeed zero.

4.4 Evaluation

For comparative evaluation of BMA and CBP-BMA, we carried out 10-fold cross-validation for each forecast point. That is, we divided the period of record into ten equal subperiods for each lead time. A single subperiod is used for validation while the rest are used for parameter estimation. Comparative evaluation is focused on addressing the following questions:

1) What is the relative skill among different input forecasts? What are their strengths and weaknesses?
2) Does BMA improve over individual input forecasts?
3) Under what conditions does BMA perform well or poorly, and why?
4) Does CBP-BMA improve over BMA? What factors impact its performance?
5) What forecast attributes do BMA and CBP-BMA improve?
6) How does performance of BMA and CBP-BMA vary with hydroclimatology and hydrologic conditions?
7) How to model priors? Does the use of priors improve performance?
8) What is the marginal value of NWM in multimodel streamflow prediction?

For evaluation metrics, we use conditional and unconditional root mean square error (RMSE), mean error (ME) and correlation coefficient (CC). The mutual information metrics are used to assess the skill of input forecasts. In this work, we focus on predicting high flows for flood forecasting and all ranges of streamflows for water supply and environmental flow forecasting.

The 95th percentile of verifying observed flow is used to evaluate the model performance for prediction of high flows for flood forecasting.

4.5 Results

In this section, we first evaluate the predictive skill of all available input forecasts in order to select the skillful predictors. We then present the evaluation results in both conditional and unconditional senses.

4.5.1 Assessment of predictive skill of input forecasts

Figs 4.3ab show the MI (Battiti, 1994) values for all headwater and downstream forecast points in the study area, respectively. The MI measures the amount of information that an input forecast contains about the verifying observed flow and can be thought of as the reduction in uncertainty in predicting the (unknown) verifying observation given the input forecast. As such, low MI indicates a small reduction in uncertainty. Large MI idicates a large reduction in uncertainty. An MI of zero means that the observed flow and input forecast are independent. Figs 4.3ab indicate that the NWM forecast carries a small amount of information about the observed

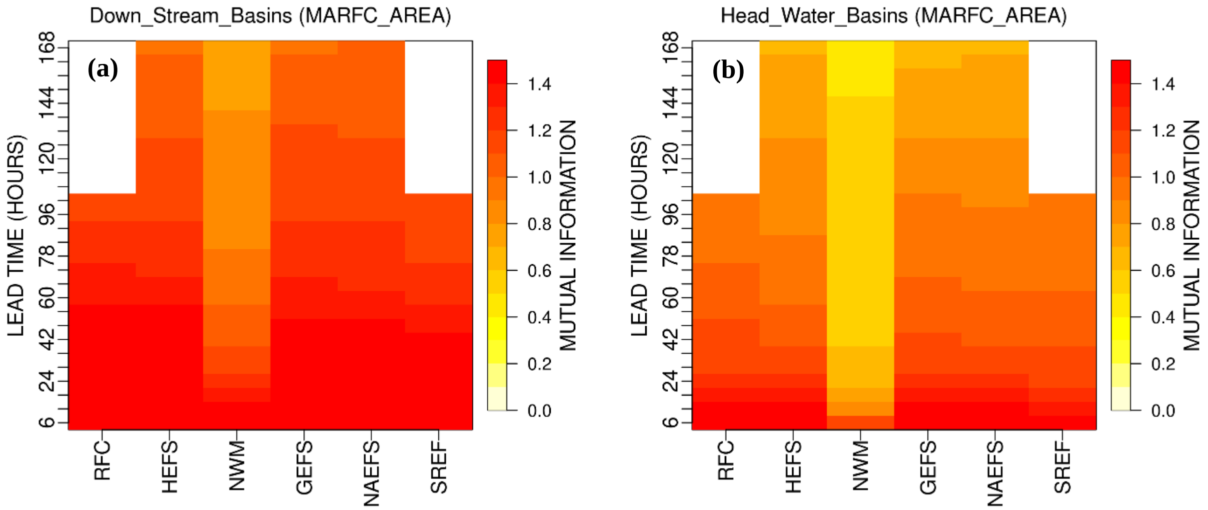


Figure 4.3: Unconditional Mutual Information of all input forecasts for total forecast horizon in a) all downstream forecast points, b) all headwater forecast points

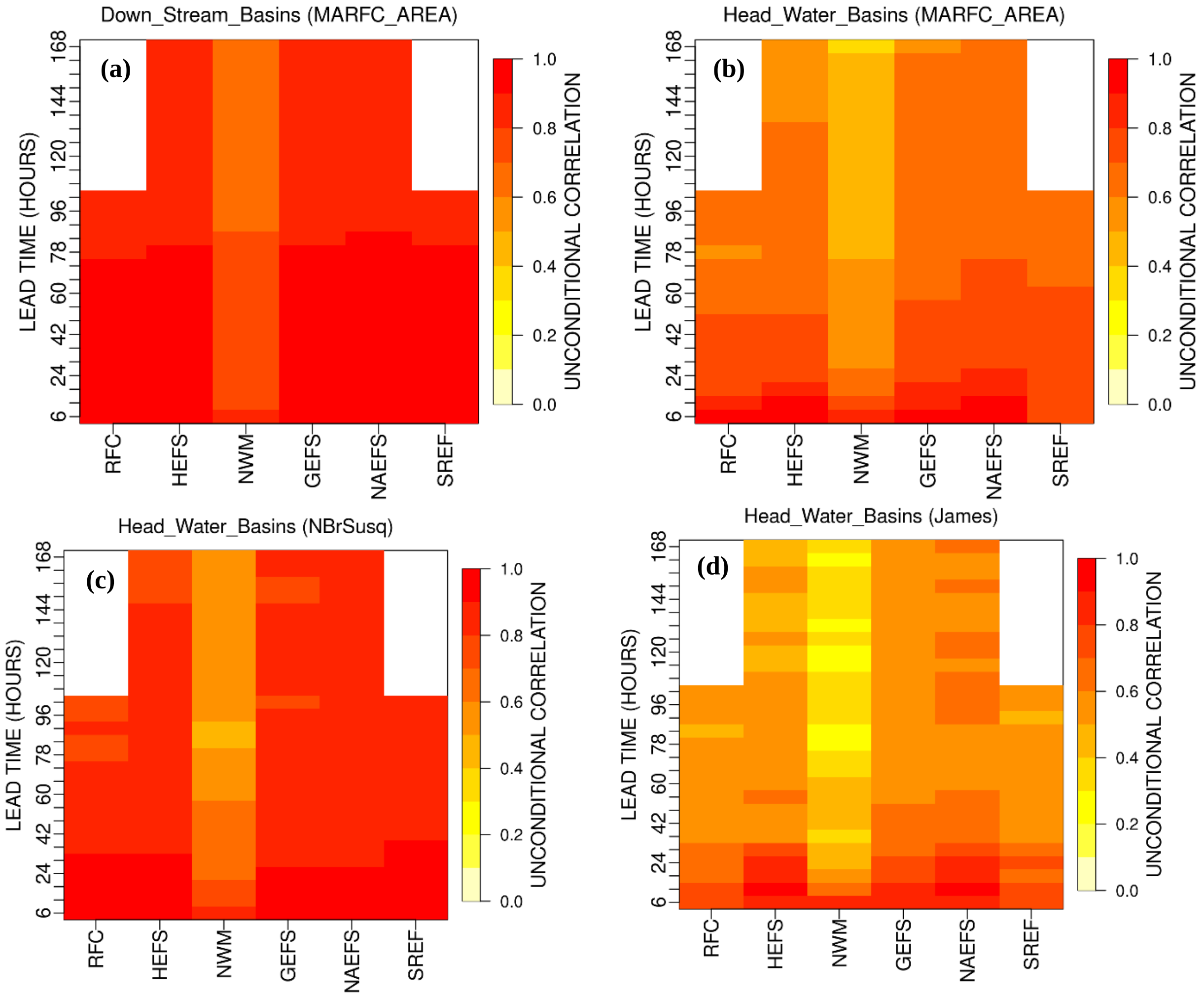


Figure 4.4: Unconditional Correlation Coefficient (CC) of all input forecasts for total forecast horizon in (a) all downstream forecast points, (b) all headwater forecast points, (c) as in (b) but for North Branch of Susquehanna forecast group, (d) as in (c) but for James forecast group

flow for both headwater and downstream forecast points. The figures show that the input forecasts have much larger predictive skills for downstream basins than headwater basins, and that the skill decreases with increasing lead time.

To compare the results of MI theory for linear dependency, we show in Figs 4.4ab the unconditional correlation coefficient of RFC, HEFS, NWM, GEFS, NAEFS, and SREF streamflow forecasts for the headwater and downstream forecast points over the entire forecast horizon. All unconditional performance or skill measures are quantified by including all ranges of observed flow. As Fig 4.3, the dark colors show a significant correlation between the observed flow and the input forecast, and the light colors indicate a small correlation. The small amount of skill is readily seen in NWM forecasts. To assess forecast skill for headwater forecast points in extreme ends of the service area, we choose the North Branch Susquehanna and James forecast groups which represent the northern- and southernmost forecast groups in the MARFC's service area, respectively (see Figs 4.4cd). Figs 4.4cd show that the predictive skill of input forecasts is significantly larger in the northern forecast group than in the southern.

Figs 4.5a-d show the unconditional RMSE and ME of the streamflow forecasts. The lead time is capped at 4.25 days so that there are six different forecasts available at all lead times. Figs 4.5ab indicate that the unconditional RMSE of NWM is significantly larger than the other forecasts for both downstream and headwater forecast points. As expected, the unconditional RMSE increases with lead times. Figs 4.5cd show the unconditional Mean Error (ME) for downstream and headwater forecast points, respectively. It is readily seen in Figs 4.5cd that the NWM forecast is unconditionally biased low, HEFS is unconditionally unbiased, and the MMEFS forecasts are unconditionally biased high. Figs 4.3 through 4.5 may be summarized as follows. In general, the skill of input forecasts varies significantly from one forecast group to another. The input forecasts are significantly more skillful in the northern forecast groups than in the southern. The skill in the southernmost forecast groups is sharply lower due to smaller skill in QPF and larger hydrologic uncertainty (Brown et al., 2014a, 2014b; Siddique et al., 2015). The RFC, HEFS, GEFS, NAEFS, and SREF streamflow forecasts are significantly more skillful than the NWM streamflow forecast. The HEFS is by far the best performing forecast in the MSE sense. The NWM forecast has consistently smaller predictive skill than all other streamflow forecasts for almost all forecast horizons.

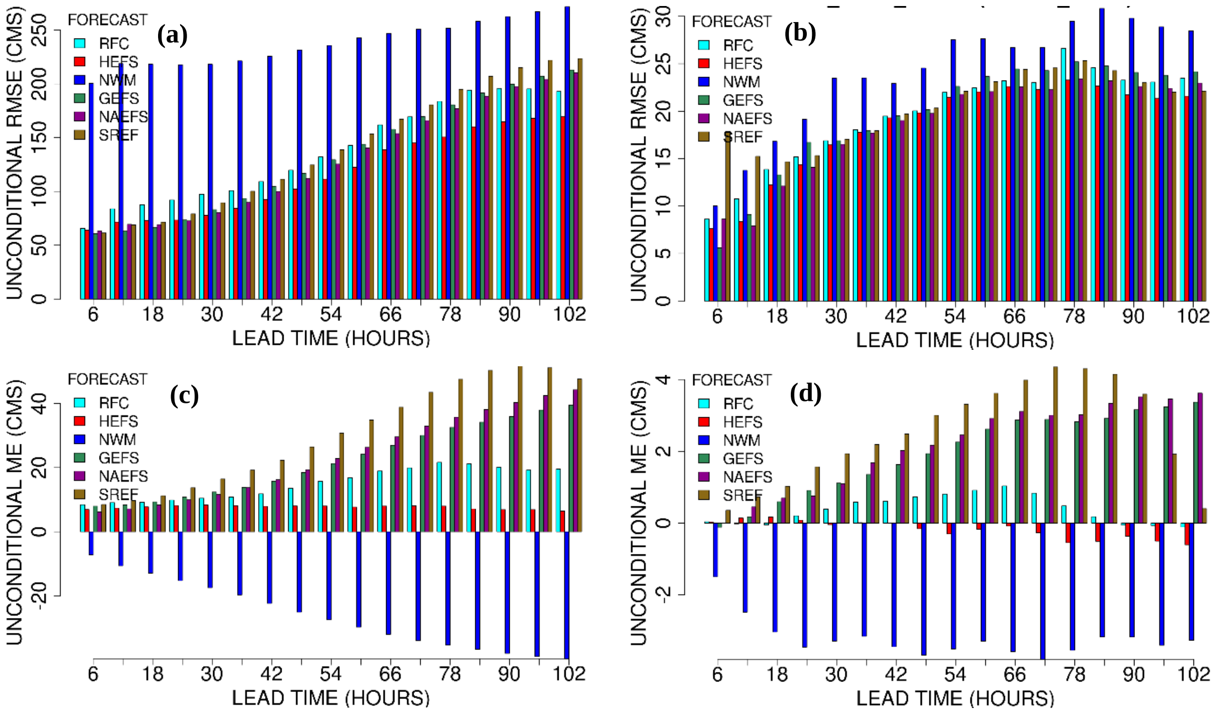


Figure 4.5: Unconditional RMSE of all input forecasts for total forecast horizon in (a) all downstream forecast points, (b) all headwater forecast points, (c) as in (a) but for ME, (d) as in (b) but for ME

