Final Report on

Towards Objective Multi-Modeling for Multi-Institutional Seasonal Water Supply Forecasting

To the

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

Award Number: NA11NWS4680002

Principle Investigator: Hamid Moradkhani

Remote Sensing and Water Resources Lab
Department of Civil and Environmental Engineering
Portland State University

June 29, 2016

A. Project Summary

Seasonal water supply outlooks, or volume of total seasonal runoff, are routinely used by decision makers in the western United States for making commitments for water deliveries, determining industrial and agriculture water allocation, and operating reservoirs. These forecasts are based primarily on statistical regression equations developed from monthly precipitation, recent snow-water equivalent, and a subset of past streamflow observations. In the Western US, the National Weather Service Northwest River Forecast Center (NWRFC) and the Natural Resources Conservation Service (NRCS) jointly issue seasonal water supply outlook forecasts of naturalized or unimpaired flow, i.e. the flow that would most likely occur in the absence of diversions. This is done using statistical and ensemble streamflow prediction (ESP) methods developed at the NWS and NRCS. In addition, water resources management entities including US Bureau of Reclamation (BOR) and Corps of Engineers (COE) use their own statistical models to issue seasonal water supply forecast.

The operational streamflow forecasting community understandably places a priority on operations rather than research and often lacks the resources to investigate new products or methods during the demands of a busy forecasting season. There is a consensus that further improvement in the forecast of atmospheric forcing as well as use of suite of hydrologic models including statistical and dynamical (i.e., physical or conceptual) models is needed for improving streamflow forecast skill at both short- and long- lead time scales. Therefore, the proposed research has an overall goal to incorporate the latest scientific findings in the area of multi-modeling that optimally combine the multi-model ensemble hydrologic forecasts. Given that a multi-model contains information from all participating models, including the less skillful ones, the question that remains is: under what conditions, a multi-model can outperform the best participating single model? In this project we make an attempt in carefully testing under what circumstances multi-model combination reduces overconfidence, i.e. ensemble spread is widened while average ensemble-mean error is reduced. This implies a net gain in prediction skill, because probabilistic skill scores penalize overconfidence.

The report is organized by highlighting major findings as published in refereed journals.

B. Towards Improved Reliability and Reduced Uncertainty of Hydrologic Ensemble Forecasts Using Multivariate Post-processing (Madadgar et al. 2014)

In this study, we addressed the drawbacks of a commonly used statistical technique, Quantile Mapping (QM), in bias correction of hydrologic forecasts. An alternative postprocessor is then introduced such that marginal distributions of observations and model simulations are combined to create a multivariate joint distribution using multivariable probability functions, the so-called copula functions. In addition to hypothetical cases, post-processing of a real case study was also tested, using a distributed parameter hydrologic model, the Precipitation Runoff Modeling System (PRMS). Several ensembles of monthly streamflow forecasts of the Sprague River basin in southern Oregon were generated with a forecast horizon of 6 months. An auxiliary index, the so called failure index (γ), was introduced to predict the overall performance of the QM technique as an ensemble post-processing method before stepping into the forecast mode. The failure index reflects the consistency of QM adjustments and corresponding observations; it varies between 0 and 1, with $\gamma = 0$ being the perfect-adjustment case. The forecast skill of QM shows that this statistical bias correction technique is not always successful in improving initial forecast trajectories. Testing 2500 hypothetical case studies indicates that the performance of the QM technique constantly degrades as γ increases. Generally, the forecast skill of the post-processed ensembles effectively improved when the multivariate postprocessor was applied, but it became even worse when QM technique was used (Figure 1). Overall, Figure 2 demonstrates that the QM method is not an effective method to adjust the original forecasts while the multivariate copula-based postprocessor is a more effective method that can be used operationally.

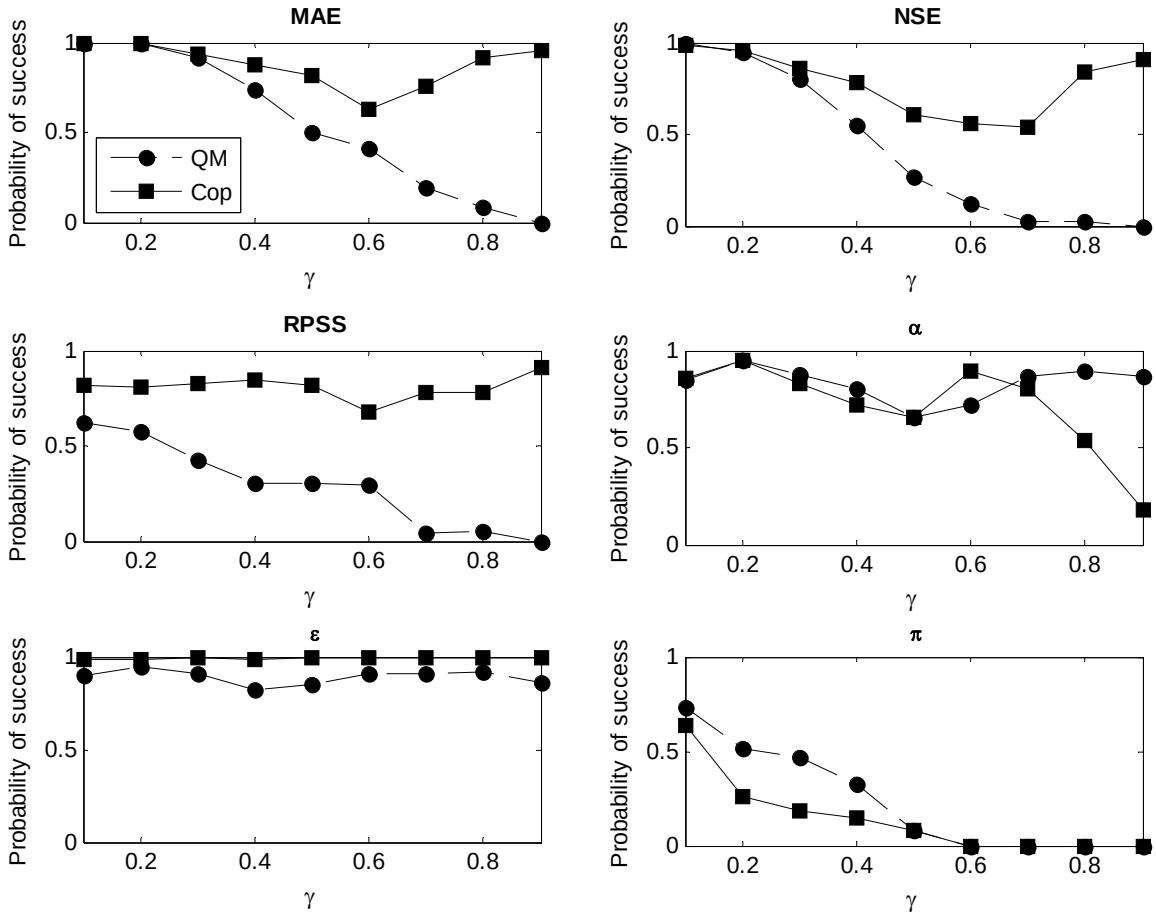

Figure 1. Probability of success against γ for point-wise (MAE and NSE) and probabilistic performance measures (RPSS, reliability (α and ε), Resolution (π)) in QM and copula-based post-processing methods. Probability of success is obtained with respect to the associated metric for different values of the failure index.

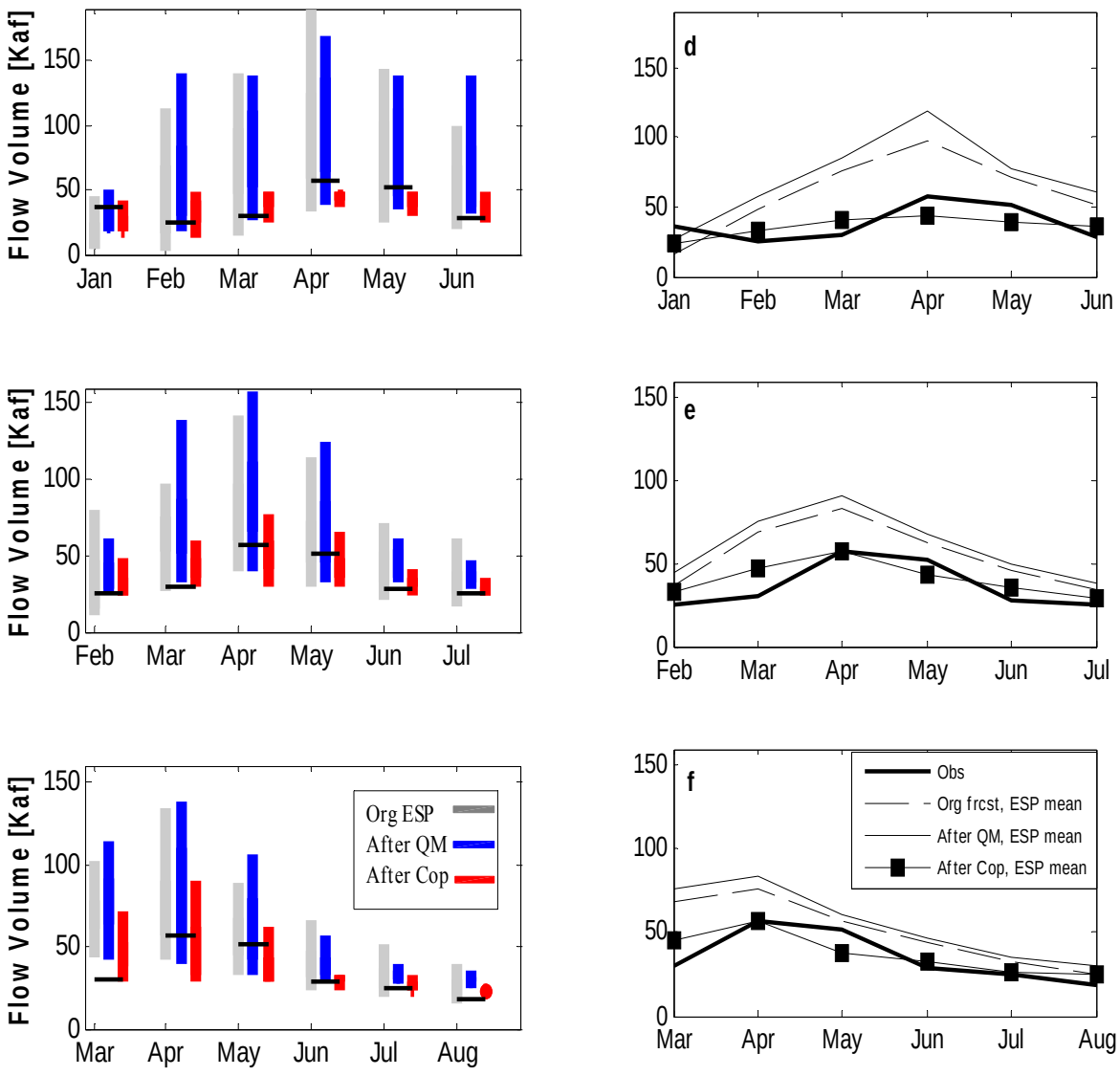

Figure 2. Comparison of the ensemble range before and after post-processing for three forecast periods in 2002 starting from a) Jan, b) Feb, and c) Mar, with the solid lines representing the monthly observations. Corresponding ESP mean are shown in subplots d-f.

C. Post-processing hydrologic model outputs to improve the objectivity of streamflow forecasts; case study of Columbia River Basin

This is an extension of the above postprocessing method to multitude of gage stations in the Pacific Northwest and validating its application toward improving accuracy of streamflow simulations. The procedure is performed for historical period of 1970–1999. Three semi-distributed hydrologic models, i.e. Variable Infiltration Capacity (VIC),

SAC-SMA, and Precipitation Runoff Modelling System (PRMS), are employed and calibrated at 1/16 degree latitude-longitude resolution for more than 100 NRNI (no regulation no irrigation) points across the Columbia River Basin (CRB) using deterministic measure, i.e. the Kling Gupta Efficiency (KGE). Results show that the new hydrologic post-processing leads to higher accuracy in streamflow simulations.

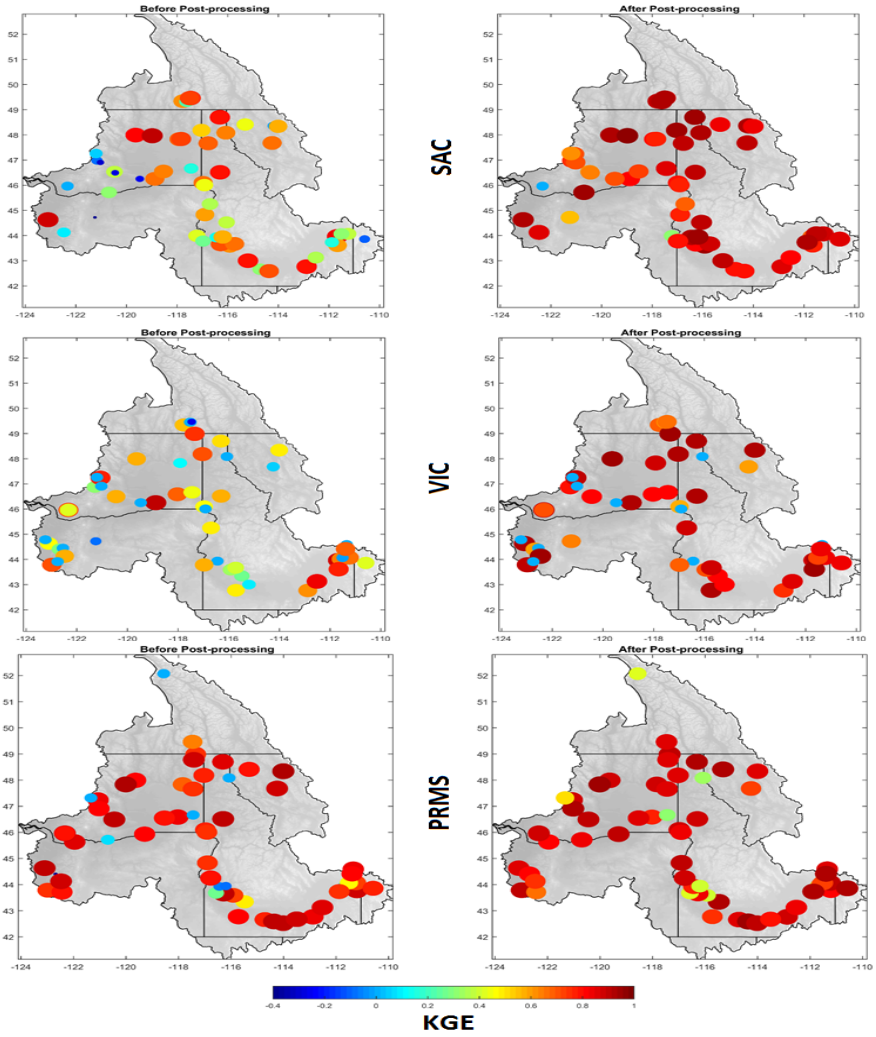

Figure 3. KGE calculated for simulated streamflow at NRNI points before (left) and after (right) post-processing.