4.5.2 Impact of skewness and heteroscedasticity

The performance of BMA depends on the predictability of streamflow. BMA assumes normality and homoscedasticity whereas streamflow is non-negative and generally skewed. As such, one may expect reduced performance for those forecast points where skewness and/or heteroscedasticity is ignificant. In this subsection, we aim to relate the two characteristics with the performance of BMA. To that end, we first assess the predictive skill of the BMA forecast for all forecast points in the entire study area. Figs 4.6ab show the number of BMA-favorable (green) and –unfavorable (red) forecast points for each forecast group. Fig 4.6a indicates the performance of the BMA for all 85 downstream forecast points. It is shown that BMA performs well for 60 forecast points whereas it performs relatively poorly for 25. Note that BMA performs well for the northern forecast groups of Chemung, North Branch Susquehanna, Delaware, Upper Main Stem Susquehanna, and Passaic whereas it performs relatively poorly for the southern forecast groups

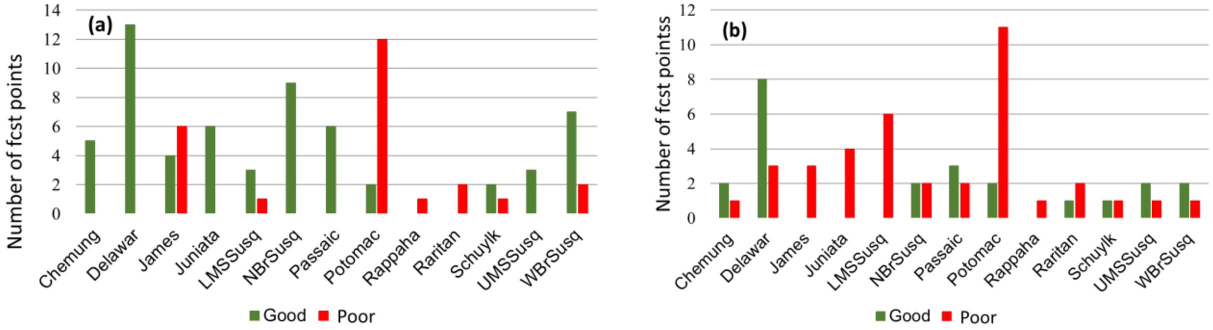


Figure 4.6: the BMA performance among different forecast groups for (a) downstream and (b) headwater

of James, Potomac, and Rappahannock. Fig 4.6b is the same as Fig 4.6a but for all 61 headwater forecast points. It is readily seen that the headwater forecast points are less favorable to BMA as only 23 forecast points are favorable to BMA. Figs. 4.7ab show the scatter plots of Absolut Mean Exponent (AME) of p-values versus the skewness of observed flow for BMA favorable and BMA unfavorable locations for lead time 48 hours, respectively. The AME is derived from the Breusch-Pagan test for heteroscedasticity of errors in a linear regression model. It measures how errors increase across the observed streamflow. The black solid curve attempts to separate the forecast points with low skewness and heteroscedasticity from the rest. Fig. 4.7a indicates that, among the 23 forecast points with favorable BMA performance, 15 forecast points correctly identified as BMA-favorable whereas the other 8 forecast points are rejected incorrectly. Fig. 4.7b shows that, among the 38 BMA-unfavorable headwater forecast points, 30 forecast points are correctly identified as BMA-unfavorable whereas only 8 forecast points are accepted incorrectly. Fig. 4.8a shows the AME vs. skewness for BMA-favorable downstream forecast points. Fig. 4.8a shows that 7 forecast points are misidentified as BMA-favorable whereas the other 53 forecast points are correctly identified as BMA-favorable. Fig. 4.8b is the same as Fig. 4.8a but for BMA-unfavorable forecast points. Fig. 4.8b shows that 18 forecast points are correctly identified as BMA-unfavorable whereas the other 7 forecast points are incorrectly identified as BMA-unfavorable.

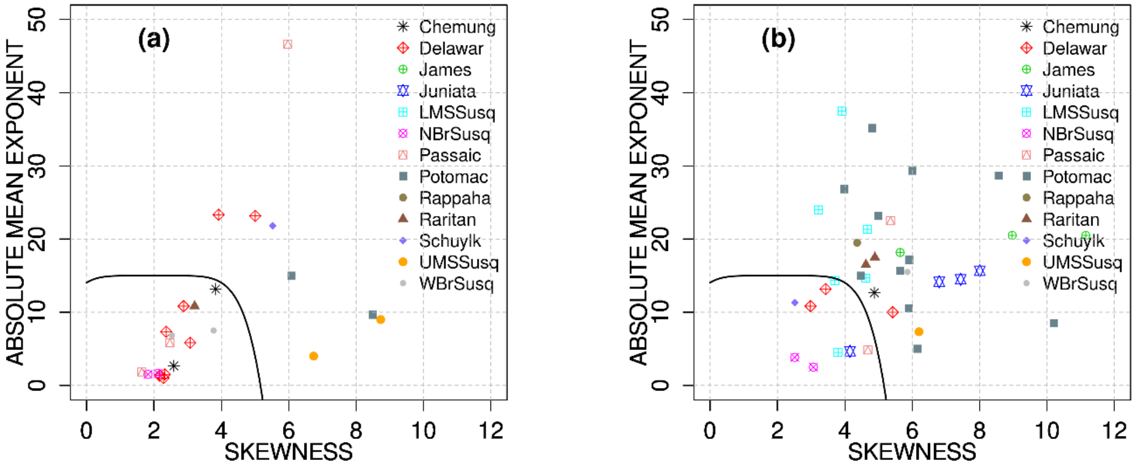


Figure 4.7: Scatterplots of AME vs. the skewness of observed flow in each of the headwater forecast points with lead time 48 hours for locations with (a) good and (b) poor BMA performance

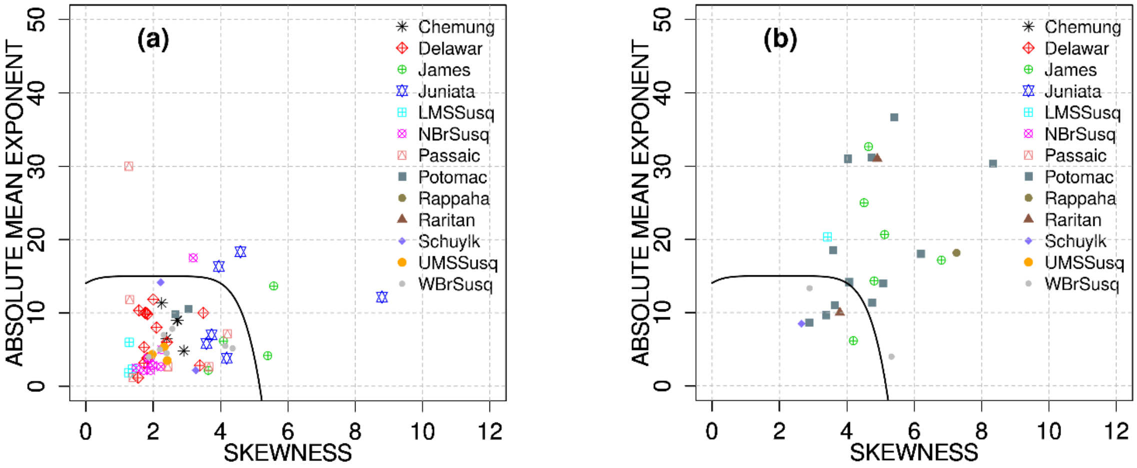


Figure 4.8: Scatterplots of AME vs. the skewness of observed flow in each of the downstream forecast points with lead time 48 hours for locations with (a) good and (b) poor BMA performance

Fig. 4.9a through 4.9c show the scatterplot of coefficient of variation (CV) versus skewness of observed streamflow, skewness versus basin size, and CV versus basin size, respectively. The blue and red symbols represent the headwater and downstream forecast points, respectively. A positive relationship is readily seen between CV and skewness in Fig. 4.9a. Expectedly, the headwater basins are more skewed and smaller in size whereas downstream basins are larger with smaller skewness. Figs 4.6-4.9 may be summarized as follows. The downstream forecast points, for which flow routing is the dominant hydrologic process, tend to be more predictable, less skewed, and

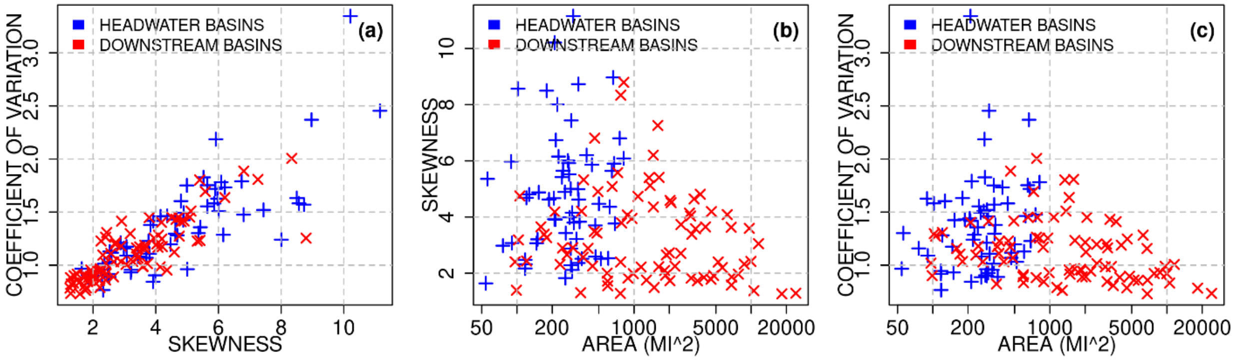


Figure 4.9: Scatterplots of (a) coefficient of variation vs. skewness, (b) skewness vs. basin area, and (c) coefficient of variation vs. basin area for all headwater and downstream basins

hence more favorable to BMA. The headwater forecast points, for which rainfall-runoff is the dominant hydrologic process, are less predictable, more skewed, and hence less favorable to BMA.

4.5.3 Impact of priors on BMA performance

This study compares two model priors available in the BMS package, the uniform and custom prior inclusion probability (PIP). A uniform model prior assumes equal weight for each of all combinations of MLR models. This study uses five different input forecasts as predictors. There are hence 32 different model combinations. A uniform model assigns a weight equals to 1/32 for each MLR model. A custom PIP assumes user-specified PIP values. In this work, they were prescribed by the posterior model probabilities obtained from forecast group-wide runs of BMA. As such, the custom PIP as implemented in this work amounts to using regional PIP values as the prior for the forecast point-specific. Figs 4.10ab show the conditional and unconditional RMSE of RFC, HEFS, GEFS, NAEFS, SREF, BMA, and CBP-BMA versus lead times where BMA and CBP-BMA used the custom PIP. Figs 4.10cd are the same as Figs 4.10ab, respectively, but used the uniform PIP. These and similar results for other forecast points indicate that the use of the custom PIP does not consistently improve over the uniform PIP, but that more extensive, larger-sample evaluation is necessary to draw firm conclusions.

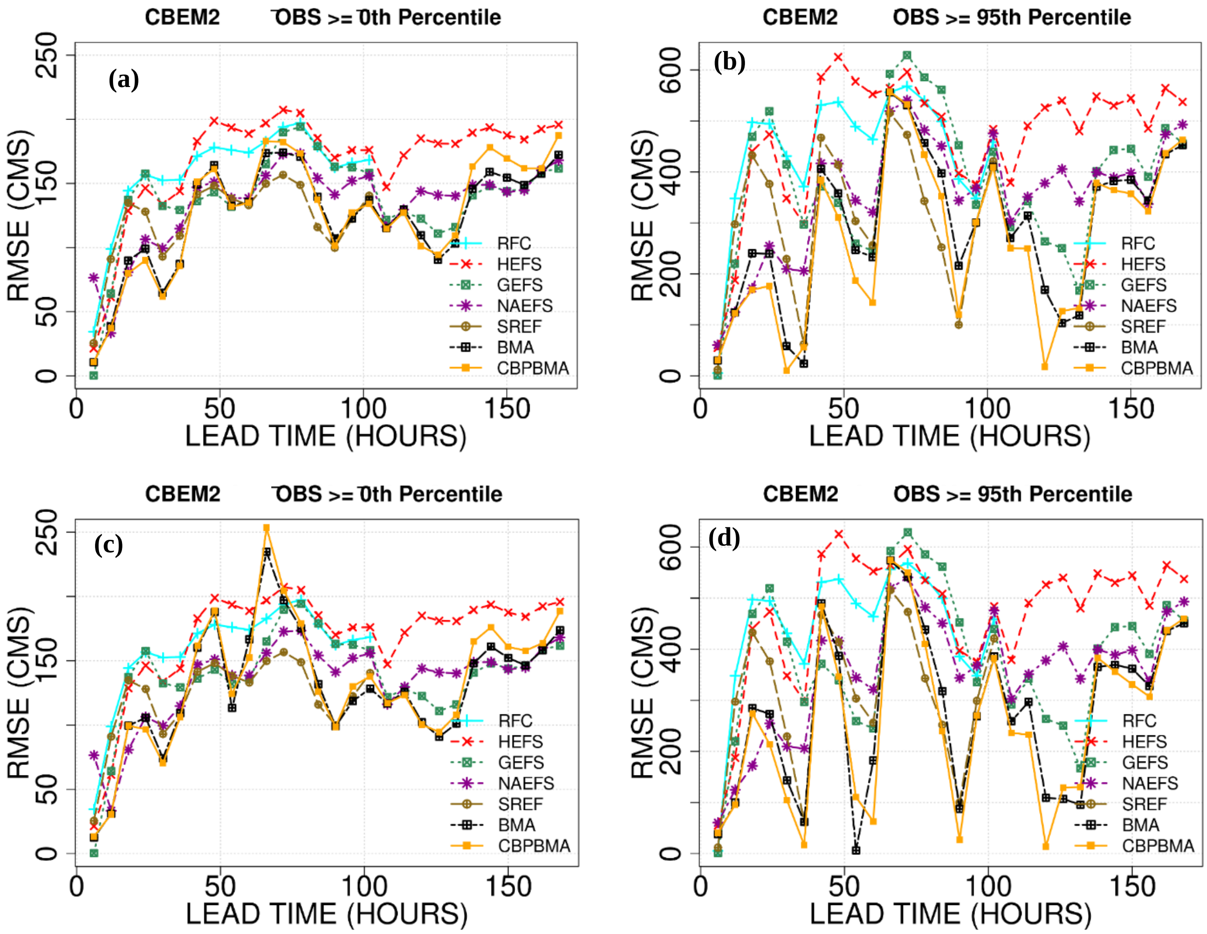


Figure 4.10: (a) the unconditional RMSE, (b) the conditional RMES vs. lead times using custom PIP as a model prior, (c) as (a) but using uniform model prior, and (d) as (b) but using uniform model prior

4.5.4 Comparative evaluation results

Figs. 4.11 ab show the correlation coefficient of RFC, HEFS, GEFS, NAEFS, SREF, BMA, and CBP-BMA forecasts for the headwater forecast points in the North Branch Susquehanna forecast group conditional on truth exceeding the amount of the x-axis for lead times of 1 and 4 days, respectively. The values on the y axis at $x = 0$ represent the correlation coefficients conditional on the verifying observed flow being nonzero, i.e., the unconditional correlation. As such, we refer to the outcomes for $x > 0$ and $x = 0$ as conditional and unconditional performance, respectively. Fig. 4.11a shows that both BMA and CBP-BMA have larger correlation with the observed than the input forecasts for all ranges of streamflow for Day-1 prediction. Among all input forecasts, HEFS has by far the best performance in terms of correlation coefficient, and the MMEFS forecasts, i.e., GEFS, NAEFS, and SREF, perform better for flows greater than 40 (cms). Fig. 4.11b is the same as Fig. 4.11a but for a lead time of 4 days. It is readily seen that correlation is significantly reduced for all forecasts at this lead time. The BMA and CBP-BMA are still the best performing forecast for flows less than 30 (cms). It is seen that NAEFS, GEFS, and SREF are perform better for larger flows whereas HEFS does not.

Fig. 4.12 is the same as Fig. 4.11 but for RMSE in logarithmic scale. The gray area shown in the figure indicates the sample size whose logarithmic axis is plotted at the right end of the figure. Similarly to Fig. 4.11a, Fig. 4.12a indicates good performance of BMA and CBP-BMA in both conditional and unconditional sense. The HEFS has the smallest RMSE among all input forecasts but for larger flows.

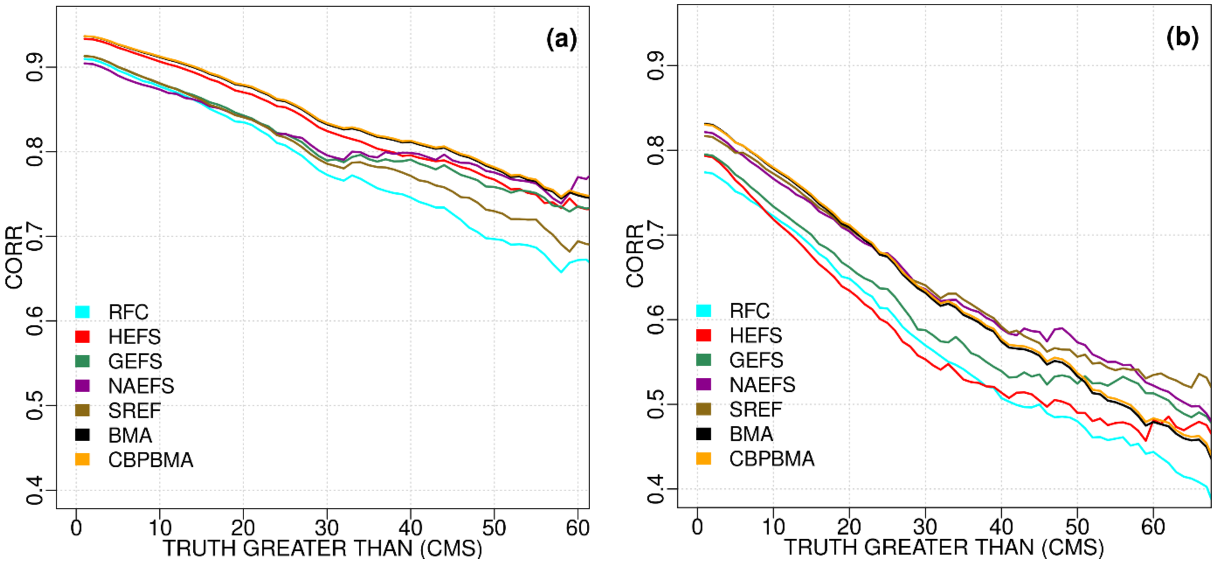


Figure 4.11: Correlation coefficient of RFC, HEFS, GEFS, NAEFS, SREF, BMA, and CBP-BMA forecasts over headwater forecast points in North Branch Susquehanna conditional on truth exceeding amount of the x-axis for lead times of (a) 1, and (b) 4 days

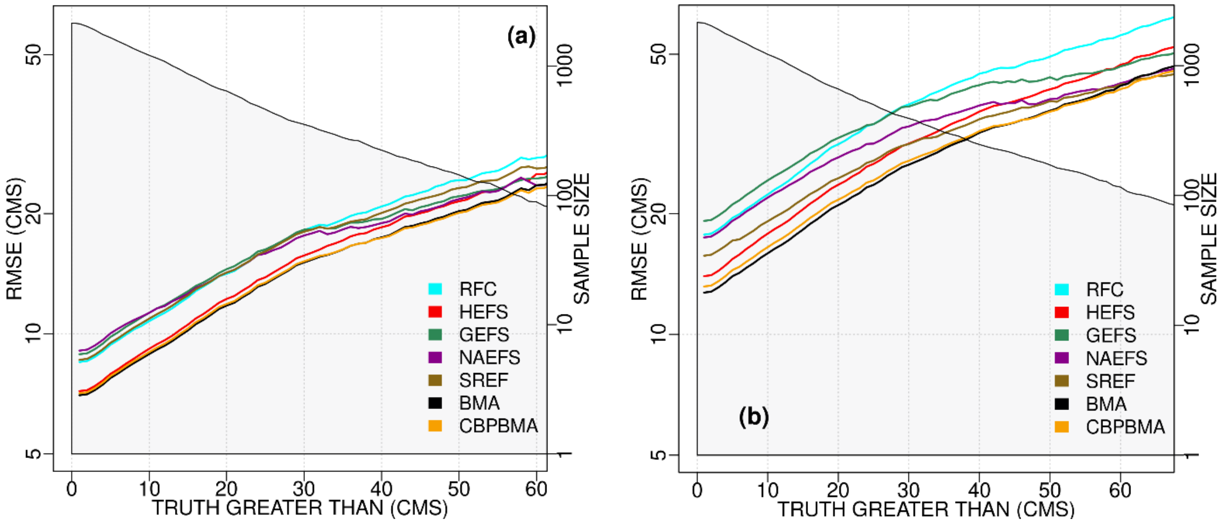


Figure 4.12: As in Fig. 4.11, but for RMSE values on the y-axis in logarithmic scale

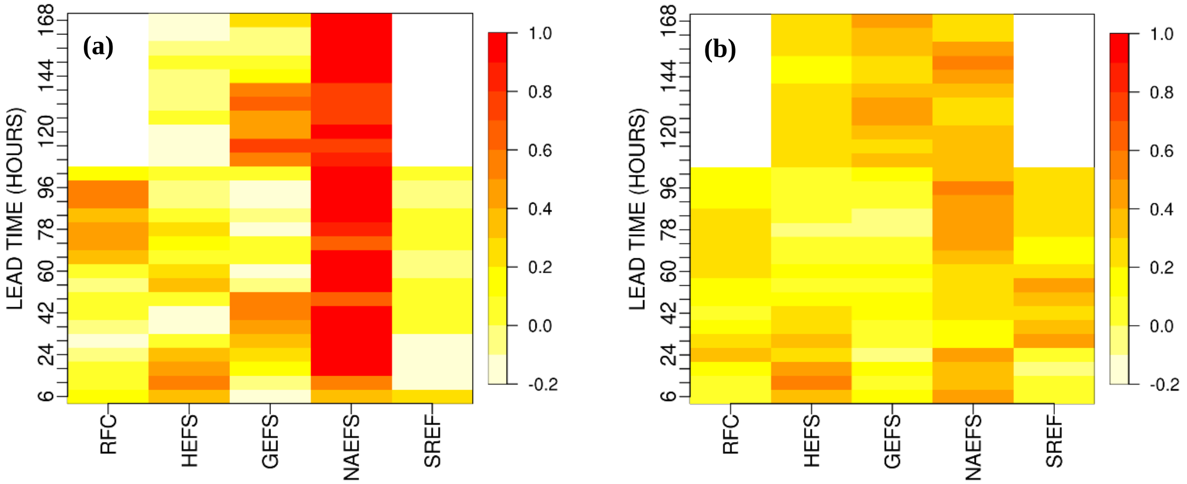


Figure 4.13: Expected value of CBP-BMA coefficients for each input forecast and each lead time for (a) James and (b) Delaware forecast groups.

To show the contribution of each input forecast as a function of lead time, Figs. 4.13ab plot the unconditional CBP-BMA coefficients for RFC, HEFS, GEFS, NAEFS, and SREF for the James and Delaware forecast groups, respectively. The former is located in the southern part of the study area, and the latter is a northern forecast group. It is seen that the NAEFS forecast contributes the most to CBP-BMA for James whereas the contributions of input forecasts are evenly spread for Delaware. The above suggests that the predictive skill among different forecasts in the northern groups is more uniform than in the southern. The GEFS forecast shows the most significant variations across the forecast lead times.