D. A probabilistic Post-processing Approach to Improve Precipitation Forecast (Khajehei and Moradkhani, 2016)

Ensemble Post processing (EPP) has become a commonly used approach to reduce the uncertainty in forcing data and hence hydrologic simulation. In this study, we introduced a Bayesian EPP approach based on copula functions (COP-EPP) to improve

the reliability of the precipitation ensemble forecast. Copula functions are capable of building joint distribution between two datasets with any level of dependency, and for any marginal distributions. These characteristics of copula functions help us generate more accurate ensemble forecast. Evaluation of COP-EPP method is carried out by comparing the performance of the generated ensemble precipitation with the outputs from an existing procedure, i.e. Mixed Type meta-Gaussian distribution, which is being used at the National Weather Service River Forecasting System (NWSRFS). Comparison is undertaken by employing three different basins with semi-arid to coastal climate (Figure 4) to study the performance of the techniques in different climate regimes. Verification indicated promising improvement in the mean ensemble using the COP-EPP for generating ensemble precipitation forecast. In order to assess the forecast skill, probabilistic measures including CRPSS, reliability, and the ROC score are employed. The results of CRPSS (Figure 5) indicate that the generated ensemble forecast from COP-EPP is more reliable and accurate in comparison to the meta-Gaussian one. Furthermore, through analysis of reliability (Figure 6), it is noticed that the copula- based method is more successful in generating the ensemble forecasts that represent extremes. The ROC score (Figure 7) indicated that both techniques are capable of generating potentially useful ensemble forecasts with high resolution; however, in the basin with higher precipitation (i.e., Rogue River Basin), COP-EPP proves to be even more superior. COP-EPP is shown to be more precise in building the ensemble precipitation forecast. In other words, results demonstrate that the copula procedure is approximately independent of spatial and temporal changes in the data.

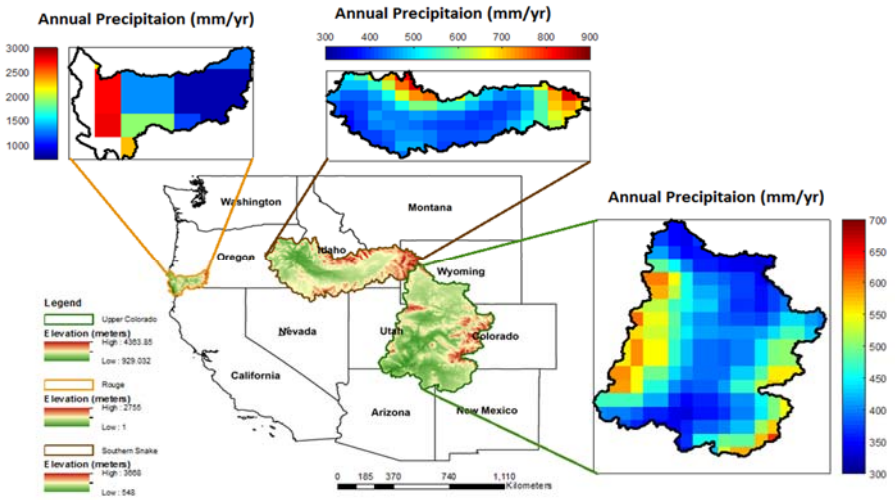

Figure 4. The location of 3 study basins in the Western USA

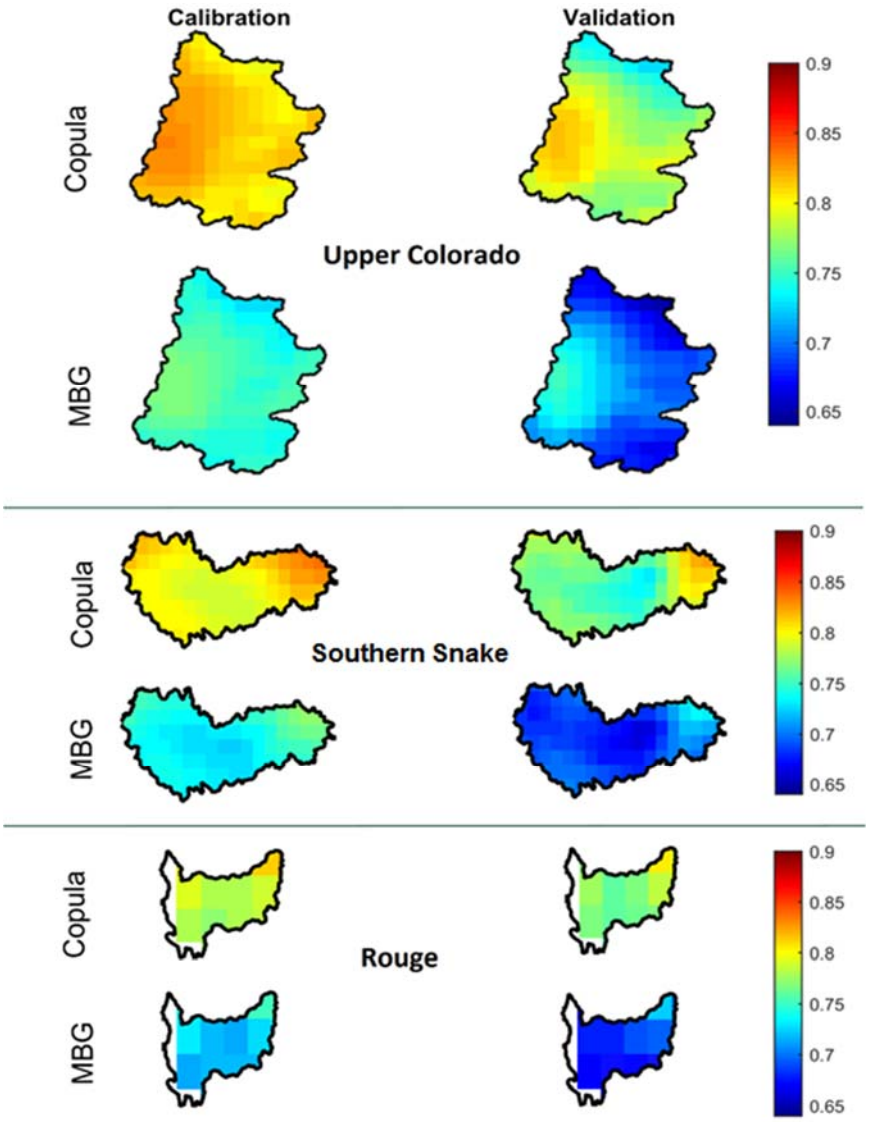

Figure 5. CRPSS measure calculated for 3 basins after two post-processing methods for calibration (left) and verification (right) periods.

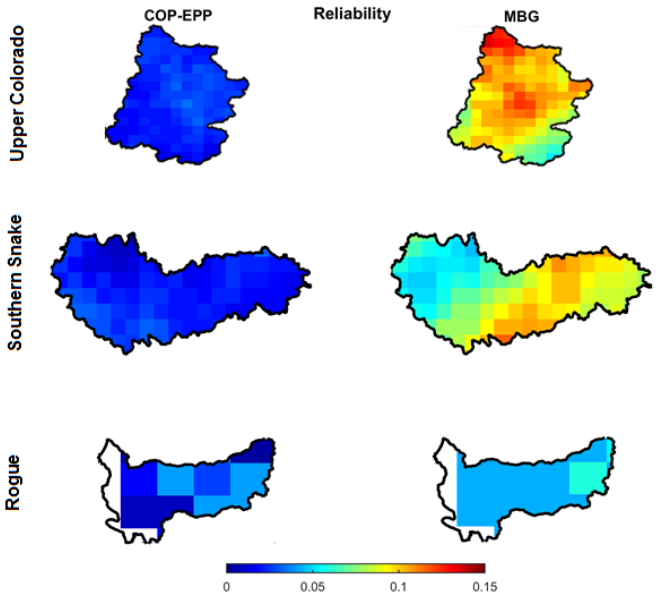

Figure 6. Reliability measure for winter precipitation (Dec, Jan, and Feb) calculated at 95th percentile of observation during the verification period (2001-2014). This measure ranges from 0 to 1 with the optimal value of 1.

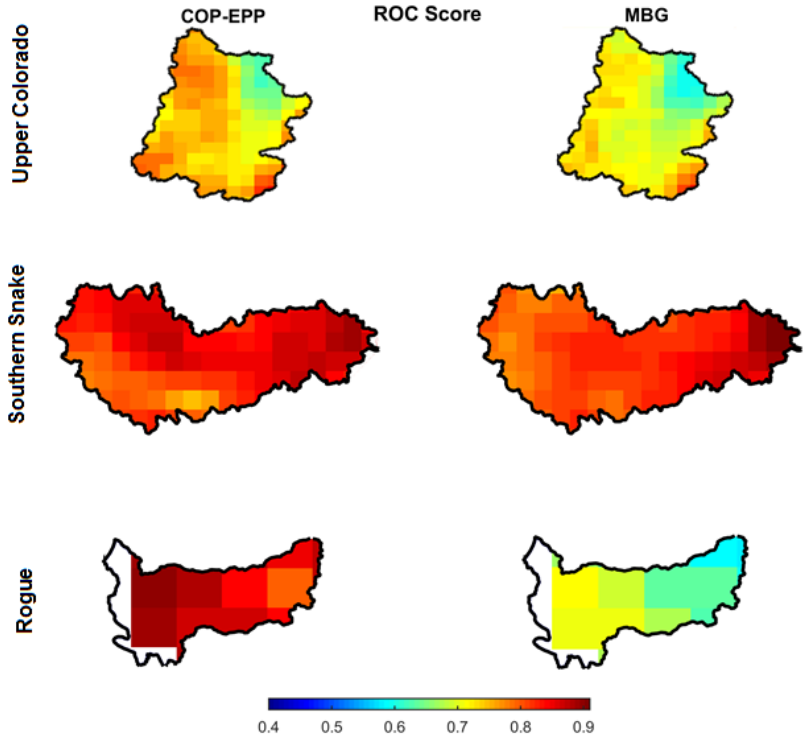

Figure 7. Assessment of forecast resolution through ROC score for winter precipitation (Dec, Jan, and Feb) during verification period (2001-2014).

E. Streamflow Forecasting Uncertainty Reduction: Multi-modeling by Integration of Bayesian Model Averaging and Data Assimilation (Parrish et al. 2012)

Multi-modeling in hydrologic forecasting has proved to improve upon the systematic bias and general limitations of a single model. This is typically done by establishing a new model as a linear combination or a weighted average of several models with weights based on individual model performance in previous time steps. The most commonly used multi modeling method; Bayesian Model Averaging (BMA) assumes a fixed probability distribution around individual models' forecast in establishing the prior and uses a calibration period to determine static weights for each individual model. More recent work has focused on sequential Bayesian model selection technique with weights that are adjusted at each time step in an attempt to accentuate the dynamics of an individual model's performance with respect to the system's response. However, these approaches still assume a fixed distribution around the individual models forecast. A new sequential Bayesian model averaging technique was developed incorporating a sliding window of individual model performance around the forecast. Additionally this technique relaxed the fixed distribution assumption in establishing the prior utilizing a data assimilation method that reflects both the performance dynamics of the models' forecasts along with their uncertainty.

a. General BMA Methodology

Consider a quantity, *y*, to be forecasted, such as the magnitude of a river flow at a particular location and time. Assume we have *k* models, $M = [M_1, M_2, ..., M_k]$ giving us independent model forecast, $Y^f = [y^{f_1}; y^{f_2}; ... y^{f_k}]$ for this quantity for time steps 1 through *T*, where $y^{f_i} = [y_1^{f_i}, y_2^{f_i}, ..., y_T^{f_i}]$. In general, the BMA procedure seeks to compute a new forecast density as a weighted average of the competing models forecasts with weights that correspond to the comparative performance of the models over some training period of observations $Y = [y_1, y_2, ..., y_T]$.

First, the BMA methodology assumes that the model forecasts are unbiased that is the $E[Y - y^{f_i}] = 0$ for each model i. Although there are numerous bias-correction methods, in this paper we incorporate a linear regression of Y on $E[y^{f_i}]$. That is,

$$Y = a_i * E[y^{f_i}] + b_i \tag{1}$$

Unique coefficients a_i and b_i for each model are determined using a least squares approximation, with the observations in the training period as the dependent variable and the forecast as the explanatory variable. These coefficients are then applied to all future model forecast. All future references to model forecast are assumed to be unbiased. Different application strategies for this technique, however, are considered on the Bayesian Modeling Averaging with a sliding window.

The forecast density for y conditioned on the models forecast, M_i, and training period of observations, Y, can be expressed according to the law of total probability as:

$$P(y_t \mid M_1, M_2 ..., M_k, Y) = \sum_{i=1}^{k} P(y_t \mid M_i, Y) P(M_i \mid Y) \tag{2}$$

where $P(y_t \mid M_i, Y)$ is defined as the posterior distribution of y based only on model M_i and the training data Y. $P(M_i \mid Y)$ is defined as the posterior probability or the relative likelihood of model M_i being correct given the training data Y.

As an illustrative example of how this process produces a multi-model forecast PDF, Figure 8 has been prepared. In this illustration, three models are considered. Panel A of the illustration shows the posterior distribution of y for each model. Panel B shows the weight defining the models relative likelihood of being the best model. The product of these weights with the distribution from panel A displays the relative contribution of the models forecast to the eventual PDF. Finally panel D shows the summation and the eventual forecast PDF (i.e., multi-model posterior distribution) for the quantity D.

The various strategies explored in this paper are based on different methods for computing these posterior distributions. For example, a characteristic of the BMA methodology is that a model forecast does not necessarily need to be probabilistic. For deterministic models, this opens up the interpretation of how the posterior distribution, $P(y_t \mid M_i, Y)$ might be defined. Previous applications have assumed that the $P(y_t \mid M_i, Y) \sim g(y_t \mid y_t^{f_i}, \sigma_i^2)$, where σ_i^2 is somehow associated with uncertainty within an individual model and g represents a Normal distribution [*Duan et al*., 2007]. However,

it is possible to relax this assumption using data assimilation techniques such as a PF, whose forecast is a distribution and can act directly as the posterior distribution of *y* given the past model predictions and observations.

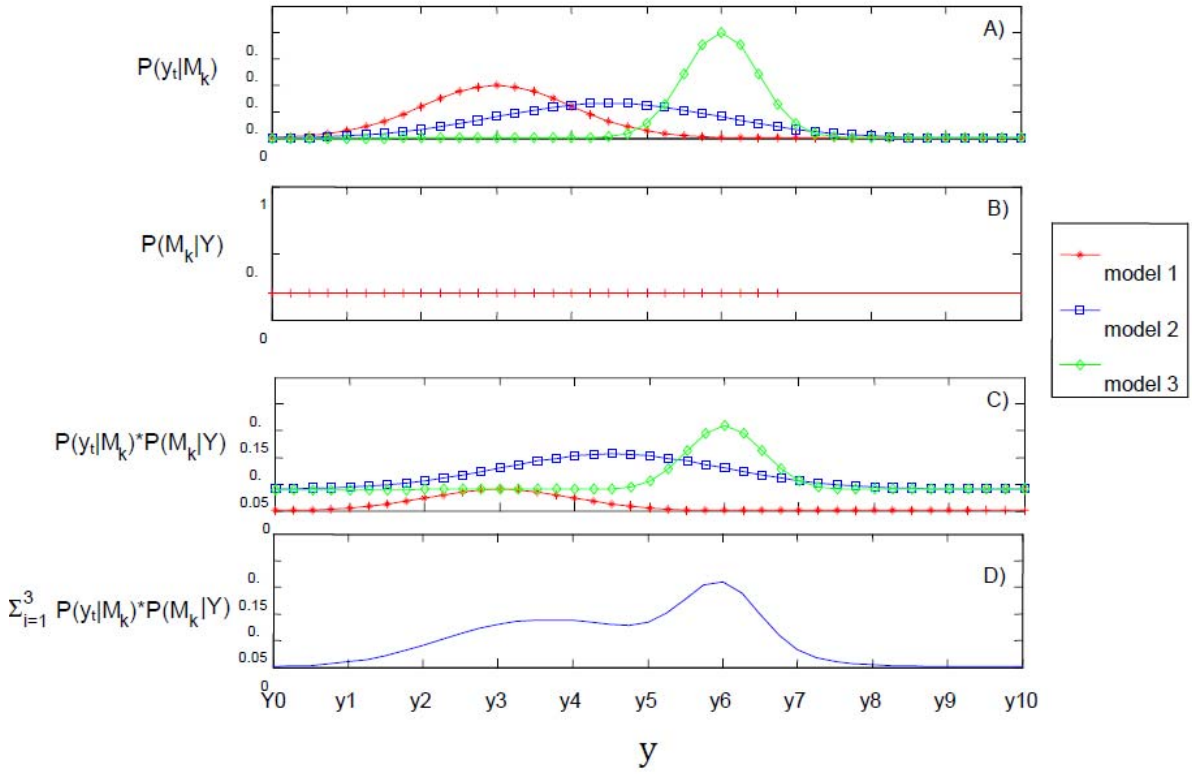

Figure 8. Implementation of Bayesian model averaging on three models: (A) posterior distribution of q for each model, (B) normalized likelihood of model giving the correct response, (C) model forecasts weighted by normalized likelihood, and (D) weighted forecast summed to form multimodel density.

b. Bayesian Model Averaging and Sequential Data Assimilation

Sequential data assimilation estimates the observational and state uncertainty as a PDF around the optimal estimation of the system state. To introduce this approach we first consider a general state-space formulation for any stochastic hydrological model.