Fig. 4.14ab plots RMSE vs. lead time of RFC, HEFS, GEFS, NAEFS, SREF, BMA, and CBP-BMA for all ranges of observed flow and observed flow exceeding the 95th percentile, respectively. Fig. 4.14a indicates that both BMA and CBP-BMA unconditionally outperform all input forecasts at all lead times. The NAEFS, as shown in Fig. 4.13a, has the best skill among all input forecasts. Fig. 4.14b shows the conditional RMSE of all forecasts for downstream forecast points in the James forecast group. As expected, BMA falls behind CBP-BMA, particularly at larger lead times where the CB is increased. The CBP-BMA outperforms all input forecasts except at larger lead times where the number of available input forecast is reduced to three. The HEFS has the smallest overall skill in the conditional sense among all input forecasts.

Figs 4.15ab show the unconditional RMSE for all headwater forecast points in dry (Apr to Oct) and wet (Nov to Mar) seasons, respectively. It is shown that BMA outperforms all forecasts in the unconditional RMSE in both dry and wet seasons. However, the margin of improvement by BMA and CBP-BMA is more significant in the wet season than in the dry season. Figs 4.15cd show the unconditional sense for all downstream forecast points in dry and wet seasons, respectively. It is seen that forecasts have the largest skill for downstream points in the wet season for which HEFS outperforms all other input forecasts owing to the greatly reduced CB resulting from spatiotemporal smoothing. The BMA and CBP-BMA outperform all input forecasts in the unconditional sense in both dry and wet seasons, and their margin of improvement is larger in the wet season than in the dry season.

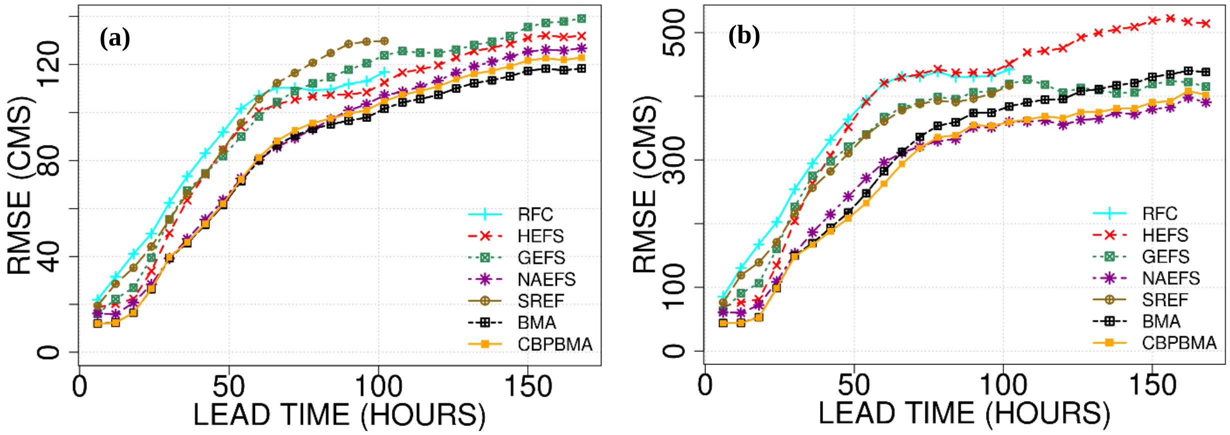


Figure 4.14: RMSE vs. lead time of RFC, HEFS, GEFS, NAEFS, SREF, BMA, and CBP-BMA (a) for all ranges of observed flow and (b) for the verifying observed flow exceeding the 95th percentile for James forecast group.

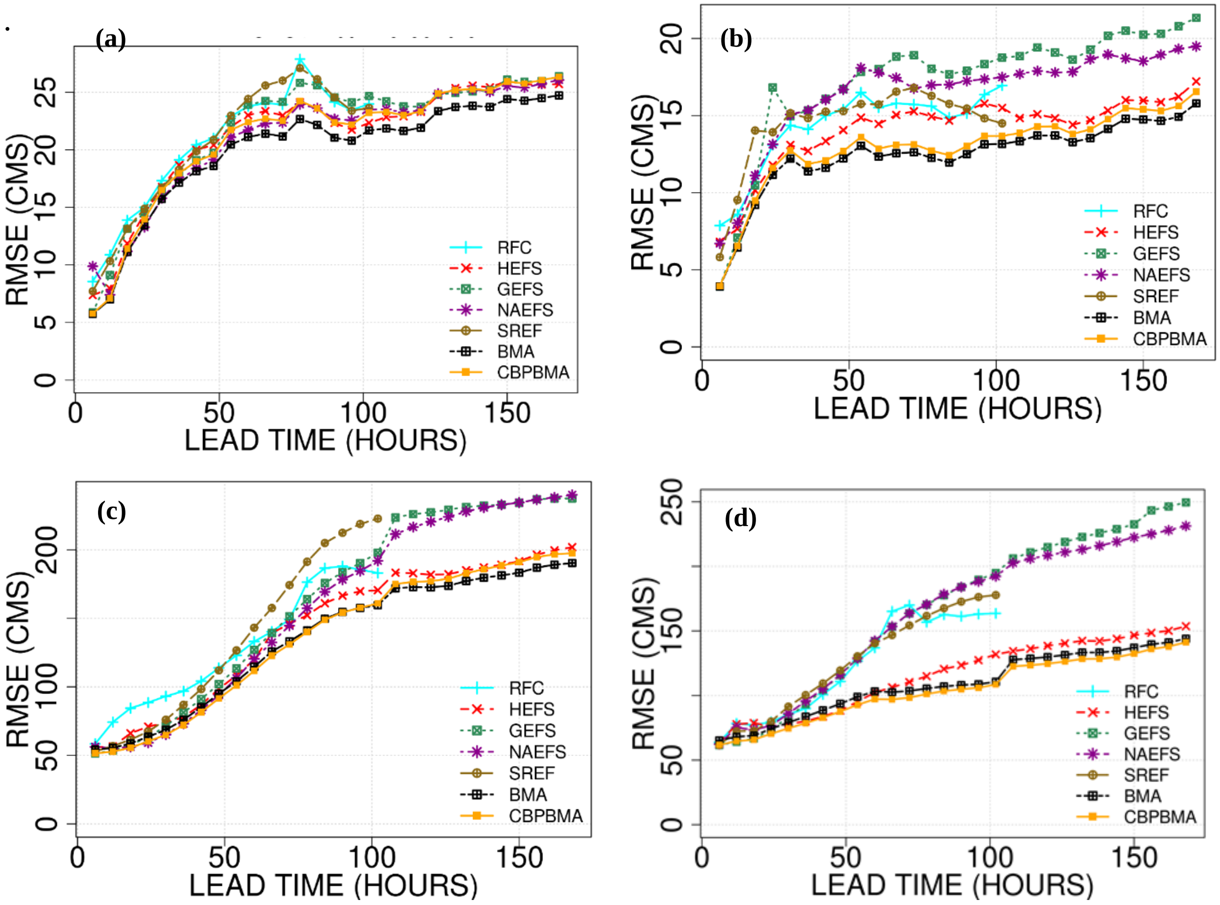


Figure 4.15: Unconditional RMSE vs. lead times of (a) dry season, (b) wet season for all headwater forecast points, (c) as (a) but for downstream forecast points, and (d) as (b) but for downstream forecast points

4.6 Conclusions and Future Research Recommendations

The principal conclusion of this work is as follows. The relative skill of input forecasts significantly varies with the range of the verifying observed flow. The effectiveness of multimodel streamflow prediction depends largely on streamflow predictability and the predictive skill in the input forecasts. It is found that skewness and heteroscedasticity are very useful indicators of the performance of BMA. The smaller the skewness is and the more homoscedastic the streamflow is, the more likely BMA and CBP-BMA perform effectively. In general, the streamflow forecasts forced by the NAEFS, and the SREF, are better-suited for flood forecasting whereas those generated with the HEFS are better-suited for water supply and environmental flow forecasting. It is shown that BMA generally outperforms the best individual forecasts in the MSE sense for all ranges of observed flow. CBP-BMA is generally superior to BMA for prediction of large flows. Generally speaking, the multimodel prediction is most effective for downstream locations in the cool season and for northern forecast groups owing to the larger predictability. Additional evaluation is needed to assess the performance of CBP-BMA further and to prescribe α optimally in real-time.

Chapter 5: General Conclusions and Future Research Recommendations

The principal conclusion from Element 1 (see Ch 2) is that, to produce multisensor estimates that are more accurate than the ingredient quantitative precipitation estimates (QPE) for all precipitation amounts, it is necessary to address the conditional bias (CB), and that adaptive conditional bias-penalized cokriging (CBPCK) improves the estimation of significant amounts of precipitation by explicitly considering the CB. It is shown that beyond the reduction in root mean square error (RMSE) due to mean-field bias (MFB) correction, both ordinary cokriging (OCK) and adaptive CBPCK additionally reduce the unconditional RMSE of radar-only QPE by 16 and 9 % over the continental US (CONUS) for the more and less predictable Oct 7-9, 2016, and Sep 13-30, 2015, events, respectively, and that adaptive CBPCK improves over OCK for estimation of hourly precipitation exceeding about 1 mm. Jointly, MFB correction and adaptive CBPCK reduce the RMSE of the radar-only QPE by about 16 to 26% for the more predictable Oct 7-9, 2016, events and by about 10 to 17% for the less predictable Sep 13-30, 2015, events for all ranges of precipitation amounts. It is shown that for the Sep 2015 events, fusing the MFB-corrected radar QPE with the MFB-corrected SCaMPR QPE reduces the unconditional RMSE of radar-only QPE by about 12% over the CONUS, whereas the reduction by MFB-corrected radar QPE alone over radar-only QPE is about 8%. For the western half of the CONUS, where the rain gauge network is sparser, and the radar QPE is less skillful, the margin of reduction increases to 6% from the above 4%. The conditional performance of the fused QPE, however, falls below that of the MFB-corrected radar QPE as the conditioning amount exceeds about 7 mm of hourly precipitation. The above suggests that the SCaMPR product should be used selectively in multisensor QPE.

The main conclusions from Element 2 (see Ch 3) are as follow. The relative performance among different input forecasts varies most significantly with the range of the verifying observed flow. Generally speaking, the streamflow forecasts forced by the North American Ensemble Forecast System, or NAEFS, and the Short-Range Ensemble Forecast, or SREF, are better-suited for flood forecasting. In contrast, those generated with the Hydrologic Ensemble Forecast Service, or HEFS, are better suited for water supply and environmental flow forecasting. It is shown that CompMLR generally outperforms the best individual forecasts in the MSE sense for observed flow exceeding the 95th percentile, for all ranges of observed flow, for less predictable headwater forecast points, and for more predictable downstream points for lead times of up to about six days. If forecasting high flows is of little importance, CB is not likely to be an issue and MLR may suffice. Large variations are seen in the MLR and CBP-MLR weights associated with different input forecasts across the 7-day forecast horizon and across different forecast points. The above suggests that selecting a single "best" forecast system for the forecast point of interest or a single "best" forecast for the day or event at hand is likely a difficult proposition. The distinguishing value of CB-aware MLR is that it produces predictions that are generally superior to the best input forecast both in the conditional sense and in the unconditional sense under wide-ranging conditions of predictability and predictive skill.

The main conclusions from Element 3 (see Ch 4) are as follows. BMA based on the Bayesian normal conjugate linear model (Walter et al. 2009) with Zellner's g-prior (Zellner 1984) generally improves over input forecasts. As with MLR, the improvement is larger and more consistent for downstream forecast points for which predictability is larger. Forecast point-specific validation results indicate that a combination of skewness and homoscedasticity is a good indicator of how effective BMA may be for multimodel streamflow prediction. It is seen that the smaller the skewness and heteroscedasticity of streamflow is at the forecast point, the more effective BMA is for multimodel streamflow prediction. To assess the impact of modeling priors, limited comparative evaluation was carried out using the posterior model probability obtained from forecast group-wide training of BMA as the prior inclusion probabilities (PIP) vs. assuming uniform model prior (i.e., all input forecasts are equally likely to be included). The results indicate that the PIP prior does not consistently improve over the uniform model prior. Given that streamflow response is generally highly flow-dependent and location-specific, it is very likely that a rather sophisticated strategy is necessary to specify the spatiotemporal domain over which the prior is to be inferred, and the likelihood function is to be constructed (Krzysztofowicz 1999). CBP-BMA generally improves over BMA for high flows (i.e., in the conditional sense) but at the expense of some deterioration in the unconditional sense (i.e., CBP-BMA performs slightly poorly vs. BMA for all flows). In this work, $\alpha = 1$ was used in CBP-BMA without explicit optimization as was done in CompMLR. Whereas optimizing α may be expected to improve conditional performance, it is unlikely to improve unconditional performance due to the fact that CBP-BMA is based on linear regression (Shen et al., 2019). Limited comparative evaluation indicates that CompMLR and CBP-BMA perform comparably. The above suggests that the nonlinear regression made of two linear regressions in CompMLR is comparable to weight-averaging of multiple Bayesian linear regressions of different combinations of input forecasts in CBP-BMA. An extensive sample evaluation is necessary, however, to assess the relative performance in different hydroclimatological regimes. The results presented in this work for Elements 2 and 3 are based on a period of record of less than two years. Systematic assessment of data requirements vs. performance, particularly for forecasting extreme floods, is necessary and that of potential benefits of regionalization for space-for-time trade-off.

Bibliography

Abel, A.B., 2017. Classical measurement error with several regressors. Wharton School of the University of Pennsylvania, National Bureau of Economic Research.

Adams, T., 2015. Verification of the NOAA/NWS MMEFS Operational Hydrologic Ensemble Forecasting System in the Ohio River Valley. World Environmental and Water Resources Congress, Austin, Tx. https://doi.org/10.1061/9780784479162.264

Adams, T.E., 2016. Flood Forecasting in the United States NOAA/National Weather Service, in: Flood Forecasting. Elsevier, pp. 249–310. https://doi.org/10.1016/B978-0-12-801884-2.00010-4

Adnan, Liang, El-Shafie, Zounemat-Kermani, Kisi, 2019. Prediction of Suspended Sediment Load Using Data-Driven Models. Water 11, 2060. https://doi.org/10.3390/w11102060

Aghakouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., and Amitai, E., 2011: Evaluation of satellite-retrieved extreme precipitation rates across the central United States. *J. Geophys. Res. Atmos.*, **116**, D02115, https://doi.org/10.1029/2010JD014741.

Ajami, N.K., Duan, Q., Gao, X., Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: Application to distributed model intercomparison project results. J. Hydrometeorol. https://doi.org/10.1175/JHM519.1

Alizadeh, B., 2019. Improving post processing of ensemble streamflow forecast for short-to-long ranges: a multiscale approach, PhD dissertation. The University of Texas at Arlington, 125 pp, https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/28663/ALIZADEH-DISSERTATION-2019.pdf?sequence=1.

Alizadeh, B., Limon, R.A., Seo, D.-J., Lee, H., Brown, J., 2020. Multiscale Postprocessor for

Ensemble Streamflow Prediction for Short to Long Ranges. J. Hydrometeorol. 21, 265–285. https://doi.org/10.1175/JHM-D-19-0164.1

Ashouri, H., Hsu, K. L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Prat, O. P., 2015: PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bull. Amer. Meteor. Soc.*, **96**, 69–83, https://doi.org/10.1175/BAMS-D-13-00068.1.

Anderson, E., 2006. Snow Accumulation and Ablation Model – SNOW-17. https://www.wcc.nrcs.usda.gov/ftpref/wntsc/H&H/snow/AndersonSnow17.pdf.

Anderson, E.A., 1973. National Weather Service river forecast system: Snow accumulation and ablation model. Silver Springs, Md. https://www.wcc.nrcs.usda.gov/ftpref/wntsc/H&H/snow/AndersonHYDRO17.pdf.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J.-L., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol. 529, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001

Baharvand, S., Jozaghi, A., Fatahi-Alkouhi, R., Karimzadeh, S., Nasiri, R., Lashkar-Ara, B., 2020. Comparative Study on the Machine Learning and Regression-Based Approaches to Predict the Hydraulic Jump Sequent Depth Ratio. Iran. J. Sci. Technol. Trans. Civ. Eng. https://doi.org/10.1007/s40996-020-00526-2

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks 5, 537–550. https://doi.org/10.1109/72.298224

Beven, K., 2001. How far can we go in distributed hydrological modelling? Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-5-1-2001

Breidenbach, J. P., D.-J. Seo, P. Tilles, and C. Pham, 2001: Seasonal variation in multiradar coverage for WSR-88D precipitation estimation in a mountainous region. Preprints, Conf. on Precipitation Extremes: Prediction, Impacts, and Responses, Albuquerque, NM. *Amer. Meteor. Soc.*, P2.38, [Available online at https://ams.confex.com/ams/annual2001/techprogram/paper_18668.htm.]

Breidenbach, J. P., D.-J. Seo, P. Tilles, and K. Roy, 1999: Accounting for radar beam blockage patterns in radar-derived precipitation mosaics for River Forecast Centers. Preprints, 15th Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Dallas, TX. *Amer. Meteor. Soc.*, 179–182.

Brown, J.D., He, M., Regonda, S., Wu, L., Lee, H., Seo, D.-J., 2014a. Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. J. Hydrol. 519, 2847–2868. https://doi.org/10.1016/j.jhydrol.2014.05.030

Brown, J.D., Seo, D.-J., 2013. Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions. Hydrol. Process. 27, 83–105. https://doi.org/10.1002/hyp.9263

Brown, J.D., Wu, L., He, M., Regonda, S., Lee, H., Seo, D.-J., 2014b. Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental design and forcing verification. J. Hydrol. 519, 2869–2889. https://doi.org/10.1016/j.jhydrol.2014.05.028

Candille, G., 2009. The Multiensemble Approach: The NAEFS Example. Mon. Weather Rev. 137, 1655–1665. https://doi.org/10.1175/2008MWR2682.1

Carl, G., Kühn, I., 2008. Analyzing spatial ecological data using linear regression and wavelet analysis. Stoch. Environ. Res. Risk Assess. 22, 315–324. https://doi.org/10.1007/s00477-

007-0117-2

Chen, C.S., Liu, C.H., Su, H.C., 2008. A nonlinear time series analysis using two-stage genetic algorithms for streamflow forecasting. Hydrol. Process. https://doi.org/10.1002/hyp.6973

Chow, V.T., Maidment, D.R., Mays, L.W., 1988. Applied Hydrology. McGraw-Hill, New York.

Chua, S. H., and Bras, R. L., 1982: Optimal estimators of mean areal precipitation in regions of orographic influence. *J. Hydrol.*, **57**, 23-48, https://doi.org/10.1016/0022-1694(82)90101-9.

Ciach, G.J., Morrissey, M.L., Krajewski, W.F., 2000. Conditional Bias in Radar Rainfall Estimation. J. Appl. Meteorol. 39, 1941–1946. https://doi.org/10.1175/1520-0450(2000)039<1941:CBIRRE>2.0.CO;2

Cocks, S. B., J. Zhang, S. M. Martinaitis, Y. Qi, B. Kaney, K. Howard. 2017: MRMS QPE performance east of the Rockies during the 2014 warm season. *J Hydrometeorol*, **18**, 761–775, https://doi.org/10.1175/JHM-D-16-0179.1

Conway, D.A., Roberts, H. V., 1983. Reverse Regression, Fairness, and Employment Discrimination. J. Bus. Econ. Stat. 1, 75. https://doi.org/10.2307/1391775

Cotter, J., 2015. USACE Flood Risk and Water Supply Activities in the Dallas Fort Worth Area, Sustainable Urban Water Workshop. The University of Texas at Arlington, Arlington, TX.

Cui, B., Toth, Z., Zhu, Y., Hou, D., 2012. Bias Correction for Global Ensemble Forecast. Weather Forecast. 27, 396–410. https://doi.org/10.1175/WAF-D-11-00011.1

Curran, J.C., 2008. Modeling future flows in the Blanco watershed under various development and rainfall scenarios. Texas J. Sci. 59(3), 209–233.

Daly, C., Neilson, R. P., and Phillips, D. L., 1994: A Statistical-Topographic Model for Mapping Climatological Precipitation over Mountainous Terrain. *J. Appl. Meteor.*, **33**, 140–158, https://doi.org/10.1175/1520-0450(1994)033<0140:astmfm>2.0.co;2.

Danandeh Mehr, A., Kahya, E., Şahin, A., Nazemosadat, M.J., 2015. Successive-station monthly streamflow prediction using different artificial neural network algorithms. Int. J. Environ. Sci. Technol. https://doi.org/10.1007/s13762-014-0613-0

David, M., Marcotte, D., Soulié, M., 1984. Conditional Bias in Kriging and a Suggested Correction, in: Geostatistics for Natural Resources Characterization. Springer Netherlands, Dordrecht, pp. 217–230. https://doi.org/10.1007/978-94-009-3699-7_13

DeFigueiredo, R., 1980. Implications and applications of Kolmogorov's superposition theorem. IEEE Trans. Automat. Contr. 25, 1227–1231. https://doi.org/10.1109/TAC.1980.1102536

DelSole, T., Tippett, M.K., 2007. Predictability: Recent insights from information theory. Rev. Geophys. 45. https://doi.org/10.1029/2006RG000202

Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., 2014. The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. Bull. Am. Meteorol. Soc. 95, 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1

Deutsch, C. V., and Journel, A. G., 1998: *GSLIB: Geostatistical software library and user's guide*. Oxford University Press, Oxford, UK, https://doi.org/10.1016/0098-3004(94)90041-8.

Diks, Cees G. H., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Environ. Res. Risk Assess. 24, 809–820. https://doi.org/10.1007/s00477-010-0378-z

Diks, Cees G.H., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Environ. Res. Risk Assess. https://doi.org/10.1007/s00477-010-0378-z

Doblas-Reyes, F.J., Deque, M., Piedelievre, J.P., 2000. Multi-model spread and probabilistic seasonal forecasts in PROVOST. Q. J. R. Meteorol. Soc. https://doi.org/10.1002/qj.49712656705

Doblas-Reyes, F.J., Hagedorn, R., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. Tellus, Ser. A Dyn. Meteorol. Oceanogr. https://doi.org/10.1111/j.1600-0870.2005.00104.x

Du, J., McQueen, J., DiMego, G., Black, T., Juang, H., Rogers, E., Ferrier, B., Zhou, B., Toth, Z., Tracton, S., 2004. 21.3 The NOAA/NWS/NCEP Short Range Ensemble Forecast (SREF) system: Evaluation of an initial condition vs multiple model physics ensemble approach. Bull. Am. Meteorol. Soc. 2329–2338.

Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007a. 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

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

Emery, X., 2006. Multigaussian kriging for point-support estimation: incorporating constraints on the sum of the kriging weights. Stoch. Environ. Res. Risk Assess. 20, 53–65. https://doi.org/10.1007/s00477-005-0004-7

Fedora, M.A., Beschta, R.L., 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

Freedman, D.A., 2005. Statistical Models: Theory and Practice. Cambridge University Press, New York.

Frost, C., Thompson, S.G., 2000. Correcting for regression dilution bias: comparison of methods for a single predictor variable. J. R. Stat. Soc. Ser. A (Statistics Soc. 163, 173–189. https://doi.org/10.1111/1467-985X.00164

Fuller, W., 1987. Measurement error models. John Wiley and Sons, Chichester.

Furl, C., Sharif, H., Zeitler, J.W., Hassan, A. El, Joseph, J., 2018. Hydrometeorology of the catastrophic Blanco river flood in South Texas, May 2015. J. Hydrol. Reg. Stud. 15, 90–104. https://doi.org/10.1016/j.ejrh.2017.12.001

Gandin, L. S., 1966: Objective analysis of meteorological fields, Translated from the Russian, Jerusalem (Israel Program for Scientific Translations). *Q. J. R. Meteorol. Soc.,* **92**, 447–447, c.

Genell, A., Nemes, S., Steineck, G., Dickman, P.W., 2010. Model selection in Medical Research: A simulation study comparing Bayesian Model Averaging and Stepwise Regression. BMC Med. Res. Methodol. 10, 108. https://doi.org/10.1186/1471-2288-10-108

Georgakakos, Konstantine P, Seo, D.-J., Gupta, H., Schaake, J., Butts, M.B., 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

Georgakakos, Konstantine P., Seo, D.J., Gupta, H., Schaake, J., Butts, M.B., 2004. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, in: Journal of Hydrology. https://doi.org/10.1016/j.jhydrol.2004.03.037

Ghazvinian, M., Seo, D.-J., Zhang, Y., 2019. Improving Medium-range Probabilistic Quantitative Precipitation Forecast for Heavy-to-extreme Events through the Conditional Bias-penalized Regression, in: AGU Fall Meeting 2019. https://agu.confex.com/agu/fm19/meetingapp.cgi/Paper/517742.