Let,

$$x_t^f = f(x_{t-1}^u, \theta, u_t) + \omega_t \tag{12}$$

$$y_t = h(x_t^f) + \upsilon_t \tag{13}$$

Where, x_t^f is a *n*-dimensional vector describing the system states forecast. $f(\bullet,\bullet)$ is the forward model that propagates the forcing data u_t into the system with updated states x_{t-1}^u from previous time step , θ is the model parameter and ω_t represents the process noise. y_t is a scalar forecast for an observation that is related to the system state through the operator $h(\bullet)$ and some observational noise υ_t.

In Bayesian model averaging the strategies determined the posterior density, $P(M_i \mid Y_T)$, by assuming a normal distribution for the posterior probability of y_t given a model M_i, $P(y_t \mid M_i, Y_T)$. Here, we relax that assumption by allowing the observation simulation, developed in the state-space formulation, to approximate this value. The advantage of using data assimilation by means of particle filtering (PF) is that it can identify the multi-modality or skew in state estimation, therefore allowing the simulated observation to be multi-modal or skewed.

c. Model Structure

To explore the merits of this approach, we utilized two conceptual rainfall runoff models, the Sacramento Soil Moisture Accounting (SAC-SMA) and the HYMOD models. SAC-SMA is a lumped rainfall runoff model with sixteen model parameters developed by *Burnash et al.,* [1973]. It remains widely used by the National Weather Service (NWS) in predicting streamflow at different time scales. HYMOD is a parsimonious model which is an extension of simple lumped storage models developed in the 1960s with only five parameters and five state variables. To calibrate the models, the Shuffle Complex Evolution algorithm – University of Arizona (SCE-UA), [*Duan et al.*, 1993] was employed. To address the uncertainty in parameter estimation, we used three distinct objective functions including the Root Mean Square Error (RMSE), Heteroscedastic Maximum Likelihood Estimator (HMLE) and the absolute BIAS. The RMSE is an appropriate measure when the measurement errors are known to be uncorrelated and homoscedastic, or when the properties of the measurement errors are unknown. On the other hand, the HMLE is a goodness of fit estimate when the measurement errors are believed to be heteroscedastic.

We compare the skill of the different Bayesian Model Averaging schemes using both point-wise and probabilistic performance measures. Table 1 outlines the basic differences of each BMA strategy. The goal of forecast verification is to summarize the relationship between a predicted value and its corresponding observation, in order to determine the effectiveness of a forecasting technique across a variety of hydrological conditions and with respect to other forecasting techniques. It is evident that a single performance measure on a single hydrologic condition is not sufficient in answering all of those questions. In this study, we calculate three performance measures associated with accuracy and skill.

Table 1 Comparison of BMA strategies evaluated

Technique	Name	Dynamic Weights	Dynamic Variance	Error Distribution	Average of Models
Static Bayesian model averaging	BMA_Static	No	No	Gaussian	6
Sequential Bayesian combination	SBC	Yes	No	Gaussian	6
Bayesian model averaging with a sliding window	BMA_SW	Yes	Yes	Gaussian	6
Sequential Bayesian	SBC_PF	Yes	Yes	Particle filter	6
Bayesian model averaging with a particle filter	BMA_PF	Yes	Yes	Particle filter	6
Sequential Bayesian combination with a particle filter	SBC_PF	Yes	Yes	Particle filter	6
Bayesian model averaging with model selection	BMA_PF_Threshold	Yes	Yes	Particle filter	Varies on threshold value

Table 2. Pointwise Performance Measures for Individual Models

	Models	Mean Bias	NSE
Deterministic model	Hymod_HMLE	6.6813	0.70652
	Hymod_PBIAS	1.3251	0.74774
	Hymod_RMSE	7.0171	0.78921
	SAC_HMLE	0.21771	0.81588
	SAC_PBIAS	1.6585	0.76208
	SAC_RMSE	4.5353	0.86809
Particle filter	Hymod_HMLE	3.2453	0.81672
	Hymod_PBIAS	−0.05295	0.78383
	Hymod_RMSE	3.144	0.83224
	SAC_HMLE	−0.28105	0.83674
	SAC_PBIAS	1.1422	0.77471
	SAC_RMSE	3.4449	0.89354

The performance assessment is conducted by first evaluating the point-wise performance measures across the entire eight year validation time period as shown in Figure 9.

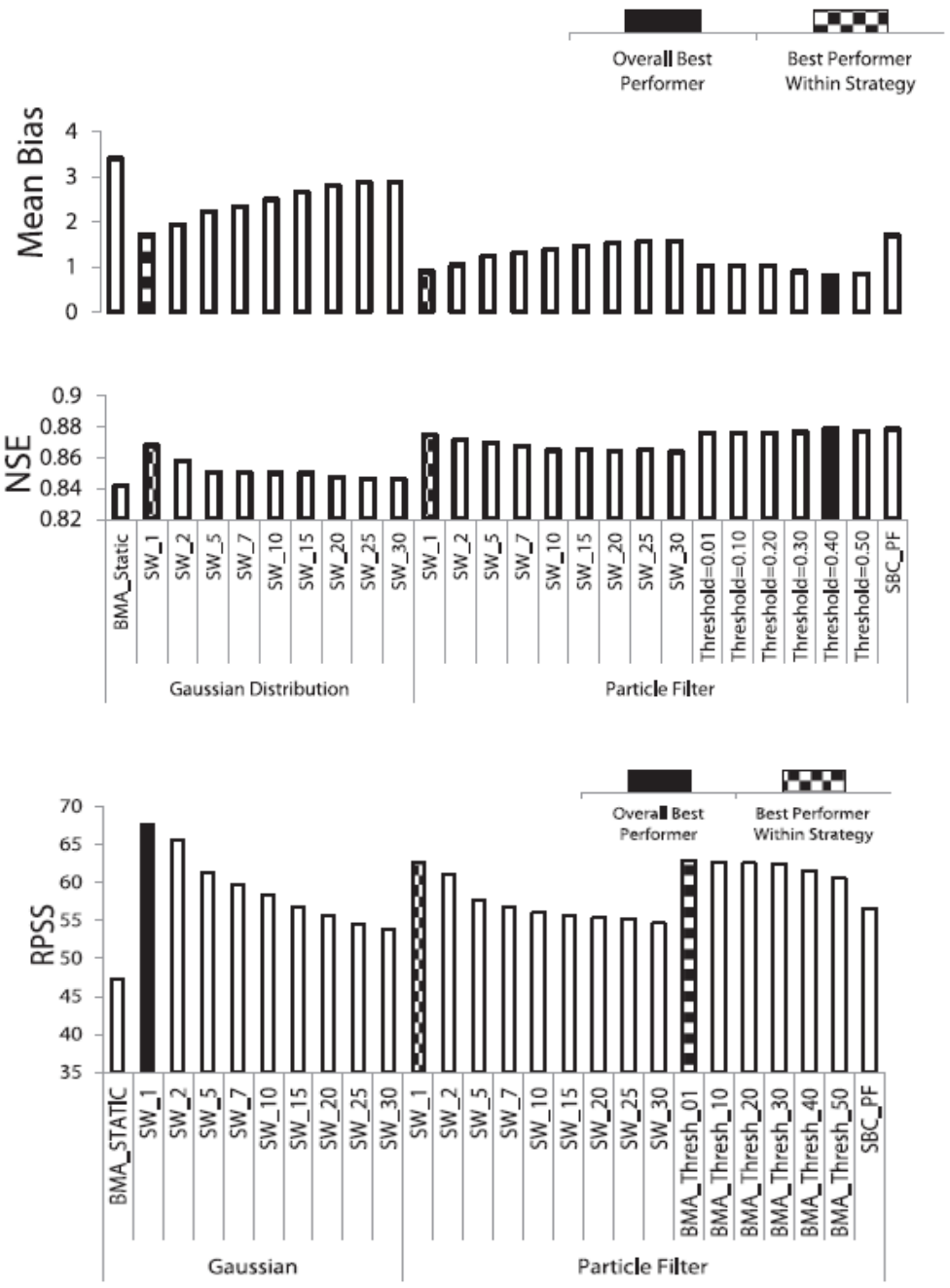

Figure 9. Pointwise and probabilistic performance measures for the complete hydrograph.

The assessment is followed by analyzing different regions of the observed hydrolograph by separating the hydrograph into different flow values for example; high, medium, and low flows. The sliding window schemes we are analyzing, however, not only address the volume of flow, but also the potential of the scheme to quickly adapt to

rapid changes in the hydrograph. For each range of volatility, different Bayesian model averaging schemes performed the best.

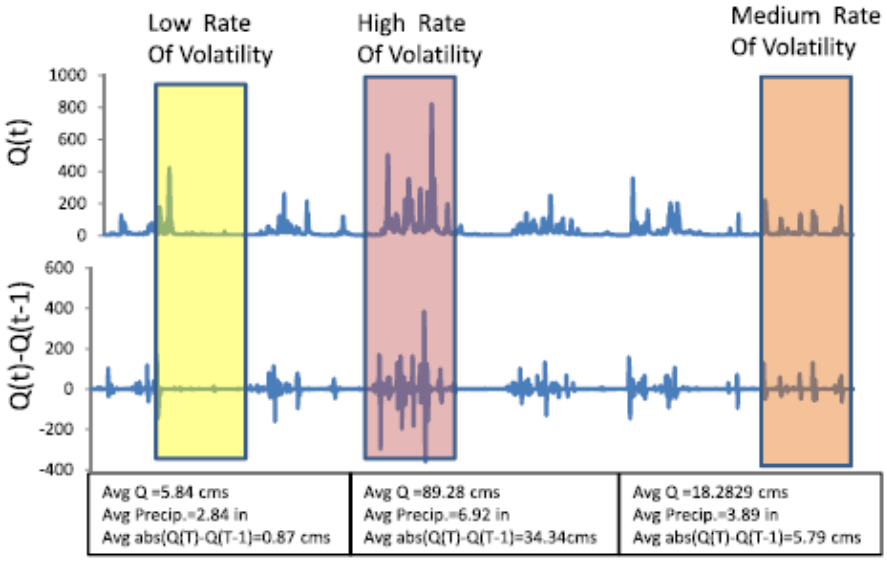

Figure 10. Volatility stages during the validation period of 1 October 1980–30 September 1988.

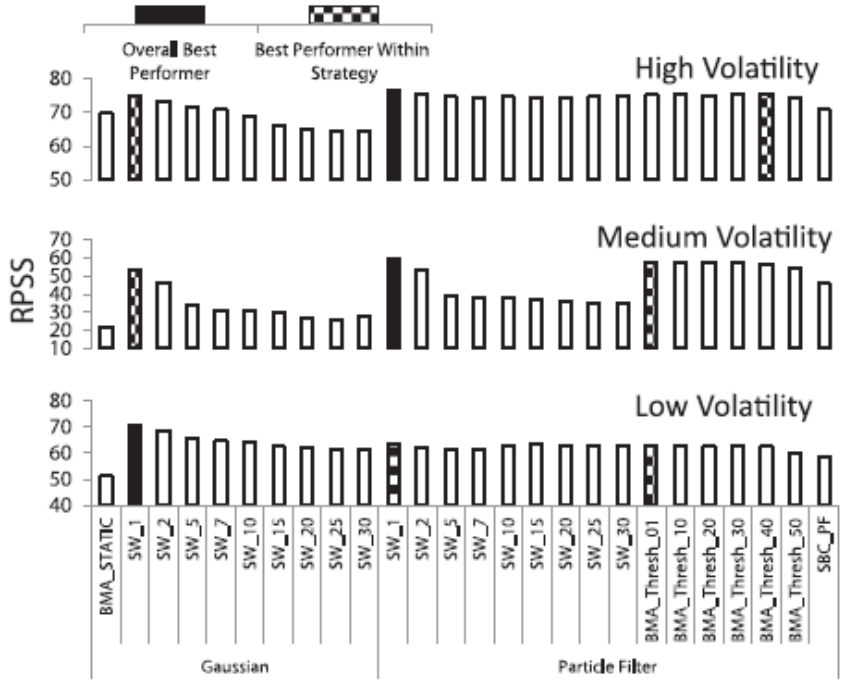

Figure 11. RPSS values for various rates of change on the hydrograph

The analysis shows that the strength of each of the different averaging techniques varies across performance metrics and volatility ranges of the hydrograph. The advantage of the multi-model averaging is most apparent in measuring probabilistic skill, but some advantage is apparent for pointwise metrics. The comparison of the multi-modeled averaged forecast with the individual models forecast illustrates that averaged forecast do not always outperform the best individual model for any particular metric. A comparison of the BMA strategies to the individual models probabilistic forecasts confirmed the strength of using model averaging on over-confident individual models. Two variations of the static BMA considered in this study are to allow for dynamic model weights and to dynamically change the set of models averaged. Both of these variations showed conflicting results when compared across all the performance measures.For pointwise metrics and RPSS values, the dynamic model weights and model uncertainty, generated by the sliding window approach, outperformed the static Bayesian Model Averaging scheme. This dynamic approach allows a necessary flexibility in gauging the confidence in a model output, with respect to the changes in the hydrograph.

Across the ranges of volatility, BMA with data assimilation strategies reduce the prediction interval by nearly fifty percent when compared to strategies utilizing the Gaussian uncertainty distribution. For low and medium ranges of volatility, this corresponds to an over-confidence in the BMA with DA strategies, capturing fewer than 95% of the observations. However for high volatility ranges, the BMA with DA approach with a sliding window of thirty nearly captured 95% of observations within its prediction interval while reducing the prediction interval width by 40% in comparison to the best BMA strategy.

F. Improving Multimodeling by integrating Multivariate functions (Copulas) to Bayesian model averaging (Madadgar, and Moradkhani, 2014)

Bayesian model averaging (BMA) is a popular approach to combine hydrologic forecasts from individual models and characterize the uncertainty induced by model structure. In the original form of BMA, the conditional probability density function (PDF) of each model is assumed to be a particular probability distribution (e.g., Gaussian, gamma, etc.). Since copula functions have shown success in different hydrologic forecasting applications, this study utilized them in model averaging to find the posterior distribution of data given model predictions. Copula functions have a flexible structure and do not restrict the shape of posterior distributions. Furthermore, copulas are effective

tools in removing bias from hydrologic forecasts. In this study, we proposed a technique that merge the copula functions with BMA (COP-BMA) to reduce the uncertainties that arises from the limitations in BMA. To compare the performance of BMA with Cop-BMA, they were applied to hydrologic forecasts from different rainfall-runoff and land-surface models. The new method, Cop-BMA, is more flexible in defining the posterior distribution and does not impose any restriction on the type of distribution. Although BMA is not theoretically limited to a certain type of posterior distribution, either the unimodal distributions such as Gaussian or gamma distribution are commonly used as posterior distributions. In contrast, Cop-BMA has a flexible structure that allows the posterior distribution to have any unimodal or multimodal shape depending on the copula function. By relaxing the assumptions on the type of posterior distribution, data transformation would not be required. Furthermore, Cop-BMA can effectively remove the bias of initial forecasts by itself and do not need any external bias-correction method.

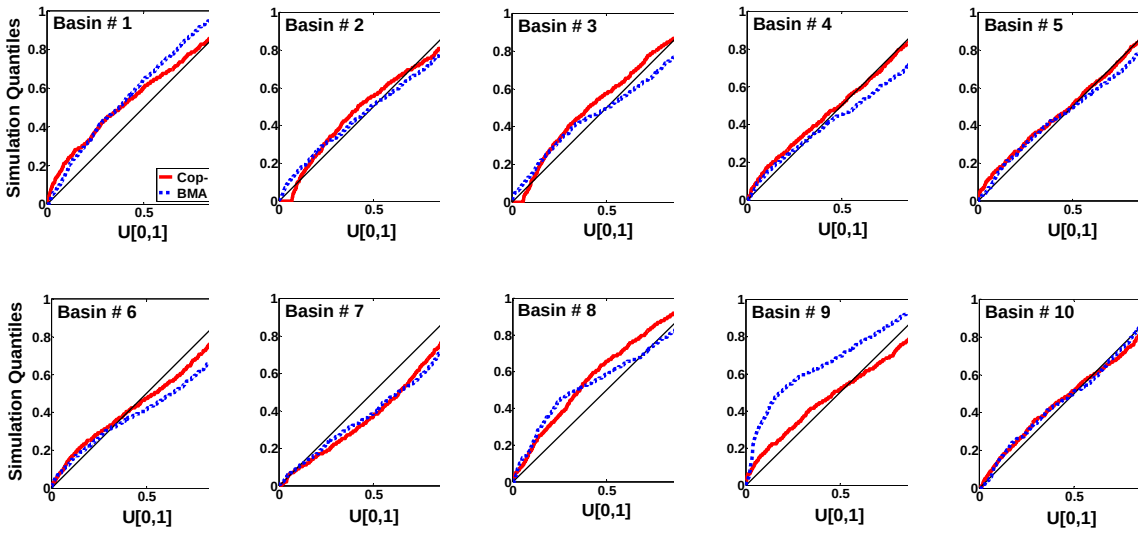

Figure 12. Comparison of predictive QQ plot produced by BMA versus Cop-BMA.

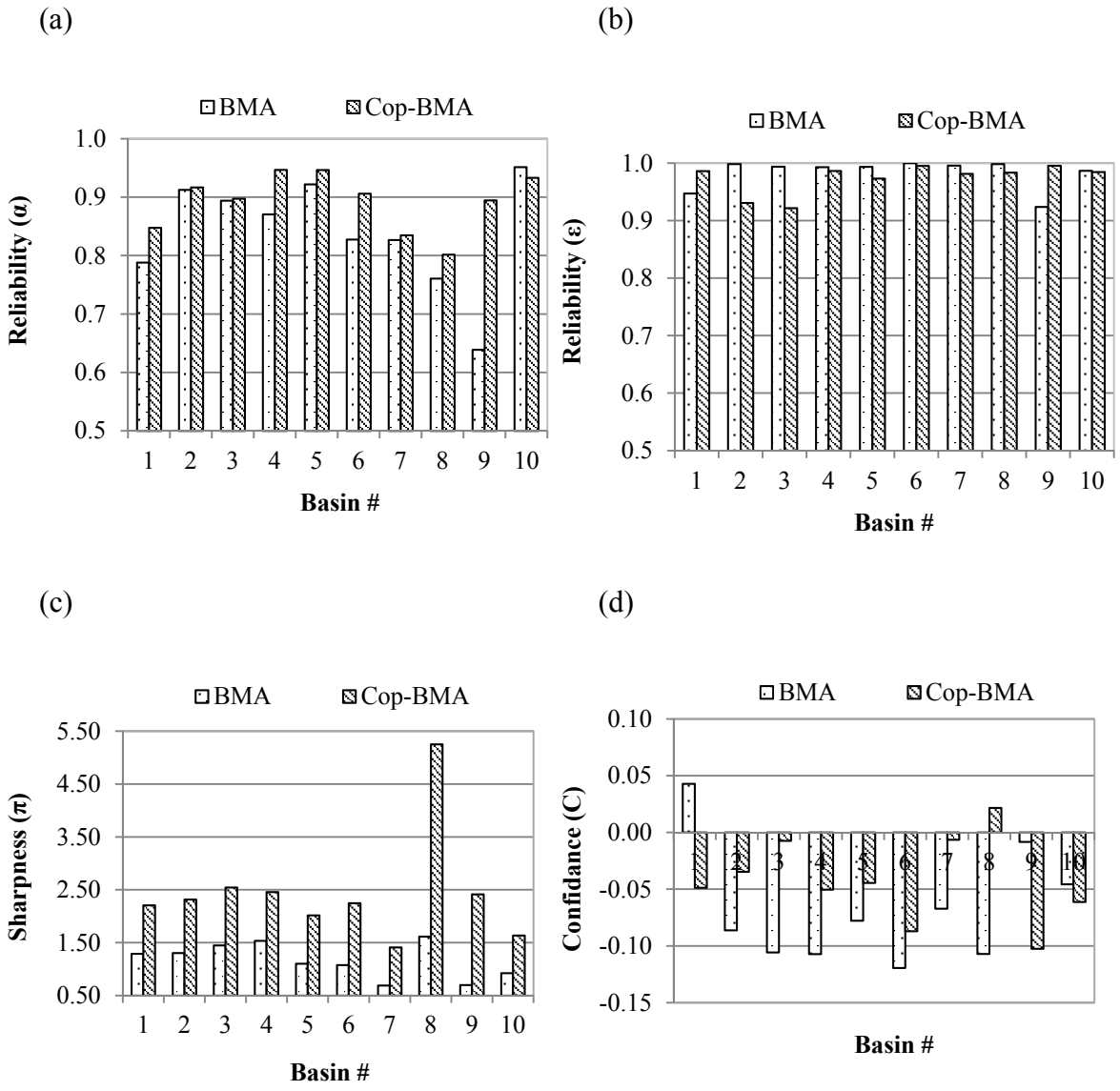

Figure 13. Comparing the performance of BMA and Cop-BMA indicated by (a, b) reliability, (c) sharpness, and (d) confidence.

G. Ensemble Combination of Seasonal Streamflow Forecasts (Najafi and Moradkhani, 2015)

To the best of knowledge of the PI, the question of how to optimally combine statistical models with dynamical models had not been investigated before. Dickinson (1975) examined combination of forecasts using a minimum variance criterion and suggested that unreliability of the estimated weights might downgrade the forecast

combination. Simple model combination has been considered in several studies through assigning equal weights to the individual forecast models. Using seven distributed hydrologic models applied to six basins, Georgakakos et al. (2004) found that the simple mean of five best models outperformed the best models in each basin. The simple mean of the individual forecasts, however, is based on the assumption that single models perform similarly (Najafi et al. 2015). This approach also does not quantify the related uncertainties in the model combination process. Other methods that weigh and combine single models based on their performances in a calibration time period have been considered using more complex approaches (Buser et al. 2010; Duan et al. 2007; Najafi et al. 2011; Parrish et al. 2012; Tebaldi and Knutti 2007).

Comparison between various model combination techniques has been performed in several studies (Ajami et al. 2006; Georgakakos et al. 2004; Shamseldin et al. 1997). Xiong et al. (2001) proposed a fuzzy system to combine forecasts of rainfall-runoff models and compared the results with simple and weighted average methods along with neural networks. Diks and Vrugt (2010) compared multi-models based on point predictors to the ones based on density forecasts such as Bayesian Model Averaging (BMA). They concluded that Granger-Ramanathan averaging (based on ordinary least squares regression) performed similar to more complex multi-model approaches such as Mallows model averaging and BMA. Viney et al. (2009) compared various combination techniques to merge ensemble daily model predictions of catchment streamflow. They concluded that multi-model ensemble predictions are generally superior to the constituent models. Bohn et al. (2010) considered two multi-model methods of simple mean and multiple linear regression (MLR) for three hydrological models over three basins. Their results showed that the individual best bias corrected model outperformed the combination of raw models. In addition, simple model average method generated smaller error reductions compared to MLR.

Although several combination techniques have been investigated in hydrologic applications, limited information is available on ensemble merging of dynamical and statistical hydrologic models, with varying complexities in seasonal streamflow forecasts.

Seasonal water supply outlooks, or volume of total seasonal runoff, are routinely used by decision makers in the Western US for making commitments for water deliveries, determining industrial and agriculture water allocation, and operating reservoirs. These forecasts are based primarily on statistical regression equations developed from monthly precipitation, recent snow-water equivalent, and a subset of past streamflow observations.

In the Western US, the National Weather Service Northwest River Forecast Center (NWRFC) and the Natural Resources Conservation Service (NRCS) jointly issue seasonal water supply outlook forecasts of naturalized or unimpaired flow, i.e. the flow that would most likely occur in the absence of diversions. This is done using statistical and ensemble streamflow prediction (ESP) methods developed at the NWS and NRCS. In addition, water resources management entities including US Bureau of Reclamation (BOR) and Corps of Engineers (COE) use their own statistical models to issue seasonal water supply forecasts.

Considering a wide range of available seasonal forecast approaches, a verifiable multi-model approach that effectively combines single forecasts and enhances the individual models is required. Combination of seasonal forecast models with varying complexities, if done through a robust framework, would reduce the sophisticated interpretation of the forecasts and ease the communication with stakeholders. Furthermore, ensemble merging of hydro-climatic extreme predictions from various methods (Halmstad et al. 2013; Najafi and Moradkhani 2014; Najafi and Moradkhani 2015) can also be considered as an effective approach to increase the reliability of predicted flood events.

This study addresses the multi-model ensemble merging of dynamical and statistical hydrologic predictions from different agencies along with simulations based on independent component analysis (Moradkhani and Meier 2010; Najafi et al. 2011) and partial least square regression. A comprehensive set of model combination techniques with varying complexities are applied over the individual forecast models to evaluate the advantages of merging seasonal forecasts. Furthermore, taking into account a wide range of model combination methods with varying complexities, we assess the relative performance of each method and the related uncertainties. In particular we intend to investigate whether complex models would necessarily outperform the simpler combination methods.

a. Study Area

Table 3 presents the characteristics of the river basins considered in this study. The first column represents the geographical locations as well as other basin specifications including the number of seasonal streamflow forecasts available for each basin. Libby has the largest drainage area of approximately 9000 mi^2, whereas Granby has the smallest area of 323mi^2. Correspondingly, the two basins experience the largest and the smallest

average seasonal flow. April is the starting month of the seasonal runoff forecasts and the ending months differ between basins.

Dworshak Dam near Ahsahka, Idaho is used to regulate annual floodwaters of the North Fork Clearwater River and for power generation. Lake Granby, located in the headwaters of the Colorado River Basin, is the largest storage reservoir in the Colorado-Big Thompson (C-BT) reservoir system, a large trans-basin water storage and delivery project. While the project was originally built for agricultural purposes, it serves multiple demands including municipal and industrial supply, hydro-power generation, recreation, and fish and wildlife. In recent years, however, water supply demands have shifted making municipal and industrial supply the main water beneficiary, rather than irrigation. Libby Dam is located on the Kootenai River in northwestern Montana, approximately 40 miles south of the US-Canadian border. It is operated by the US Army Corps of Engineers as a multi-purpose project for hydropower, flood control, and recreation. Project operations also incorporate water quality and quantity targets in support of fisheries and environmental objectives. The Rogue River is located in southwestern Oregon which drains the area between the Cascade Mountains and the Pacific Ocean. Flow has been regulated since February 1977 by Lost Creek Lake including a slight regulation by Fish Lake and Emigrant Lake. There are many diversions for irrigation upstream from the *station.*

Table 3. River basin characteristics

Name	**Dworshak, ID**	**Lake Granby, CO**	**Libby, MT**	**Rogue, OR**
USGS ID	13340000	09019500	12301933	14359000
Latitude	46.478	40.121	48.401	42.437
Longitude	116.257	105.9	115.319	122.986
Drainage Area (mi^2)	5,507	323	8,985	2,053
Datum of Gage (ft. above NGVD29*)	990.80	7,960	2,100	1,121.78
Average seasonal Runoff (kaf)	2462.1[1]	218.6[2]	6168.9[3]	816.3[4]
Time Period	1981_2000	1981_2005	1988_2002	1981_1997
Number of Forecast Models	4	5	3	4

[1,2] Seasonal Streamflow from April_July

[3] Seasonal Streamflow from April_August

[4] Seasonal Streamflow from April_September

*National Geodetic Vertical Datum of 1929

b. Ensemble Streamflow Prediction and Statistical Forecast

In the United States, the National Weather service (NWS), through its network of River Forecast Centers, and the Natural Resources Conservation Service (NRCS) provide operational water supply forecasts (Twedt et al. 1977). The traditional approach for forecasting water supply volumes has been based primarily on statistical regression equations developed from monthly precipitation, recent snow water equivalent (snow water equivalent measurements at 1st of the month), and a subset of past streamflow observations, to predict streamflow volumes. This method provides reliable long lead forecasts (monthly to seasonal) with the exception of occasional failures in the extreme event years (Day 1985). The regression-based seasonal streamflow forecasts are incapable of providing information about different sources of uncertainties in their forecasts due to their mathematical structure (Twedt et al. 1977). In addition, incorporation of new sources of data (satellite observation) and new methods (e.g., data assimilation) within the regression framework is difficult, in part, because the forecast models require training over a long time series of historical observations (Wood and Lettenmaier 2006). Moreover, the regression method may be inappropriate in a non-stationary climate (due to the effects of climate change) (Cayan et al. 2001; Hamlet et al. 2005; Wood and Lettenmaier 2006). Ensemble streamflow prediction (ESP) has been proposed to address these restrictions (Day 1985). In ESP, the calibrated hydrologic model is run for a sufficiently long period until the forecast time in order to obtain the watershed initial condition. The model is then driven by resampled historical datasets to generate an ensemble of possible future streamflow in the basin. Each historical data is treated as a realization of the atmospheric forcing, which is used to simulate the streamflow trajectories. This grants ESP a firm ground for considering the uncertainty pertained to the future climate, which might be the major component of forecast uncertainty in some seasons (Najafi et al. 2012; Wood and Schaake 2008).

c. Forecast Models

Table 4 provides a summary of the seasonal streamflow forecast models that were available for model combination. Not all models existed for each basin. Results of the ensemble streamflow prediction (ESP), for example, were only accessible for the Lake Granby basin. A short description of each forecast model along with the corresponding river forecast center is also provided. Commonly, the statistical forecasts are based on the principal component analysis, which provides uncorrelated signals used as predictors in a

multivariable linear regression model. The statistical forecasts based on partial least squares regression (PLSR), independent component analysis (ICA) and Z-score regression were also added to the current forecasts. The analyses based on these methods are discussed in more details by (Moradkhani and Meier 2010; Najafi et al. 2010).