Ghazvinian, M., Zhang, Y., Seo, D.-J., 2020. A Nonhomogeneous Regression-Based Statistical Postprocessing Scheme for Generating Probabilistic Quantitative Precipitation Forecast. J. Hydrometeorol. 21, 2275–2291. https://doi.org/10.1175/JHM-D-20-0019.1

Ghazvinian, M., Zhang, Y., Seo, D.-J., He, M., Fernando, N., 2021. A novel hybrid artificial neural network - Parametric scheme for postprocessing medium-range precipitation forecasts. Adv. Water Resour. 151, 103907. https://doi.org/10.1016/j.advwatres.2021.103907

Gochis, D.J., Barlage, M., Dugger, A., FitzGerald, K., Karsten, L., McAllister, M., McCreight, J., Mills, J., A. RafieeiNasab, L. Read, K. Sampson, D. Yates, W.Y., 2018. The WRF-Hydro modeling system technical description, (Version 5.0). NCAR Technical Note. 107 pages. Available online at: https://ral.ucar.edu/sites/default/files/public/WRFHydroV5TechnicalDescription.pdf.

Goel, P.K., Zellner, A., 1986. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti. North-Holland, Amsterdam.

Goldberger, A.S., 1984. Reverse Regression and Salary Discrimination. J. Hum. Resour. 19, 293–318. https://doi.org/10.2307/145875

Goodarzi, E., Ziaei, M., Hosseinipour, E.Z., 2014. Introduction to Optimization Analysis in Hydrosystem Engineering, Topics in Safety, Risk, Reliability and Quality. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-04400-2

Goudenhoofdt, E., and Delobbe, L., 2009: Evaluation of radar-gauge merging methods for quantitative precipitation estimates. *Hydrol. Earth Syst. Sci.*, **13**, 195-203, https://doi.org/10.5194/hess-13-195-2009.

Gourley, J. J., Hong, Y., Flamig, Z. L., Wang, J., Vergara, H., and Anagnostou, E. N., 2011:

Hydrologic Evaluation of Rainfall Estimates from Radar, Satellite, Gauge, and Combinations on Ft. Cobb Basin, Oklahoma. *J. Hydrometeor.*, **12**, 973–988, https://doi.org/10.1175/2011jhm1287.1.

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. J. Forecast. 3, 197–204. https://doi.org/10.1002/for.3980030207

Graziano, T., Clark, E., Cosgrove, B., Gochis, D., 2017. Transforming National Oceanic and Atmospheric Administration (NOAA) water resources prediction. 31st Conf. on Hydrology, Seattle, WA, Amer. Meteor. Soc., 2A.2, https://ams.confex.com/ams/97Annual/webprogram/Paper314016.html.

Green, C.A., Ferber, M.A., 1984. Employment Discrimination: An Empirical Test of Forward Versus Reverse Regression. J. Hum. Resour. 19, 557. https://doi.org/10.2307/145948

Guertin, K., 1984. Correcting Conditional Bias, in: Geostatistics for Natural Resources Characterization. Springer Netherlands, Dordrecht, pp. 245–260. https://doi.org/10.1007/978-94-009-3699-7_15

Habib, E., Haile, A. T., Tian, Y., and Joyce, R. J., 2012: Evaluation of the High-Resolution CMORPH Satellite Rainfall Product Using Dense Rain Gauge Observations and Radar-Based Estimates. *J. Hydrometeor.*, **13**, 1784–1798, https://doi.org/10.1175/jhm-d-12-017.1

Habib, E., Larson, B.F., Graschel, J., 2009. Validation of NEXRAD multisensor precipitation estimates using an experimental dense rain gauge network in south Louisiana. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2009.05.010

Habib, E., Qin, L., Seo, D.-J., Ciach, G. J., and Nelson, B. R., 2013: Independent Assessment of Incremental Complexity in NWS Multisensor Precipitation Estimator Algorithms. *J. Hydrol. Eng.*, **18**, 143–155, https://doi.org/10.1061/(asce)he.1943-5584.0000638

Hamill, T.M., Bates, G.T., Whitaker, J.S., Murray, D.R., Fiorino, M., Galarneau, T.J., Zhu, Y., Lapenta, W., 2013. NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. Bull. Am. Meteorol. Soc. 94, 1553–1565. https://doi.org/10.1175/BAMS-D-12-00014.1

Hassanzadeh, Y., Ghazvinian, M., Abdi, A., Baharvand, S., Jozaghi, A., 2020. Prediction of short and long-term droughts using artificial neural networks and hydro-meteorological variables. Atmos. Ocean. Phys. https://doi.org/arXiv:2006.02581v1 [physics.ao-ph] (available at https://arxiv.org/abs/2006.02581v1)

Hausman, J., 2001. Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. J. Econ. Perspect. 15, 57–67. https://doi.org/10.1257/jep.15.4.57

Hoerl, A.E., 1962. Application of Ridge Analysis to Regression Problems. Chem. Eng. Prog. 58, 54–59.

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634

Hoeting, Jennifer A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: A tutorial. Stat. Sci. https://doi.org/10.1214/ss/1009212519

Hoeting, Jennifer A, Madigan, D., Raftery, A.E., Volinsky, C.T., Raferty, A.E., Volinshy, C.T., 1999. Bayesian Model Averaging: A Tutorial (with discussion). Stat. Sci.

Huffman, G. J., D. T. Bolvin, and E. J. Nelkin, 2017: Integrated Multi-satellitE Retrievals for GPM (IMERG) Technical Documentation, Technical Documentation, Lab. for Atmos., NASA Goddard Space Flight Cent. and Sci. Syst. and Appl. Inc, [Available at: https://pmm.nasa.gov/sites/default/files/document_files/IMERG_doc.pdf]

Huffman, G.J., D.T. Bolvin, E.J. Nelkin, D.B. Wolff, R.F. Adler, G. Gu, Y. Hong, K.P. Bowman, and E.F. Stocker, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *J. Hydrometeor*, **8**, 38–55, https://doi.org/10.1175/JHM560.1.

Hughes, M.D., 1993. Regression Dilution in the Proportional Hazards Model. Biometrics 49, 1056. https://doi.org/10.2307/2532247

ICHARM, 2009. Global trends in water related disasters: an insight for policymakers. International Centre for Water Hazard and Risk Management (UNESCO), Tsukuba, Japan, http://www. icharm.pwri.go.jp.

Jahan, A., Edwards, K.L., Bahraminasab, M., 2016. Multi-criteria Decision Analysis for Supporting the Selection of Engineering Materials in Product Design. Butterworth-Heinemann Elsevier Ltd, Oxford.

Joliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley, 254 pp.

Jolliffe, I.T., Stephenson, D.B. (Eds.), 2011. Forecast Verification. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9781119960003

Journel, A. G., and Ch. J. Huijbregts, 1978: *Mining geostatistics*. Academic Press, 600pp.

Joyce, R. J., Janowiak, J. E., Arkin, P. A., and Xie, P., 2004: CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution. *J. Hydrometeor.*, **5**, 487–503, https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2.

Jozaghi, A., Alizadeh, B., Hatami, M., Flood, I., Khorrami, M., Khodaei, N., Ghasemi Tousi, E., 2018. A Comparative Study of the AHP and TOPSIS Techniques for Dam Site Selection

Using GIS: A Case Study of Sistan and Baluchestan Province, Iran. Geosciences 8, 494. https://doi.org/10.3390/geosciences8120494

Jozaghi, A., Ghazvinian, M., Seo, D., Zhang, Y., Wells, E., Reed, S., Fresch, M., 2020. Improving Water Forecasting with Bayesian Averaging of Multiple forecasts. 34th Conference on Hydrology, Boston, MA, Amer. Meteor. Soc., J26.3, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/369034.

Jozaghi, A., Nabatian, M., Noh, S., Seo, D.-J., Tang, L., Zhang, J., 2019. Improving Multisensor Precipitation Estimation via Adaptive Conditional Bias–Penalized Merging of Rain Gauge Data and Remotely Sensed Quantitative Precipitation Estimates. J. Hydrometeorol. 20, 2347–2365. https://doi.org/10.1175/JHM-D-19-0129.1

Jozaghi, A., Shamsai, A., 2017. Application of Geospatial Information System and Technique for Order Preference by Similarity to Ideal Solution for sitting water reservoirs Case study: South of Sistan&Balouchestan Province. Sci. Res. Q. Geogr. Data 25(100), pp5-15. https://doi.org/10.22131/sepehr.2017.24802

Jozaghi, A., Shen, H., Ghazvinian, M., Seo, D.-J., Zhang, Y., Welles, E., Reed, S., 2021a. Multi-Model Streamflow Prediction Using Conditional Bias-Penalized Multiple Linear Regression. Submitted to Stoch. Environ. Res. Risk Assess.

Jozaghi, A., Shen, H., Seo, D., Zhang, Y., Welles, E., Reed, S., 2021b. Improving Flood Forecasting via Conditional Bias-Penalized Bayesian Model Averaging. Amer. Meteor. Soc., 35th Conference on Hydrology, New Orleans, LA.

Kalinga, O. A., and T. Y. Gan, 2010: Estimation of rainfall from infrared-microwave satellite data for basin-scale hydrologic modelling. *Hydrol. Processes,* **24**, 2068–2086, https://doi.org/10.1002/hyp.7626.

Kasiviswanathan, K.S., Sudheer, K.P., 2017. Methods used for quantifying the prediction uncertainty of artificial neural network based hydrologic models. Stoch. Environ. Res. Risk Assess. https://doi.org/10.1007/s00477-016-1369-5

Kelly, K.S., Krzysztofowicz, R., 1997. A bivariate meta-Gaussian density for use in hydrology. Stoch. Hydrol. Hydraul. 11, 17–31. https://doi.org/10.1007/BF02428423

Keum, J., Coulibaly, P., 2017. Information theory-based decision support system for integrated design of multivariable hydrometric networks. Water Resour. Res. 53, 6239–6259. https://doi.org/10.1002/2016WR019981

Khan, U.T., Valeo, C., He, J., 2013. Non-linear fuzzy-set based uncertainty propagation for improved DO prediction using multiple-linear regression. Stoch. Environ. Res. Risk Assess. 27, 599–616. https://doi.org/10.1007/s00477-012-0626-5

Khanmohammadi, N., Rezaie, H., Montaseri, M., Behmanesh, J., 2018. The application of multiple linear regression method in reference evapotranspiration trend calculation. Stoch. Environ. Res. Risk Assess. 32, 661–673. https://doi.org/10.1007/s00477-017-1378-z

Kim, B., Seo, D.-J., Noh, S.J., Prat, O.P., Nelson, B.R., 2018. Improving multisensor estimation of heavy-to-extreme precipitation via conditional bias-penalized optimal estimation. J. Hydrol. 556, 1096–1109. https://doi.org/10.1016/j.jhydrol.2016.10.052

Kim, D., Nelson, B., and Seo, D.-J., 2009: Characteristics of Reprocessed Hydrometeorological Automated Data System (HADS) Hourly Precipitation Data. *Wea. Forecasting*, **24**, 1287–1296, https://doi.org/10.1175/2009waf2222227.1.

Kitzmiller, D., Miller, D., Fulton, R., Ding, F., 2013. Radar and multisensor precipitation estimation techniques in national weather service hydrologic operations. J. Hydrol. Eng. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000523

Kondragunta, C., D. Kitzmiller, D.-J. Seo, and K. Shrestha, 2005: Objective Integration of Satellite, Rain Gauge, and Radar Precipitation Estimates in the Multisensor Precipitation Estimator Algorithm. *19th Conf. on Hydrology, AMS Annual Meeting,* 2.8.

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

Kuligowski, R. J., 2002: A Self-Calibrating Real-Time GOES Rainfall Algorithm for Short-Term Rainfall Estimates. *J. Hydrometeor.*, **3**, 112-130, https://doi.org/10.1175/1525-7541(2002)003<0112:ascrtg>2.0.co;2.

Kuligowski, R. J., Li, Y., and Zhang, Y., 2013: Impact of TRMM data on a low-latency, high-resolution precipitation algorithm for flash-flood forecasting. *J. Appl. Meteor. Climatol.*, **52**, 1379- 1393, https://doi.org/10.1175/JAMC-D-12-0107.1.