Table 4. Seasonal streamflow forecast models used for multi-modeling.

Name	Description	Forecast Center
ESP	Ensemble Streamflow Prediction: NWS*	National Weather Service River Forecast Center
CSWS	Principal Component Regression (PCR): NWRFC*	Natural Resources Conservation Service
zcbrt	Z-score with cubic transformations on predictors: NRCS*	Natural Resources Conservation Service
PLSR	Partial Least Square Regression	Portland State University
ICA	Independent Component Analysis	Portland State University
Zsimflow	Z-score Regression	Portland State University
SWS	Statistical Water Supply	US Army Corps of Engineers
PCR2011	2011 Libby Apr-Aug water supply forecast using PCR: USACE*	US Army Corps of Engineers
PCR2004	2004 Libby Apr-Aug water supply forecast using PCR: USACE	US Army Corps of Engineers
MW1986	Morrow-Wortman Libby water supply forecast (Split-Basin Regression)	US Army Corps of Engineers

* **NWS**: National Weather Service; **NWRFC**: Northwest River Forecast Center; **NRCS**: Natural Resources Conservation Service; **USACE**: United States Army Corps of Engineers;

Using the available models from different agencies along with the simulation results from PLSR, ICA, and Z-score, the performance of each seasonal streamflow forecast was analyzed in each study area (Table 5). Four performance measures were considered for this purpose: BIAS, RMSE, Nash Sutcliffe Efficiency (NSE) and Kling-Gupta efficiency (KGE) (Gupta et al. 2009).

The models given in Table 5 is sorted based on NSE values. The NSE, RMSE and KGE values mostly agree for all basins, except KGE for ESP at Lake Granby. BIAS generally agrees with other measures except for the Rogue River basin. Overall, the seasonal forecast models perform satisfactorily with NSE values over ~0.8 except for Granby with NSEs over 0.5. One-to-one comparison between the individual forecasts is

not possible, since data obtained from different agencies might belong to different operational stages (e.g. operational mode, leave-one-out cross validation, and after-the-fact reforecast).

Table 5. Performance measures of the seasonal streamflow forecast models for a) Dworshak b) Granby c) Libby and d) Rogue river basins.

a)

Model	NSE	KGE	RMSE	BIAS (%)
Zsimflow	0.87	0.92	293.91	0.54
SWS-NRCS	0.86	0.90	303.31	0.88
SWS-NWS	0.83	0.91	326.34	0.48
SWS-USACE	0.82	0.90	343.98	2.33

b)

Model	NSE	KGE	RMSE	BIAS (%)
ICA	0.66	0.71	39.66	2.2
PLSR	0.64	0.69	40.73	2.32
zcbrt	0.61	0.69	42.65	2.07
CSWS	0.51	0.51	47.54	0.63
ESP	0.5	0.68	48	10.01

c)

Model	NSE	KGE	RMSE	BIAS (%)
PC2010	0.88	0.92	537.3	0.97
PC2004	0.88	0.94	542.19	0.48
MW1986	0.82	0.88	654.38	2.06

d)

Model	NSE	KGE	RMSE	BIAS (%)
ICA	0.96	0.95	59.02	1.59
Zsimflow	0.89	0.92	95.39	1.24
SWS-NWRFC	0.86	0.82	105.78	-0.46
SWS-NRCS	0.79	0.77	132.71	-1.12

d. Model Combination Strategies

A summary of the model combination techniques used in this study is shown in Table 6. We categorize the models into three segments according to their degrees of complexities. Simple models consist of mean, median and linear regression, and complex

models include Bayesian model averaging methods which are probabilistic approaches that require more complicated parameter estimation approaches. Intermediate segments include methods based on information criteria and principal component analysis. The segmentation allows for comparison of multi-model performances regarding their complexities.

e. Simple Methods

Simple average and median are the two basic averaging methods that are applied on all of the seasonal forecast model results. The mean of the forecast models is calculated as $MM = \frac{1}{N}\sum_{i=1}^{N} X_i$ where N is the total number of models and X_i is the seasonal runoff predicted by the ith model. Assigning equal weights to all forecast models, however, ignores the fact that some models would perform better than the others.

In the Bates-Granger (Bates and Granger 1969) approach, the individual models are empirically weighted based on their errors in the calibration period:

$$w_f = \frac{\hat{\sigma}_f^{-2}}{\hat{\sigma}_1^{-2} + \hat{\sigma}_2^{-2} + \ldots + \hat{\sigma}_N^{-2}}, \tag{14}$$

where $\hat{\sigma}_f^2$ is the variance of the forecast error of model "f" over the calibration period. It assumes that the performances of the individual forecast models would not change, having a constant variance of residuals over time, and also the forecast models are unbiased.

In the Granger Ramanathan Average approach (constrained and unconstrained), the linear regression model of the forecasts was created in the form of:

$$y_t = \sum_{i=1}^{k} \beta_i X_{i,t} + \nu_t, \tag{15}$$

where ν_t is a white noise with a Gaussian distribution. X represents individual seasonal forecasts and the model combination result is given by y. The scale parameter β is calculated during the process of calibration period using the available time series of the observational data (y) and the forecasts (X). In this study two GRA approaches were considered which included the application of the estimated scaling factors (β), based on the ordinary least square method, directly to the forecast models for the test period. In

this scenario, negative values of the scaling factors could be generated. In the second approach, the estimated βs were converted to $w_i = abs(\beta_i) / \sum_i abs(\beta_i)$. The resulting weights (w_i) were then applied to the corresponding test datasets.

Generalized Linear Model (MacCullagh and Nelder 1989), is a natural extension of the simple linear regression model. In GLM instead of modeling the expected response directly as a function of the linear predictors, i.e. $\mu = E(Y_t) = \sum_{i=1}^{k} \beta_i X_{i,t}$, a function of $g(\mu_t) = \sum_{i=1}^{k} \beta_i X_{t,i}$ is considered. $g(.)$ is a smooth and invertible linearizing function called a link function and is considered to be logarithmic in this study. Hence, the GLM is regarded as a linear model for a transformation of the expected response or a nonlinear model for the response. In this study, we assume that the components of the response vector follow a gamma distribution.

f. Intermediate Methods

Principal component regression (PCR) combines principal component analysis (PCA) with multiple linear regressions. PCA is a technique that creates new uncorrelated variables (principal components or PCs) by projecting the original predictors onto an orthogonal space. Principal components are generated from linear combinations of the predictor variables, which are obtained by multiplying each predictor variable with the corresponding loading and summing the results. The loadings used in the linear equations are the elements of the eigenvectors, which are calculated from the sample variance-covariance matrix of the standardized predictor variables (Moradkhani and Meier 2010).

Similar to PCR, the partial least square regression (PLSR) also provides uncorrelated new predictors that are linear combinations of the original predictor variables. In the PLSR method, in addition to the predictors (which are accounted for in the PCR technique) the predictand is also considered to generate the principal components.

Methods based on information Criteria (Burnham and Anderson 2002) were used as model selection approaches that deal with the trade-off between the goodness of fit and the complexity of the model. The likelihood of model "g_i" given the observed dataset "y" can be calculated by:

$$L(g_i|y) \propto \exp\left(-\frac{1}{2}\Delta_i\right), \tag{16}$$

where $\Delta_i = AIC_i - AIC_{\min}$ and AIC_{min} is the minimum of the N different AIC_i values (eq. 17), therefore, the best model results in $\Delta = 0$. The Akaike Information Criterion (AIC) is obtained from:

$$AIC = -2\log\left(L(\Theta|y)\right) + 2K, \tag{17}$$

where $L(\Theta|y)$ is the likelihood of model parameters given observed data "y" which measures the model fit . According to information criteria model weights are given by:

$$w_i = \frac{L(g_i|y)}{\sum_{r=1}^{R} L(g_r|y)}, \tag{18}$$

"R" is the total number of models, and "K" is a penalty term which increases by the size of the model (number of parameters).

corrected AIC (AICc) (Hurvich and Tsai 1989) is defined as:

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}, \tag{19}$$

with "n" representing the calibration sample size. For large n, AICc converges to AIC. Bayesian information criterion (BIC) and Consistent Akaike information criterion with Fisher information (CAICF) are given by Eqs. 20-21:

$$BIC = -2\log(L) + K.\log(n), \tag{20}$$

$$CAICF = -2\log(L) + K\{\log(n) + 2\} + \log\left|I(\hat{\theta})\right|, \tag{21}$$

where $\log\left|I(\hat{\theta})\right|$ is the natural logarithm of the determinant of the estimated Fisher information matrix. *Bozdogan (1988)* proposed a criterion which is close to CAICF and is based on a concept of complexity.

Mallows's Cp, considers the minimum mean squared error model selection for regression, and is completely discussed by Mallows (1995). Considering the $n \times k$ matrix of X; P as any subset of $K = \{1, 2\ldots k\}$; X_P the sub-matrix of X and $PE_P = \left(X\beta - X_P\hat{\beta}_P\right)^2$

$$C_P = \frac{RSS_P}{\hat{\sigma}^2} - n + 2p, \tag{22}$$

where $RSS_P = \left(y - X_P \hat{\beta}_P\right)^2$ and $\hat{\sigma}^2 = \frac{RSS_K}{(n-k)}$.

Considering the methods based on information criteria, predictions of the test cases were performed by two approaches which included the averaging of a small number of selected models (represented by 'm') or using the model averaged coefficients (represented by 'avgcf') (Burnham and Anderson 2002).

g. Complex Methods

In Bayesian Model Averaging (BMA) each ensemble member 'X_i' is used in an individual probability distribution function $p(y|X_i)$, which is the PDF of the hydro-climate variable 'y' conditional on model X_i being the best ensemble member. The posterior probability of each model X_i given the observed data O (i.e. $p(X_i|O)$) is considered as the weight associated with that model reflecting its relative performance in the training period. Hence the BMA weights are probabilities and are summed up to unity. The resulting PDF associated with 'y' is a weighted average of each constituent PDF.

Table 6. Summary of the multi-modeling averaging methods.

	Multi-Modeling Approach	**Acronym**
Simple Models	Equal Weight Average	EWA
	Median	Med
	Bates-Granger	BGA
	Constrained Granger Ramanathan Average	CGRA
	Unconstrained Granger Ramanathan Average	UCGRA
	Generalized Linear Model	GLM
Intermediate Models	Principal Component Regression (1:4 components)	PCR
	Partial Least Square Regression (1:4 components)	PLSR
	Akaike Information Criteria	AIC
	corrected Akaike Information Criteria	AICc
	Bayesian Information Criteria	BIC
	Consistent Akaike information criterion with Fisher information	CAICF
	Mallows's Cp Information Criteria	Cp
	Bozdogan's Index of Informational Complexity	ICOMP
Complex Models	Bayesian Model Average in Linear Regression Model	BMA-LR
	Bayesian Model Average in Generalized Linear Model	BMA-GLM
	Bayesian Model Average with EM algorithm	BMA-EM
	Bayesian Model Average with MCMC algorithm	BMA-MCMC

h. Bootstrapping

A bootstrap approach is used to verify the performance of multi-model ensemble merging techniques. Considering M individual forecast models, and P years of available observed datasets, each ensemble merging technique is trained based on 70% of sampled datasets to perform the multi-model averaging and predict the remaining time series of observed seasonal streamflow data. This process is repeated for 200 times (with sample replacement) resulting in different predictions of each observed flow (y_i) based on varying sets of training datasets. The means of predictions for each seasonal streamflow are then taken as the final multi-model prediction (Jiang and Simon 2007).

i. Results

Figure 14 shows the root mean square error (RMSE) and BIAS of all multi-models and individual forecast models for each river basin (each panel). Regarding Dworshak, four of the simple multi-model ensemble averaging techniques and one of the complex techniques outperformed the best individual model forecast according to RMSE which include simple average, median, Bates-Granger and constrained linear regression along with BMA optimized by expectation maximization (BMA-EM). Increasing the number of principal components in PCR and PLSR has resulted in increased RMSEs. GLM and BMA-LR1 performed weakly while intermediate methods generally showed average performance. The individual model forecasts are all positively biased. The multi-model ensemble averaging methods which showed improvements based on RMSE, generally improved the BIAS as well. It should be noted that BIAS is a measure of correspondence between the average forecast and the average observed streamflow. Therefore, improving BIAS alone does not indicate the good performance of the multi-modeling approach. For example, although PCR4 and PLSR4 have considerably improved based on the BIAS, the corresponding RMSEs have increased indicating that their accuracies are reduced.

Regarding Lake Granby, similar to Dworshak, the simple average, median, Bates-Granger and constrained linear regression methods showed the best results. Although they did not outperform the best individual forecast model, their performances were close. PCR1, PCR2 and BMA-EM also performed better than most of the individual models according to RMSE measure. Increasing the number of principal components in PCR and PLSR resulted in increased RMSE values. Also GLM showed a weak performance compared to the other models. All the individual model forecasts are

positively biased as well. The multi-model ensemble averaging methods generally improved the overall individual model forecast biases.

Regarding the Libby basin with only three forecast models, the simple average, median and BMA-EM methods outperformed the best individual forecasts, while Bates-Granger and constrained linear regression showed close performances to the best forecast. Generally the intermediate and complex multi-models showed weak performances according to the RMSE accuracy measure. Similar to Dworshak and Granby, the individual model forecasts are, in average, positively biased.

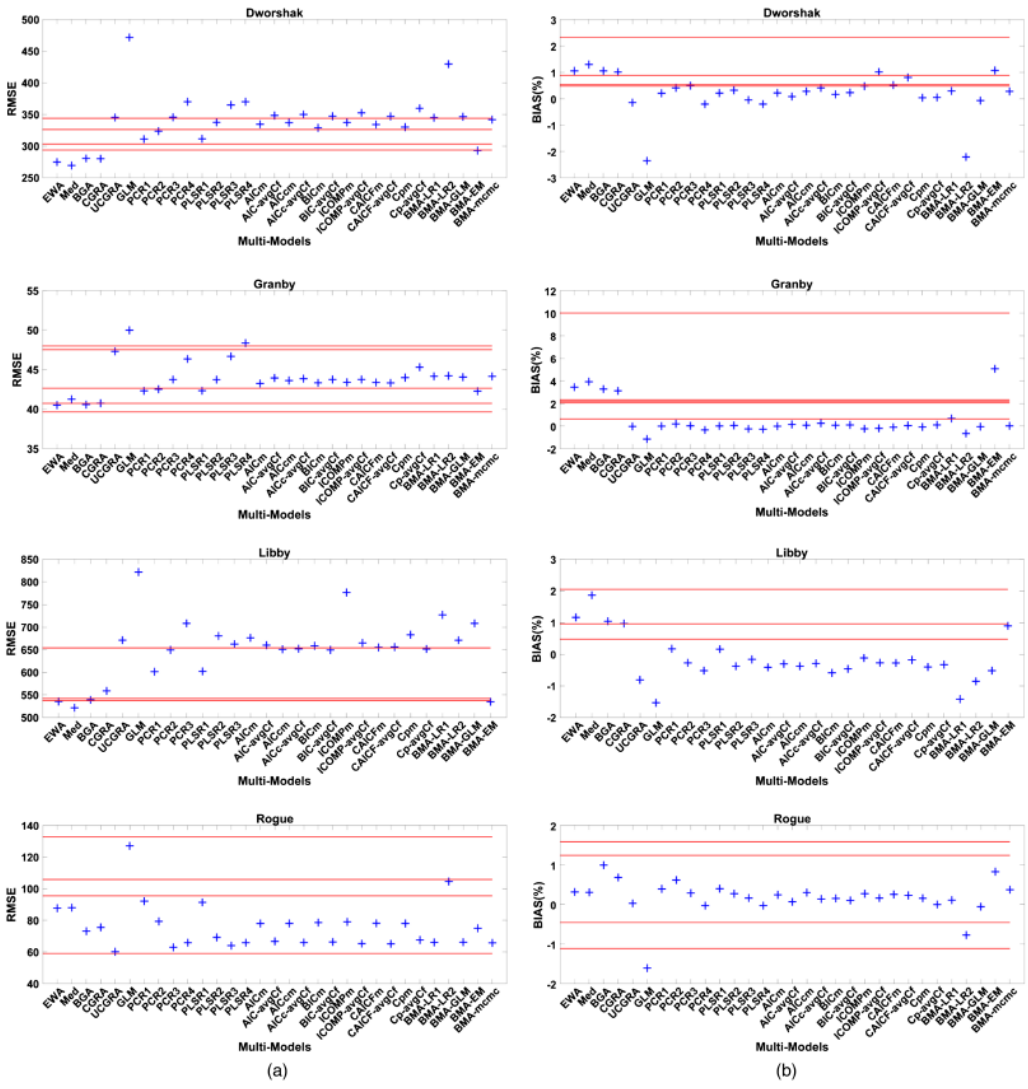

Figure 14. Comparison between model combination techniques (plus sign) and individual forecast models (horizontal lines) for each river basin based on (a) RMSE expressed in kaf; (b) percent Bias (1 kaf is approximately 1.23 million m^3)

Multi-model ensemble averaging methods showed significant improvements over the individual model forecasts for the Rogue River basin. In addition to the simple average, median, Bates-Granger, constrained linear regression and BMA-EM methods, which showed considerable improvements in the other basins, other multi-modeling approaches based on information criteria and Bayesian Model Averaging, including BMA-GLM, BMA-MCMC, and BMA-LR, outperformed most of the individual model forecasts. Several intermediate and complex models outperformed the simple multi-models as well, except for UCGRA. Here, increasing the number of principal components in PCR and PLSR improved the performance by decreasing the RMSE values. Contrary to the other basins, the individual forecasts of the Rogue basin show both positive and negative average biases, providing sufficient spread around the observed flow.

The multi-model performances depend upon the weights that are assigned to each seasonal forecast model. Figure 15 shows that the relative distribution of weights assigned by two distinct multi-modeling methods (simple vs. complex) varies between the basins. In Dworshak and Granby the weights show different patterns; for example, BMA-EM assigns a large weight on ESP model forecast in Granby basin while the Bates Granger model (BGA) assigns the lowest weight. However, in Rogue basin, both BMA and BGA assign similar weights to the individual forecasts with ICA receiving the highest.

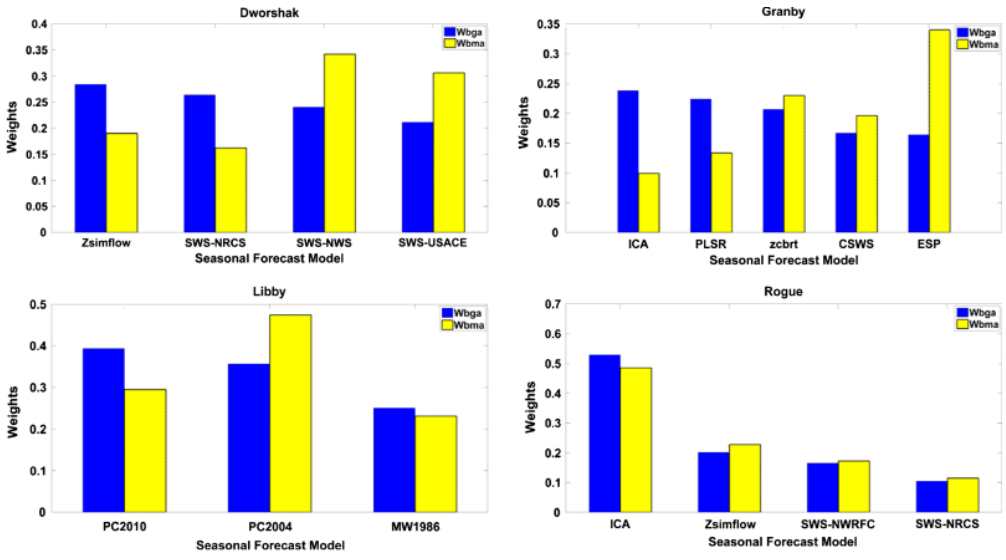

Figure 15. Weights corresponding to individual models based on Bates-Granger and Bayesian model averaging methods; individual forecasts are sorted from the best to worst for each basin

In the Bates-Granger method the individual models are weighted based on their errors in the calibration period as shown in Figure 15 (individual models are sorted from left to right). However the weights in BMA-EM reflect the probability of each model given the observed data which are obtained from maximum likelihood estimation. Results of Figure 15 shows that the weights assigned to the individual models based on BMA-EM approach do not necessarily correspond to their errors in the training period. Previous studies showed that individual models which perform well in the calibration period do not necessarily enhance the multi-model results (Viney et al. 2009), therefore assigning larger weights to best models does not guarantee the better performance of the multi-model approach.

Model Combination methods commonly provide uncertainty bounds except for simple mean and median. The 90% confidence intervals of three multi-model ensemble averaging methods with different complexities are shown over the entire periods (Figure 16). Regarding the Dworshak River basin, most observed flows lie within the ranges of combination approaches. BMA-LR and AIC show smaller uncertainty ranges compared with GLM. Highest GLM uncertainty ranges occur in 1996, 1997 and 1999. In 1984, 1985 and 1995 the observed flow is outside the confidence ranges of all multi-models. For the Granby basin BMA-LR and AIC models perform similarly and show smaller ranges of uncertainties compared to GLM. Similar to the results given for Dworshak and Granby, the GLM uncertainties are large in the Libby basin and AIC and BMA generally well capture the observed flows as in 1999. As for other basins, most of the observational values are within the ranges of the simulations with respect to the Rogue basin. The GLM again shows the largest uncertainty range; however, BMA and AIC generally outperform the GLM.

The performances of the multi-models in different categories of simple, intermediate and complex are also compared for each basin for the last ten years of the forecasts (Figure 17). Each box plot shows the ranges of the multi-model results for each category along with the parameter uncertainties from bootstrapping. Results show satisfactory performance of multi-model forecasts in general and that different models with different complexities approximately perform similarly. Generally, in circumstances where the simple models were incapable of capturing the observations, the more complex models also showed similar behaviour; examples are Dworshak-1999, Granby-2002, Libby-2000, and Libby-2002. However, it should be noted that complex models such as Bayesian

Model Averaging method provide a probabilistic approach to quantify the between model and within model uncertainties corresponding to the resulting forecasts.

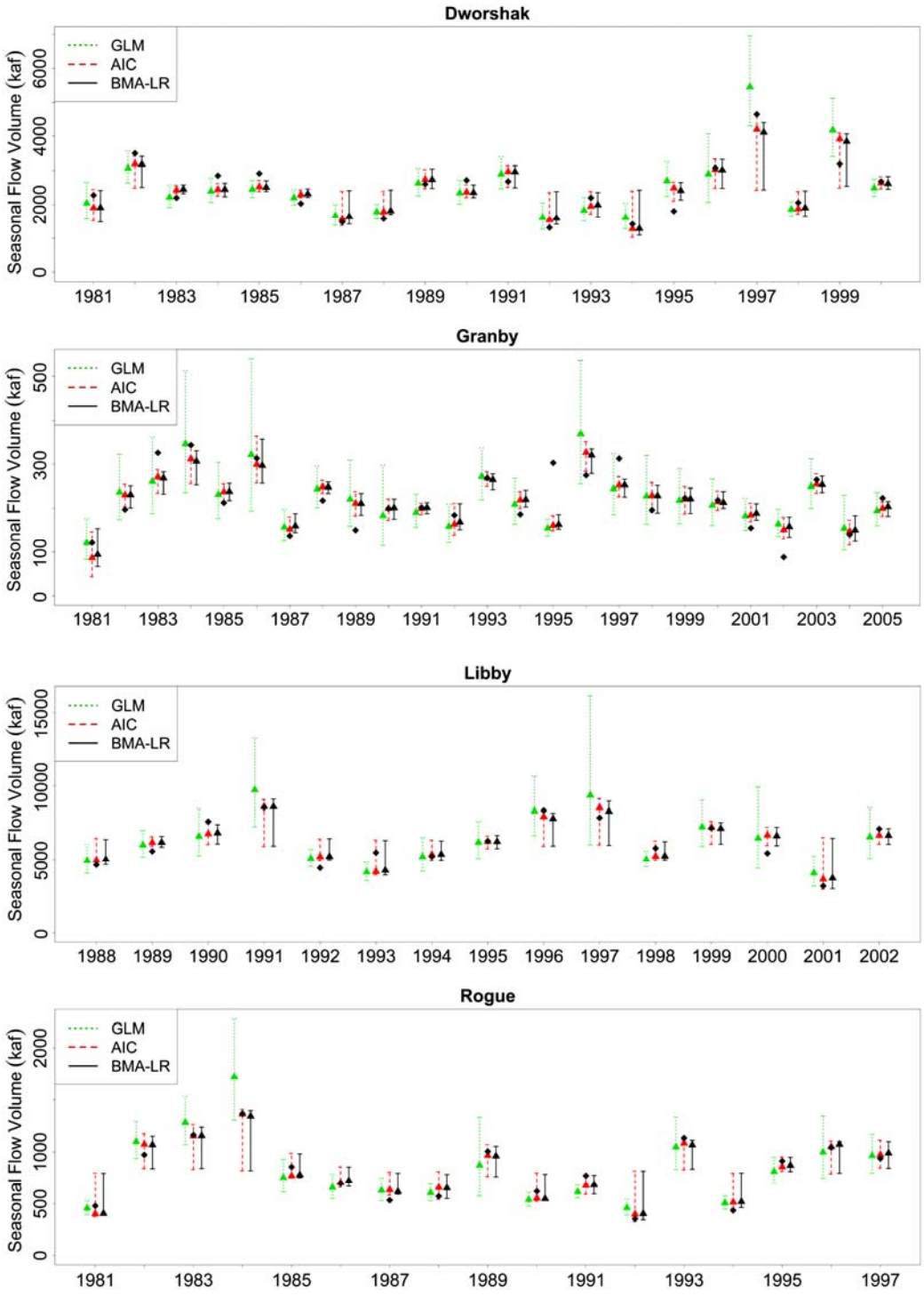

Figure 16. Ninety percent multimodel predictive bounds for each basin; observed seasonal flow is shown by a black diamond

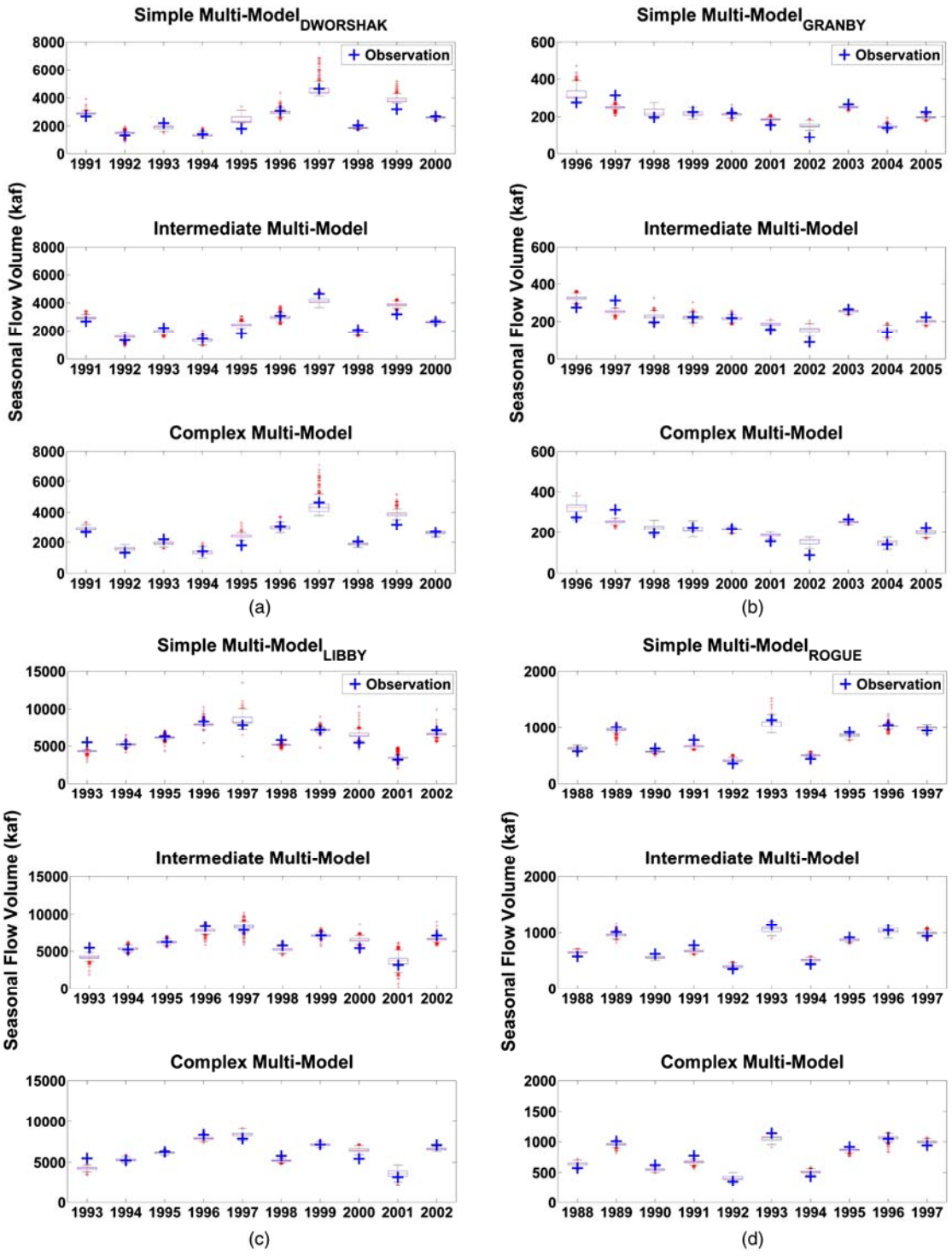

Figure 17. Comparison between the performance of multimodels with different complexities: (a) Dworshak; (b) Granby; (c) Libby; (d) Rogue

The distributions of all model combination approaches (i.e. the median forecasts from combined models along with the bootstrapping uncertainty ranges) are shown for Dworshak basin in Figure 18. Results show that individual model predictions with more spread around the observations tend to enhance the model combination estimates as in 1994, 1996 and 2000. In situations where all individual forecasts are positively or negatively biased the resulting multi-model ensemble average is incapable of accurately simulating the observations as in 1995, and 1999.