Kuligowski, R. J., 2010: *GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document for Rainfall Rate (QPE)*. NOAA/NESDIS/STAR Algorithm Theoretical Basis Doc., version 2.0, 44 pp. [Available online at http://www.goes-r.gov/products/ATBDs/baseline/Hydro_RRQPE_v2.0_no_color.pdf.]

Lashkarara, B., Baharvand, S., Najafi, L., 2021. Study the Performance of Data-driven Models to Predict the Scour Depth caused by the Aerated Vertical Jet. Irrig. Sci. Eng. 0(4), 79–89. https://doi.org/10.22055/JISE.2021.36599.1959

Lee, H., Shen, H., Noh, S.J., Kim, S., Seo, D.-J., Zhang, Y., 2019. Improving flood forecasting using conditional bias-penalized ensemble Kalman filter. J. Hydrol. 575, 596–611. https://doi.org/10.1016/j.jhydrol.2019.05.072

Levi, M.D., 1973. Errors in the Variables Bias in the Presence of Correctly Measured Variables. Econometrica. https://doi.org/10.2307/1913819

Liu, J., Yuan, D., Zhang, L., Zou, X., Song, X., 2016. Comparison of Three Statistical Downscaling Methods and Ensemble Downscaling Method Based on Bayesian Model Averaging in Upper Hanjiang River Basin, China. Adv. Meteorol. 2016, 1–12. https://doi.org/10.1155/2016/7463963

Liu, Y., Ji, Y., Liu, D., Fu, Q., Li, T., Hou, R., Li, Q., Cui, S., Li, M., 2021. A new method for runoff prediction error correction based on LS-SVM and a 4D copula joint distribution. J. Hydrol. 598, 126223. https://doi.org/10.1016/j.jhydrol.2021.126223

Liu, Y., Ye, L., Qin, H., Hong, X., Ye, J., Yin, X., 2018. Monthly streamflow forecasting based on hidden Markov model and Gaussian Mixture Regression. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2018.03.057

Liu, Y.R., Li, Y.P., Huang, G.H., Zhang, J.L., Fan, Y.R., 2017. A Bayesian-based multilevel factorial analysis method for analyzing parameter uncertainty of hydrological model. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2017.08.048

Lu, X., Ju, Y., Wu, L., Fan, J., Zhang, F., Li, Z., 2018. Daily pan evaporation modeling from local and cross-station data using three tree-based machine learning models. J. Hydrol. 566, 668–684. https://doi.org/10.1016/j.jhydrol.2018.09.055

Luo, X., Yuan, X., Zhu, S., Xu, Z., Meng, L., Peng, J., 2019. A hybrid support vector regression framework for streamflow forecast. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2018.10.064

Madadgar, S., Moradkhani, H., 2014. Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. Water Resour. Res. https://doi.org/10.1002/2014WR015965

Maddox, R. A., Zhang, J., Gourley, J. J., and Howard, K. W., 2002: Weather Radar Coverage over the Contiguous United States. *Wea. Forecasting*, **17**, 927–934, https://doi.org/10.1175/1520-0434(2002)017<0927:wrcotc>2.0.co;2.

Madigan, D., Raftery, A.E., 1994. Model selection and accounting for model uncertainty in graphical models using occam’s window. J. Am. Stat. Assoc. https://doi.org/10.1080/01621459.1994.10476894

Makkonen, L., 2006. Plotting positions in extreme value analysis. J. Appl. Meteorol. Climatol. https://doi.org/10.1175/JAM2349.1

Meng, E., Huang, S., Huang, Q., Fang, W., Wu, L., Wang, L., 2019. A robust method for non-stationary streamflow prediction based on improved EMD-SVM model. J. Hydrol. 568, 462–478. https://doi.org/10.1016/j.jhydrol.2018.11.015

Michaels, S., 2015. Probabilistic forecasting and the reshaping of flood risk management. J. Nat. Resour. Policy Res. https://doi.org/10.1080/19390459.2014.970800

Muhammad, A., Stadnyk, T.A., Unduche, F., Coulibaly, P., 2018. Multi-model approaches for improving seasonal ensemble streamflow prediction scheme with various statistical post-processing techniques in the Canadian Prairie Region. Water (Switzerland). https://doi.org/10.3390/w10111604

Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Mon. Weather Rev. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

Napolitano, G., Serinaldi, F., See, L., 2011. Impact of EMD decomposition and random initialisation of weights in ANN hindcasting of daily stream flow series: An empirical examination. J. Hydrol. 406, 199–214. https://doi.org/10.1016/j.jhydrol.2011.06.015

National Weather Service, 2016. Meteorological Ensemble Forecast Processor (MEFP) User’s Manual. Office of Hydrologic Development, Silver Spring, MD. https://www.nws.noaa.gov/oh/hrl/general/HEFS_doc/MEFP_Users_Manual.pdf.

National Weather Service, 2015. Ensemble Postprocessor (EnsPost) User’s Manual. Office of

Hydrologic Development, Silver Spring, MD, https://www.nws.noaa.gov/oh/hrl/general/HEFS_doc/EnsPost_Users_Manual.pdf. Accessed 04 April, 2015.

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

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

National Weather Service, 2006. NWSRFS User Manual Documentation. Office of Hydrologic Development, Silver Spring, MD. https://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/htm/warnpdf.php.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and Implications of Stage IV Quantitative Precipitation Estimates for Product Intercomparisons. *Wea. Forecasting*, **31**, 371-394, DOI: http://dx.doi.org/10.1175/WAF-D-14-00112.1.

Nelson, B.R., Seo, D.-J., Kim, D., 2010. Multisensor Precipitation Reanalysis. J. Hydrometeorol. 11, 666–682. https://doi.org/10.1175/2010JHM1210.1

Noji, E.., Lee, C.., 2005. Disaster preparedness. In H. Frumkin, ed. Environmental health: from global to local. 1st ed. San Francisco, CA: Jossey-Bass Publishers. 745-780.

Okamoto, K., Ushio, T., Iguchi, T., Takahashi, N., and Iwanami, K., 2005: The Global Satellite Mapping of Precipitation (GSMaP) project. *International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3414-3416, https://doi.org/10.1109/IGARSS.2005.1526575.

Olea, R.A., 2012. Building on crossvalidation for increasing the quality of geostatistical modeling. Stoch. Environ. Res. Risk Assess. 26, 73–82. https://doi.org/10.1007/s00477-011-0496-2

Pan, G., 1998. Smoothing effect, conditional bias and recoverable reserves. Can Inst Min Met. Bull 91(1019), 81–86.

Policelli, F., Hubbard, A., Jung, H.C., Zaitchik, B., Ichoku, C., 2019. A predictive model for Lake Chad total surface water area using remotely sensed and modeled hydrological and meteorological parameters and multivariate regression analysis. J. Hydrol. 568, 1071–1080. https://doi.org/10.1016/j.jhydrol.2018.11.037

Prat, O. P., and B. R. Nelson, 2013: Precipitation contribution of tropical cyclones in the southeastern United States from 1998 to 2009 using TRMM satellite data. *J. Climate*, **26**, 1047–1062, https://doi.org/10.1175/JCLI-D-11-00736.1.

Prat, O. P., B. R. Nelson, S. E. Stevens, E. N., D.-J. Seo, B. Kim, J. Zhang, and Y. Qi, 2015: Merging Radar Quantitative Precipitation Estimates (QPEs) from the High-resolution NEXRAD Reanalysis over CONUS with Rain-gauge Observations. *2015 AGU Fall Meeting*, H24E-04, December 14-18 2015, San Francisco, CA.

Prat, O. P., B. R. Nelson, S. E. Stevens, and D.-J. Seo, 2014: Long-term large-scale bias-adjusted precipitation estimates at high spatial and temporal resolution derived from the National Mosaic and Multi-sensor QPE (NMQ/Q2) precipitation reanalysis over CONUS. *ERAD 2014, AGU Fall Meeting*, A11A-3007, December 15-19 2014, San Francisco, CA.

Rafieeinasab, A., Norouzi, A., Seo, D.-J., Nelson, B., 2015. Improving high-resolution quantitative precipitation estimation via fusion of multiple radar-based precipitation products. J. Hydrol. 531, 320–336. https://doi.org/10.1016/j.jhydrol.2015.04.066

Raftery, A.E., 1995. Bayesian Model Selection in Social Research. Sociol. Methodol. https://doi.org/10.2307/271063

Raftery, A.E., 1993. Bayesian model selection in structural equation models. Test. Struct. Equ. Model.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. https://doi.org/10.1175/MWR2906.1

Raftery, A.E., Hoeting, J., Volinsky, C.T., Painter, I., Yeung, K.Y., 2015. BMA: Bayesina model averaging [WWW Document]. http://cran.r-project.org/package=BMA. https://doi.org/10.1080/02664760902899774

Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian Model Averaging for Linear Regression Models. J. Am. Stat. Assoc. https://doi.org/10.1080/01621459.1997.10473615

Rahman, A.S., Khan, Z., Rahman, A., 2020. Application of independent component analysis in regional flood frequency analysis: Comparison between quantile regression and parameter regression techniques. J. Hydrol. 581, 124372. https://doi.org/10.1016/j.jhydrol.2019.124372

Rajagopalan, B., Lall, U., Zebiak, S.E., 2002. Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. Mon. Weather Rev. https://doi.org/10.1175/1520-0493(2002)130<1792:CCFTRA>2.0.CO;2

Regonda, S.K., Rajagopalan, B., Clark, M., Zagona, E., 2006. A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin. Water Resour. Res. https://doi.org/10.1029/2005WR004653

Regonda, S.K., Seo, D.J., Lawrence, B., Brown, J.D., Demargne, J., 2013. Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts - A Hydrologic Model Output Statistics (HMOS) approach. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2013.05.028

Roe, J., Dietz, C., Restrepo, P. et al, 2010. NOAA's Community Hydrologic Prediction System. 2th Federal Interagency Hydrologic Modeling Conference, Las Vegas, NV. https://acwi.gov/sos/pubs/2ndJFIC/Contents/7E_Roe_12_28_09.pdf.

Sahoo, G.B., Ray, C., De Carlo, E.H., 2006. Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. J. Hydrol. 327, 525–538. https://doi.org/10.1016/j.jhydrol.2005.11.059

Sahoo, G.B., Schladow, S.G., Reuter, J.E., 2009. Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. J. Hydrol. 378, 325–342. https://doi.org/10.1016/j.jhydrol.2009.09.037

Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., Seo, D.J., 2007. Precipitation and temperature ensemble forecasts from single-value forecasts. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hessd-4-655-2007

Schoppa, L., Disse, M., Bachmair, S., 2020. Evaluating the performance of random forest for large-scale flood discharge simulation. J. Hydrol. 590, 125531. https://doi.org/10.1016/j.jhydrol.2020.125531

Schweppe, F.C., 1973. Uncertain dynamic systems. Englewood Cliffs, N.J., Prentice Hall.

Seber, G.A.F., Wild, C.J., 1989. Nonlinear Regression, Wiley Series in Probability and Statistics.

John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/0471725315

Seo, D.-J., 2006. A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hessd-3-1987-2006

Seo, D.-J., 1996: Nonlinear estimation of spatial distribution of rainfall - An indicator cokriging approach. *Stoch. Hydrol. Hydraul.*, **10**, 127, https://doi.org/10.1007/BF01581763.

Seo, D.-J., 1998a: Real-time estimation of rainfall fields using radar rainfall and rain gage data. *J. Hydrol*., **208**, 37–52, https://doi.org/10.1016/S0022-1694(98)00141-3.

Seo, D.-J., 1998b: Real-time estimation of rainfall fields using rain gage data under fractional coverage conditions. *J. Hydrol*., **208**, 25–36, https://doi.org/10.1016/S0022-1694(98)00140-1.

Seo, D.-J., 2013: Conditional bias-penalized kriging (CBPK). *Stoch. Environ. Res. Risk Assess.*, **27**, 43-58, https://doi.org/10.1007/s00477-012-0567-z.

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

Seo, D.-J., Breidenbach, J. P., and Johnson, E. R., 1999: Real-time estimation of mean field bias in radar rainfall data. *J. Hydrol.*, **223**, 131-147, https://doi.org/10.1016/S0022-1694(99)00106-7

Seo, D.-J., Breidenbach, J., Fulton, R., Miller, D., and O’Bannon, T., 2000: Real-Time Adjustment of Range-Dependent Biases in WSR-88D Rainfall Estimates due to Nonuniform Vertical Profile of Reflectivity. *J. Hydrometeor.*, **1**, 222-240. https://doi.org/10.1175/1525-7541(2000)001<0222:rtaord>2.0.co;2.