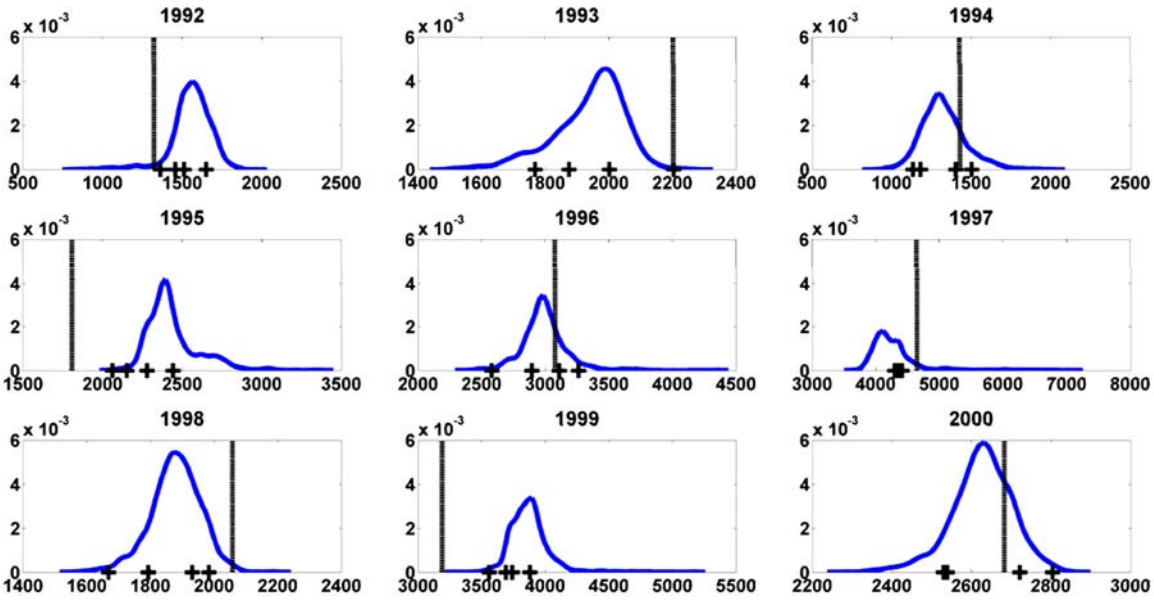

Figure 18. Comparison between model combination strategies (distributions) and individual model forecasts (plus sign) with observed flow (vertical lines) for Dworshak River basin

Increasing the number of individual models is one way to increase this spread, as shown for Granby River basin in Figure 19; however care should be taken to prevent over-fitting of model combination parameters. Considering the Libby basin (Figure 20), both proper spread of individual simulations (as in 1994) and their accuracies (1995) have positively affected the multi-modeling performance. In 1998, although multi-modeling spread is significantly wide it does not include the observation because of inaccurate forecasts and their narrow spread.

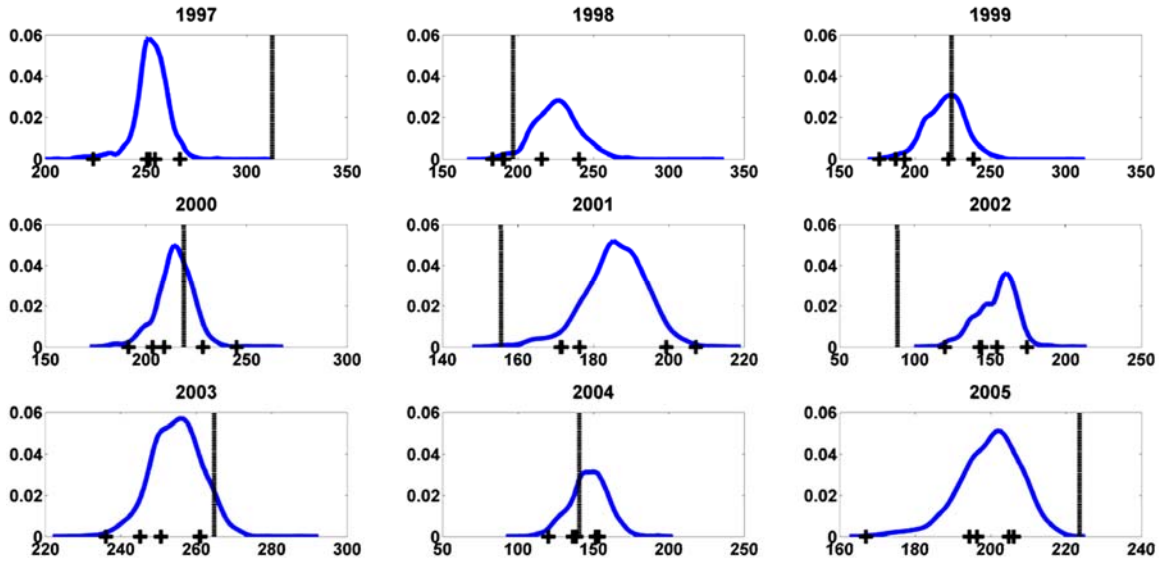

Figure 19. Comparison between model combination strategies (distributions) and individual model forecasts (plus sign) with observed flow (vertical lines) for Granby River basin

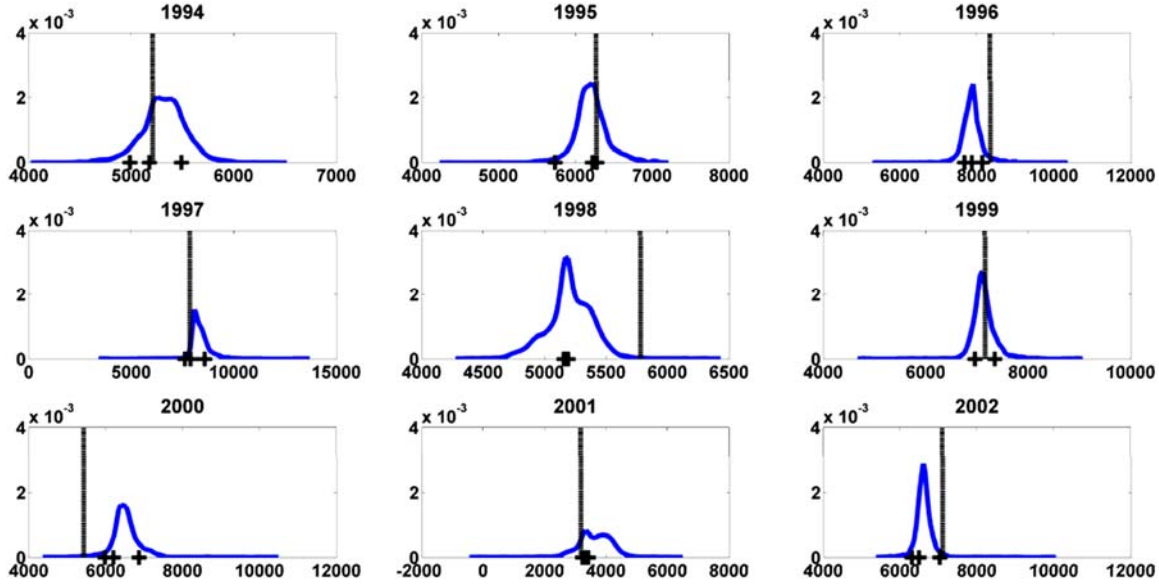

Figure 20. Comparison between model combination strategies (distributions) and individual model forecasts (plus sign) with observed flow (vertical lines) for Libby River basin.

Regarding the Rogue basin, individual forecast spread shows an important role in multi-modeling performance (Figure 21). While fewer individual models present accurate results, the wide spread between models has improved the overall performances of the multi-modeling techniques, in particular the intermediate and complex models as discussed previously. We also note that parameter uncertainty ranges arising from bootstrapping approach are small compared to variations between model averaging techniques (not shown).

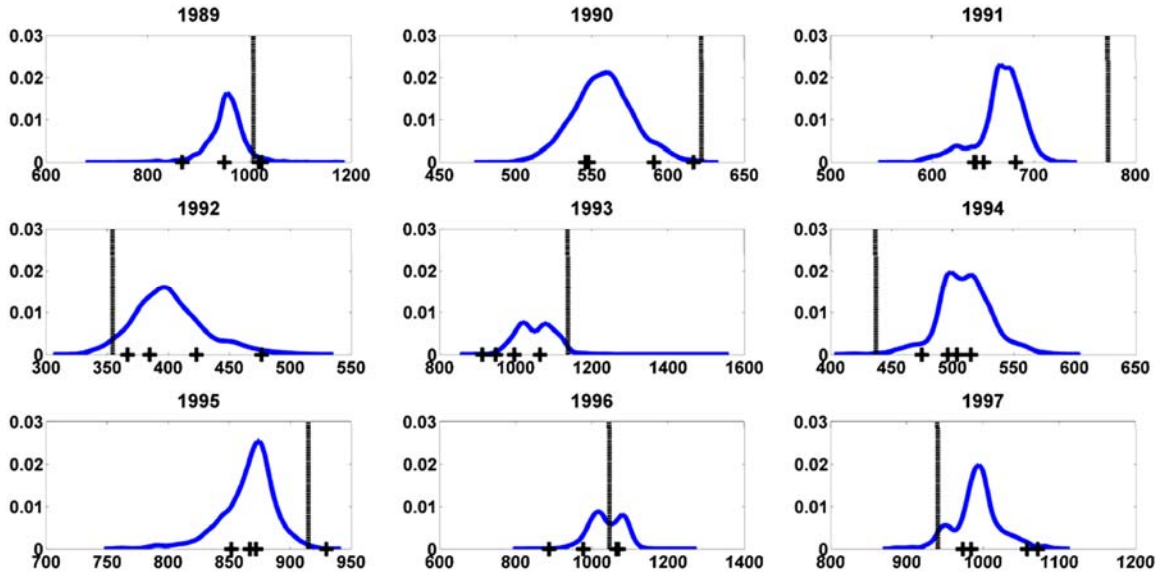

Figure 21. Comparison between model combination strategies (distributions) and individual model forecasts (plus sign) with observed flow (vertical lines) for Rogue River basin

Analyses showed that multi-model ensemble merging of seasonal forecasts issued by different agencies can be considered as an effective approach to increase the accuracy and reliability of seasonal forecasts. Performance of the multi-modeling techniques varied depending on the study region, number of individual forecast models and their performances. In three of the studied regions, where the overall biases of all the individual model forecasts were positive, the performances of more complex multi-modeling methods were inferior compared with the ones in Rogue basin in which both positive and negative forecast biases existed. The bootstrapping approach used in this study provided reliable multi-modeling assessments and quantified the parameter uncertainties.

Results of this study suggest that simple averaging techniques such as mean, median, Bates-Granger and constrained multiple linear regression along with the complex BMA-EM approach, generally outperform most of the individual model forecasts as well as other multi-model ensemble averaging techniques. In several circumstances, as in Dworshak, they also outperformed the best individual forecast model. In Rogue basin with the individual forecast having both positive and negative biases, most of the multi-model averaging techniques, in particular the intermediate and complex methods, performed satisfactorily and improved the majority of the individual model forecasts. In addition, the information based methods performed similarly to one another and were less sensitive to the choice of the penalty functions. Simple multi-modeling methods such as GLM provided larger uncertainty bounds compared with the information based methods and BMA.

Although GLM performance is weaker than the other multi-model averaging techniques and most of the individual models, it does not show a significantly poor performance. However, the GLM performance is weaker compared to the others and not recommended for multi-modeling purposes based on the results of this study (by assuming a gamma distribution with a logarithmic link function). We also note that few other multi-modeling techniques including PCR4 and PLSR4 as well as BMA-LR and methods based on information criteria might provide results that are weaker than the worst individual forecast.

Model combination of seasonal forecast models is influenced by the availability of sufficient observations. Care should be taken to avoid over-parameterization of the multi-modeling techniques. Provision of sufficiently long data series would result in better optimizations of the multi-model parameters. The results of the model combination techniques presented here also depend on choosing the widely used RMSE as the performance measure. Moreover, most of the models were calibrated based on minimization of the squared error except for BMA in which individual model weights were obtained from maximum likelihood.

H. Integration of Multimodeling Framework with FEWS/CHPS

The Flood Early Warning System (FEWS) provides a framework for deploying and passing time series information, intended for use in models utilized for flood forecasting. Within this framework, FEWS passes information across a “Published Interface”, via xml files, to allow the end-user to adapt any model to their system. This has been performed

for the National Weather Service to create CHPS with the OHDFewsAdapter (documented at ftp://hydrology.nws.noaa.gov/pub/CHPS/For_Software_Developers/). This adapter is the insertion point in CHPS, translating the Published Interface files into a framework more conducive to running Office of Hydrologic Development (OHD) models. As shown in Figure 22, the OHDFewsAdapter takes the files passed through the Published Interface, instantiates the model driver, and then the driver passes any other necessary information through xml or text files to the models itself. This allows any model, whether or not it is written in Java, to be adapted to the CHPS framework.

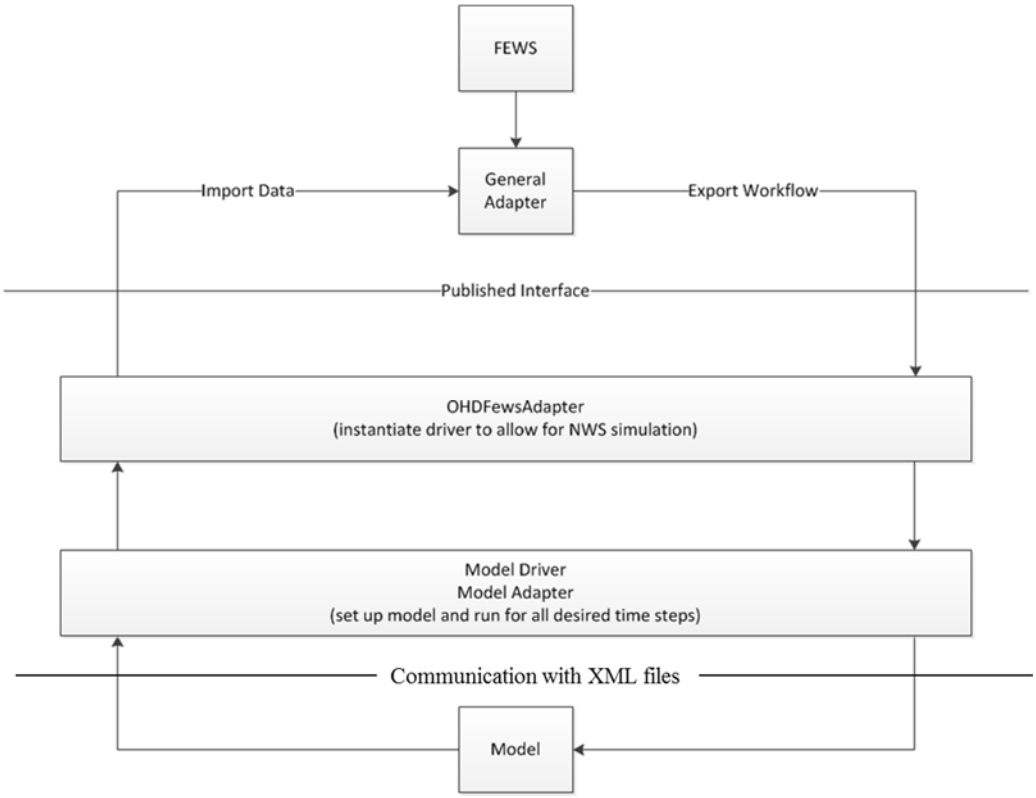

Figure 22. Flow chart of FEWS and CHPS

a. Challenges of Bringing Multi-modeling into CHPS

FEWS and CHPS are developed to run a single model simulation at a time. In order to make this system as modular, and therefore flexible, as possible, FEWS/CHPS pass much of their information through xml files. By passing information through xml files, via the FEWS published interface, a subsequent model can be adapted to an existing system, without any additional software development to FEWS or CHPS. The flow of data in the FEWS/CHPS system is shown in Figure 22. In this figure, xml files are used to pass information across the published interface, and between the OHDFewsAdapter and the actual model drivers.