Seo, D.-J., Demargne, J., Wu, L., Liu, Y., Brown, J.D., Regonda, S., Lee, H., 2010. Hydrologic ensemble prediction for risk-based water resources management and hazard mitigation. Joint Federal Interagency Conf. on Sedimentation and Hydrologic Modeling, Las Vegas, NV, NOAA, 26 pp. [Available online at www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/publications_presentations/Seo_et_al_JFIC_June2010.pdf.].

Seo, D.-J., Herr, H.D., Schaake, J.C., 2006. A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hessd-3-1987-2006

Seo, D.-J., Perica, S., Welles, E., and Schaake, J. C., 2000: Simulation of precipitation fields from probabilistic quantitative precipitation forecast. *J. Hydrol.*, **239**, 203-229, https://doi.org/10.1016/S0022-1694(00)00345-0.

Seo, D.-J., Saifuddin, M.M., Lee, H., 2018a. Conditional bias-penalized Kalman filter for improved estimation and prediction of extremes. Stoch. Environ. Res. Risk Assess. 32, 183–201. https://doi.org/10.1007/s00477-017-1442-8

Seo, D.-J., Saifuddin, M.M., Lee, H., 2018b. Correction to: Conditional bias-penalized Kalman filter for improved estimation and prediction of extremes. Stoch. Environ. Res. Risk Assess. https://doi.org/10.1007/s00477-018-1626-x

Seo, D.-J., Saifuddin, M.M., Lee, H., 2018c. Correction to: Conditional bias-penalized Kalman filter for improved estimation and prediction of extremes. Stoch. Environ. Res. Risk Assess. 32, 3561–3562. https://doi.org/10.1007/s00477-018-1626-x

Seo, D.-J., Seed, A., Delrieu, G., 2010: Radar and multisensor rainfall estimation for hydrologic applications, In Rainfall: State of the Science, Geophysical Monograph Series, F. Y. Testik, and M. Gebremichael, eds., **191**, 79-104, AGU, Washington, D. C, https://doi.org/10.1029/2010GM000952.

Seo, D.-J., Siddique, R., Zhang, Y., Kim, D., 2014. Improving real-time estimation of heavy-to-extreme precipitation using rain gauge data via conditional bias-penalized optimal estimation. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2014.09.055

Seo, D.-J., Smith, J. A., 1996: On the Relationship Between Catchment Scale and Climatological Variability of Surface-Runoff Volume. *Water Resour. Res.*, **32**, 633– 643, https://doi.org/10.1029/95WR03641.

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, C.E., Weaver, W., 1949. The mathematical theory of communication (Urbana, IL, University of Illinois Press.

Shen, H., Lee, H., Seo, D.-J., 2019. Adaptive Conditional Bias-Penalized Kalman Filter for Improved Estimation of Extremes and its Approximation for Reduced Computation. Signal Processing. https://doi.org/arXiv:1908.00482 [eess.SP]

Shen, H., Seo, D.J., Lee, H., Liu, Y., Noh, S., 2020. Improving Flood Forecasting Using Conditional Bias-Aware Assimilation of Streamflow Observations and Dynamic Assessment of Flow-Dependent Information Content. Submitt. to J. Hydrol.

Shih, S.F., Shih, W.F.P., 1979. Application of ridge regression analysis to water resources studies. J. Hydrol. 40, 165–174. https://doi.org/10.1016/0022-1694(79)90095-7

Siddique, R., Mejia, A., Brown, J., Reed, S., Ahnert, P., 2015. Verification of precipitation forecasts from two numerical weather prediction models in the Middle Atlantic Region of the USA: A precursory analysis to hydrologic forecasting. J. Hydrol.

https://doi.org/10.1016/j.jhydrol.2015.08.042

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

Smith, J. A., W. F. Krajewski, 1992: Estimation of the mean-field bias of radar rainfall estimates. *J. Appl. Meteor.*, **30**, 397–412, https://doi.org/10.1175/1520-0450(1991)030<0397:EOTMFB>2.0.CO;2.

Smith, J. A., D.-J. Seo, M. L. Baeck, and M. D. Hudlow, 1996: An intercomparison study of NEXRAD precipitation estimates. *Water Resour. Res.*, **32**, 2035–2045, https://doi.org/10.1029/96WR00270.

Smith, R., 2015. A mutual information approach to calculating nonlinearity. Stat 4, 291– 303. https://doi.org/10.1002/sta4.96

Smith, T. M., P. A. Arkin, J. J. Bates, and G. J. Huffman, 2006. Estimating Bias of Satellite-Based Precipitation Estimates. *J. Hydrometeor.*, **7**, 841–856, https://doi.org/10.1175/JHM524.1.

Sorooshian, S., Hsu, K. L., Gao, X., Gupta, H. V., Imam, B., and Braithwaite, D., 2000: Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bull. Amer. Meteor. Soc.*, **81**, 2035–2046, https://doi.org/10.1175/1520-0477 (2000)081<2035:EOPSSE>2.3.CO;2.

Supriya, P., Krishnaveni, M., Subbulakshmi, M., 2015. Regression Analysis of Annual Maximum Daily Rainfall and Stream Flow for Flood Forecasting in Vellar River Basin. Aquat. Procedia 4, 957–963. https://doi.org/10.1016/j.aqpro.2015.02.120

Tang, L., D.-J. Seo, M. Nabatian, J. Zhang, K. W. Howard, and D. Kitzmiller, 2019: Comparative Evaluation of Merging and Local Bias Correction for Radar-Gauge QPE. *AMS*

Annual Meeting, Phoenix, AZ.

Tang, G., Long, D., Behrangi, A., Wang, C., Hong, Y., 2018. Exploring Deep Neural Networks to Retrieve Rain and Snow in High Latitudes Using Multisensor and Reanalysis Data. Water Resour. Res. https://doi.org/10.1029/2018WR023830

Theil, H., 1961. Economic forecasts and policy. North-Holland Pub. Co., Amsterdam.

Tibshirani, R., 1997. The lasso method for variable selection in the cox model. Stat. Med. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tikhonov, A.N., Arsenin, V.Y., 1977. Solutions of Ill-Posed Problems. Winston and Sons, Washington DC.

Tikhonov, A.N., Goncharsky, A. V., Stepanov, V. V., Yagola, A.G., 1995. Numerical Methods for the Solution of Ill-Posed Problems, Numerical Methods for the Solution of Ill-Posed Problems. https://doi.org/10.1007/978-94-015-8480-7

Tiwari, M.K., Adamowski, J., 2013. Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. Water Resour. Res. https://doi.org/10.1002/wrcr.20517

Toth, Z., Kalnay, E., 1997. Ensemble forecasting at NCEP and the breeding method. Mon. Weather Rev. https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2

Turk, F. J., and Miller, S. D., 2005: Toward improved characterization of remotely sensed precipitation regimes with MODIS/AMSR-E blended data techniques. *IEEE Transactions on Geoscience and Remote Sensing,* **43**, 1059-1069, https://doi.org/10.1109/TGRS.2004.841627.

Um, M.-J., Yun, H., Jeong, C.-S., Heo, J.-H., 2011. Factor analysis and multiple regression between topography and precipitation on Jeju Island, Korea. J. Hydrol. 410, 189–203. https://doi.org/10.1016/j.jhydrol.2011.09.016

Vanmarcke, E., 1983: *Random fields: Analysis and synthesis*. ISBN 0-262-72045-0. Cambridge, Massachusetts, USA: The MIT Press, pp. 382.

Vasiloff, S. V., D. Seo, K. W. Howard, J. Zhang, D. H. Kitzmiller, M. G. Mullusky, W. F. Krajewski, E. A. Brandes, R. M. Rabin, D. S. Berkowitz, H.E. Brooks, J. A. McGinley, R. J. Kuligowski, and B. G. Brown, 2007: Improving QPE and Very Short Term QPF: An Initiative for a Community-Wide Integrated Approach. *Bull. Amer. Meteor. Soc.*, **88**, 1899–1911, https://doi.org/10.1175/BAMS-88-12-1899.

Viallefont, V., Raftery, A.E., Richardson, S., 2001. Variable selection and Bayesian model averaging in case-control studies. Stat. Med. 20, 3215–3230. https://doi.org/10.1002/sim.976

Vicente, G. A., Scofield, R. A., and Menzel, W. P., 1998: The Operational GOES Infrared Rainfall Estimation Technique. *Bull. Amer. Meteor. Soc.*, **79**, 1883- 1898, https://doi.org/10.1175/1520-0477(1998)079<1883:TOGIRE>2.0.CO;2.

Vogel, R.M., Yaindl, C., Walter, M., 2011. Nonstationarity: Flood Magnification and Recurrence Reduction Factors in the United States1. JAWRA J. Am. Water Resour. Assoc. 47, 464–474. https://doi.org/10.1111/j.1752-1688.2011.00541.x

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. https://doi.org/10.1029/2005WR004838

Wang, D., Zhang, W., Bakhai, A., 2004. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. Stat. Med. 23, 3451–3467. https://doi.org/10.1002/sim.1930

Wang, H., Wang, C., Wang, Y., Gao, X., Yu, C., 2017. Bayesian forecasting and uncertainty quantifying of stream flows using Metropolis–Hastings Markov Chain Monte Carlo algorithm. J. Hydrol. 549, 476–483. https://doi.org/10.1016/j.jhydrol.2017.03.073

Wilks, D.S., 2006. Statistical Methods in the Atmospheric Sciences. 2nd Edition. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Wilks, D.S., 2005. Statistical Methods in the Atmospheric Sciences: 2nd Edition. Academic Press, London, UK. https://doi.org/10.1002/met.16

Wu, L., Seo, D.J., Demargne, J., Brown, J.D., Cong, S., Schaake, J., 2011. Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2011.01.013

Xu, J., Anctil, F., Boucher, M.A., 2019. Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2019.124002

Yaseen, Z.M., El-Shafie, A., Afan, H.A., Hameed, M., Mohtar, W.H.M.W., Hussain, A., 2016. RBFNN versus FFNN for daily river flow forecasting at Johor River, Malaysia. Neural Comput. Appl. 27, 1533–1542. https://doi.org/10.1007/s00521-015-1952-6

Young, C.B., Bradley, A.A., Krajewski, W.F., Kruger, A., Morrissey, M.L., 2000. Evaluating NEXRAD multisensor precipitation estimates for operational hydrologic forecasting. J. Hydrometeorol. https://doi.org/10.1175/1525-7541(2000)001<0241:ENMPEF>2.0.CO;2

Yun, W.T., Stefanova, L., Mitra, A.K., Vijaya Kumar, T.S.V., Dewar, W., Krishnamurti, T.N., 2005. A multi-model superensemble algorithm for seasonal climate prediction using

DEMETER forecasts. Tellus, Ser. A Dyn. Meteorol. Oceanogr. https://doi.org/10.1111/j.1600-0870.2005.00131.x

Zellner, A., 1986. On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions". In Goel, P.; Zellner, A. (eds.). Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, Studies in Bayesian Econometrics and Statistics. New York: Elsevier. pp. 233–243. ISBN 978-0-444-87712-3.

Zeugner, S., 2011. Bayesian Model Averaging with BMS for BMS Version 0.3.0, https://cran.r-project.org/web/packages/BMS/index.html

Zhang, J, K. Howard, C. Langston, B. Kaney, Y. Qi, L. Tang, H. Grams, Y. Wang, S. Cocks, S. Martinaitis, A. Arthur, K. Cooper, J. Brogden, and D. Kitzmiller, 2016: Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bull. Amer. Meteor. Soc.,* **97**, 621-638, http://dx.doi.org/10.1175/BAMS-D-14-00174.1.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) System: Description, results, and future plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338, https://doi.org/10.1175/2011BAMS-D-11-00047.1

Zhang, Y., Wu L., Scheuerer M., Schaake J., Kongoli C., 2017. Comparison of Probabilistic Quantitative Precipitation Forecasts from Two Postprocessing Mechanisms. J. Hydrometeor., 18, 2873–2891, https://doi.org/10.1175/JHM-D-16-0293.1.

Zhu, Y., Toth, Z., 2008. NAEFS and NCEP Global Ensemble. Presentation for National DOH Workshop, https://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/events_announce/NAEFS_Yuejian_Zhu.pdf.