Data flow for Multi-modeling which was developed as part of Data Assimilation system differs from that of FEWS. The model simulations are performed in an ensemble loop, during each time-step. This requires the model simulations to be stopped at each time-step, and subsequent models run before moving on to a further time-step. Since FEWS ships whole time-series' to a model, one at a time, updates in a sequential manner become quite challenging. If a model is to be run one step at a time, in an ensemble fashion, the computational demand will become excessive. Due to the number of xml files that need to be passed at each model run, for each ensemble member, it is infeasible to run ensemble simulations without significant software development on the CHPS side.

b. Development of Multi-modeling within CHPS

A class for running an ensemble model was developed, which is called the StocasticModel. This class houses all of the instantiated models, and will run or manipulate the Data Assimilation Driver when told to do so. Within this class, a sampling utility, named samplingUtils, will perform all of the error sampling necessary to run an ensemble simulation. It also applies the DA algorithm when an observation is provided. As of the focus in this study, this class is capable of either performing the multi-model averaging with statistic weights (Simple Model Averaging) or dynamic weights (BMA), whenever it is necessary (Figure 23).

```
//apply bayes law to estimate the weights
private void sequentialBayesLaw() {

    // loop through the ensemble members
    for (int EnsembleMember = 0; EnsembleMember < EnsembleSize; EnsembleMember++) {

        // loop through the locations in the likelihood object
        for (String locationStr : likelihood.get(EnsembleMember).keySet()) {

            // replace the weight with the produce of the likelihood and weight for the current ensemble member and location
            Weight.get(EnsembleMember).put(
                    locationStr,
                    likelihood.get(EnsembleMember).get(locationStr)
                            * Weight.get(EnsembleMember).get(locationStr));

            //estimate the sum of the weight
            if (EnsembleMember == 0) { //reset the SumWeight map with the updated weight at ensemble member 0
                SumWeight.put(locationStr,
                        Weight.get(EnsembleMember).get(locationStr));

            } else { //add the current ensemble member weight to the SumWeight map
                SumWeight.put(locationStr,
                        Weight.get(EnsembleMember).get(locationStr)
                                + SumWeight.get(locationStr));

            }

        }

    }
```

Figure 23. Implementing BMA into StochasticModel

```
<parameter id="MODEL_COMBINATIONS">
    <table>
        <columnTypes A="string" B="string" C="string"/>
        <row A="SQIN" B="SYCO3" C="BMA"/>
    </table>
</parameter>
<parameter id="MODEL_COMBINATIONS_SQIN">
    <table>
        <columnTypes A="string" B="string"/>
        <row A="HYMOD_SYCO3" B="0.5"/>
        <row A="UNITHG_SYCO3" B="0.5"/>
    </table>
</parameter>
```

Figure 24. Parameters introduced for Multi-modeling

c. Example Study

One case study is performed here to verify the utility of the CHPS Multi-modeling framework developed. This is conducted in the Johnson Creek in Northern Oregon using the observed precipitation and temperature data, estimated potential evapotranspiration data and model parameters provided by the Northwest River Forecast Center (NWRFC). Here, the simulations are performed from October 1st 1980 to September 30th 1989 providing 10 years of analysis for calculating performance metrics.

In this basin, the NWRFC routinely provides forecasts of flow at the Sycamore gaging station (SYCO3). Experiment was performed using SACSMA and HYMOD

models in SYCO3. Evaluation of BMA was done by comparing the results with the ones from multi-modeling with static weights (0.5 for each model). In this scenario, 6- hourly forecasting experiment is performed. The comparison of Simple Model Averaging (SMA) and the true streamflow is presented in Figure 25. The comparison of Bayesian Model Averaging (BMA) and the true streamflow is presented in Figure 26. Note that this figure only takes a short time window from the 10 year simulation period to make differences between the forecast and truth more visible. From Figure 25, it can be seen that the SMA is biased low, in comparison to the true streamflow. After applying the BMA (Figure 26), the forecasts are shifted towards the truth, indicating that the BMA reduces error. Further evidence of this reduction in error is provided in Table 6. This suggests that the BMA improves short-term streamflow forecasting in comparison to the SMA.

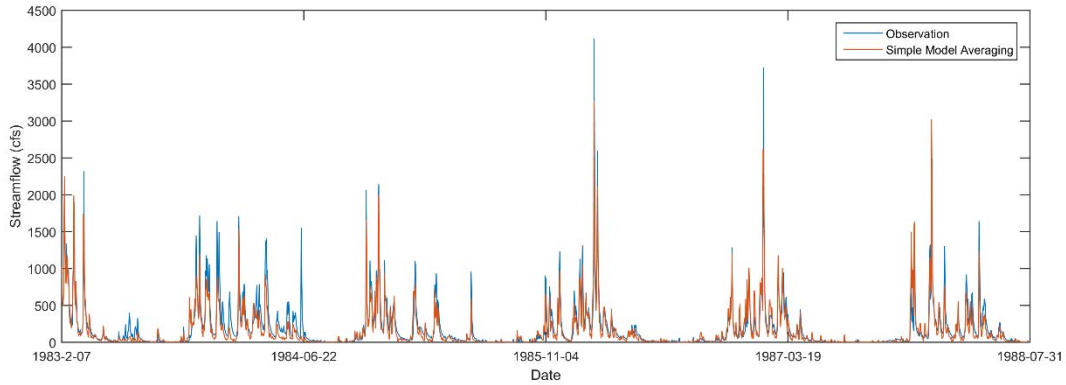

Figure 25. Streamflow forecasting experiment with Simple Model Averaging (SMA) for Johnson Creek at Sycamore (SYCO3)

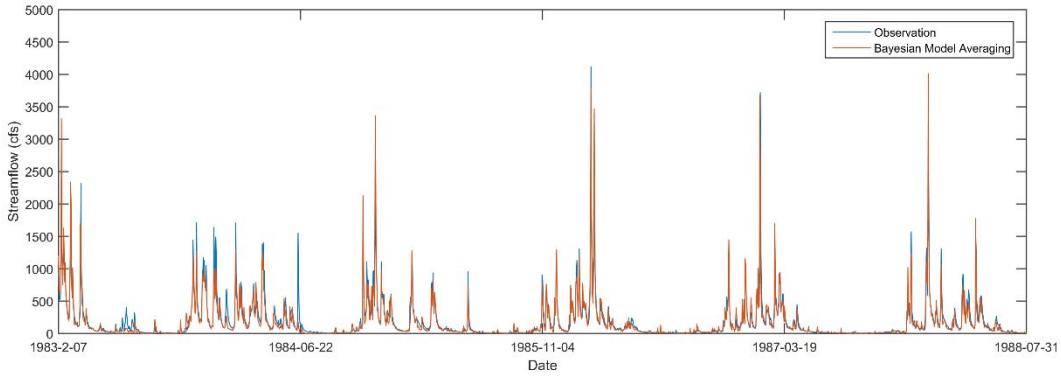

Figure 26. Streamflow forecasting experiment with Bayesian Model Averaging (BMA) for Johnson Creek at Sycamore (SYCO3)

Table 6. Comparison of performance measures from both SMA and BMA

Measures	Simple Averaging	BMA
KGE	0.71	0.79
NSE	0.65	0.77

I. Publications, Presentations, and Technology Transfer Activities

Several in person communications were conducted to increase the relevance of this work to operational forecasters. In addition, dissemination of research findings was carried out through refereed publications, invited presentations and other presentations (in particular at FEWS day User workshop organized in Portland, OR), and meetings.

a. Refereed Publications

- Khajehei, S. and H. Moradkhani (revision submitted), Towards an Improved Ensemble Precipitation Forecast: A Probabilistic Post-processing Approach, *J. of Hydrology*.

- Najafi, M.R. and H. Moradkhani (2015), Towards Ensemble Combination of Seasonal Streamflow Forecasts, *Journal of Hydrologic Engineering,* 10.1061/(ASCE)HE.1943-5584.0001250.

- Madadgar, S. and H. Moradkhani (2014), Improved Bayesian Multi-modeling: Integration of Copulas and Bayesian Model Averaging, *Water Resources Research,* 50, 9586–9603, DOI: 10.1002/2014WR015965.

- DeChant C.M., and H. Moradkhani (2014), Toward a Reliable Prediction of Seasonal Forecast Uncertainty: Addressing Model and Initial Condition Uncertainty with Ensemble Data Assimilation and Sequential Bayesian Combination, *Journal of Hydrology*, special issue on Ensemble Forecasting and data assimilation, DOI: 10.1016/j.jhydrol.2014.05.045.

- DeChant, C.M. and H. Moradkhani (2014), Hydrologic Prediction and Uncertainty Quantification, *Handbook of Engineering Hydrology*, CRC press, Taylor & Francis Group, pp. 387–414.

- Moradkhani, H., C.M. DeChant and S. Sorooshian (2012), Evolution of Ensemble Data Assimilation for Uncertainty Quantification using the Particle Filter-Markov Chain Monte Carlo Method, *Water Resources Research,*48,W12520.

- Madadgar, S., Moradkhani, H., and Garen, D. (2014), Towards Improved Post-processing of Hydrologic Forecast Ensembles, *Hydrological Processes*, 28 (1), 104-122, doi: 10.1002/hyp.9562.

- Parrish, M., Moradkhani, H., and DeChant C.M. (2012), Towards Reduction of Model Uncertainty: Integration of Bayesian Model Averaging and Data Assimilation, *Water Resources Research*, 48, W03519, doi:10.1029/2011WR011116.

- Najafi, M.R., Moradkhani, H., and Piechota, T., (2012), Ensemble Streamflow Prediction: Climate Signal Weighting vs. Climate Forecast System Reanalysis" *Journal of Hydrology*, 442 –443, 105–116.

b. Invited Presentations

- How to Enhance Hydroclimate Forecasting for Water Resources and Emergency Management, USACE, *Engineering Research and Development Center, March 2016.*
- Hydrologic Forecasting, Data Assimilation and Post-processing, FEWS user's day workshop, Portland, March 2016.
- Enhancing Seasonal Forecasting by Multi Modeling, University of Oulu, *Finland,* February 2016.
- Remotely Sensed Satellite Data Assimilation: State-of-the art of Ensemble Inference in Hydrogeophysical Applications, *University of New South Wales,* June 2015.
- Enhancing Drought Forecasting Reliability in Presence of Uncertainties, *AGU Fall Meeting*, December 2014.
- Toward a More Effective Postprocessing of Hydrologic Forecasts by Copula-Embedded Bayesian, *American Geophysical Union*, San Francisco, December 2014.

- State-of-the-Art of Ensemble Land Data Assimilation in Hydrometeorological Forecasting: the Value of Remotely-Sensed Satellite Observation, *10th International Civil Engineering Conference,* Tabriz, May 2015.
- Combined Data Assimilation and Multimodeling for Seasonal Hydrologic Forecasting- A more Complete Characterization of Uncertainty, CAHMDA/DAFOH workshop, Austin, TX, September 2014.
- Merging Hydrologic Data Assimilation with multi-modeling for operational forecasting, *Beijing Normal University*, Beijing, China, May 2014.
- Quantifying the Uncertainty in the Assessment of Climate Change Impact on Hydrologic Extremes using Hierarchical Bayesian Modeling, Society for Industrial and Applied Mathematics (SIAM), symposium on Uncertainty Quantification and Reduction in Environmental Fluids, Savannah, Georgia, March 2014.
- Assessment of Climate Change Impact on the Hydrology of the Columbia River Basin Using Multi-modeling, International Columbia Basin Climate and Hydrology Assessment Workshop, Portland, Oregon, January 2014.
- Towards Improved Ensemble Hydrologic Forecasting:Postprocessing or Data Assimilation? *University of Arizona, Department of Hydrology and Water Resources*, Tucson, Arizona, October 2013.
- The value of in-situ and Remotely Sensed Data Assimilation for Hydrometeorologic Forecasting: From Theory to Operation, *Eawag, Swiss Federal Institute of Aquatic Science and Technology*, Zurich, Switzerland, May 2013.
- Recent Advancement in Ensemble Data Assimilation for Reduction of Uncertainty in *WATGLOBE*, *Terrestrial Hydrologic Modeling, at Terrestrial Water Cycle Observation and Modeling from Space, Land Data Assimilation*, Chinese Academy of Science, Beijing, China, April 2013.
- A New Postprocessing method for improved ensemble forecasts, AGU, San Francisco, December 2012.
- The Pursuit of Hydrologic Data Assimilation: Robustness and Reliability in State and Parameter Estimation, AGU, San Francisco, December 2012.

- Combining Statistical and Ensemble Streamflow Predictions to Cope with Consensus Forecast, AGU, San Francisco, December 2012.
- Toward Improved Reliability of Seasonal Hydrologic Forecast: Accounting for Initial Condition and State-Parameter Uncertainties, AGU, San Francisco, December 2012.
- Towards Improved Reliability and Reduced Uncertainty of Hydrologic Ensemble Forecasts Using Multivariate Post-processing, 2012 International Workshop on Hydrological Ensemble Prediction Experiment, *Beijing Normal University,* Beijing, China, October, 2012.
- The Quest of Ensemble Data Assimilation: From Theory to Operation, *Hydrologic Research Center,* San Diego, October, 2012.
- Advances in Ensemble Data Assimilation With Operational Implications, 2nd International Workshop on Data Assimilation for Operational Hydrologic Forecasting and Water Resources Management, *Incheon, Korea,* September 2012.
- Ensemble Forecasting, Statistical Forecasting or Multimodeling? From Theory to Operation*, Hydrologic Research Center*, San Diego, CA, October, 2012.

c. Other Presentations

- Post-processing of multi-hydrologic model simulations for improved streamflow projections, EGU General assembly, Vienna, 2016
- Accounting for combined effect of initial condition and model uncertainty in seasonal forecasting through data assimilation, HEPEX workshop on Ensemble for better hydrological forecasts, Quebec, Canada, June 2016.
- From meteorological to hydrological postprocessing: the quest for an effective approach, HEPEX workshop on Ensemble for better hydrological forecasts, Quebec, Canada, June 2016.
- Reducing Uncertainties of Hydrologic Model Predictions Using a New Ensemble Pre-Processing Approach, AGU fall Meeting, December 2015.

- Dynamically Evolving Models for Dynamic Catchments: Application of the Locally Linear Dual EnKF to a Catchment with Land Use Change, AGU fall Meeting, December 2015.
- Comparison of Two Global Sensitivity Analysis Methods for Hydrologic Modeling over the Columbia River Basin, AGU fall Meeting, December 2015.
- Reducing Uncertainties of Hydrologic Model Predictions Using a New Ensemble Pre-Processing Approach
- Dynamically Evolving Models for Dynamic Catchments: Application of the Locally Linear Dual EnKF to a Catchment with Land Use Change
- Comparison of Two Global Sensitivity Analysis Methods for Hydrologic Modeling over the Columbia River Basin
- Evaluating the Potential Use of Remotely Sensed Soil Moisture Data for Agricultural Drought Risk Monitoring
- A Regional Bayesian Hierarchical Model for Flood Frequency Analysis In Oregon, *AGU Chapman Conference*, Portland, July 2013.
- Evaluating Multi-Modeling Techniques with Varying Complexities for Seasonal Hydrologic Forecasts, *AGU Chapman Conference*, Portland, July 2013.
- Understanding the effects of initial condition and model structural uncertainty in seasonal hydrological forecasts with data assimilation and Bayesian model averaging, *AGU Chapman Conference*, Portland, July 2013.
- A Probabilistic Framework for predicting the Spatial Variation of Future Droughts, *AGU Chapman Conference*, Portland, July 2013.
- Toward Improving the Multi-modeling Hydrologic Forecasting: Integration of Data Assimilation and Bayesian Model Averaging, at *Operational River Flow And Water Supply Forecasting*, Vancouver, British Columbia V6B 5K3, Canada , October 2011.
- Implication of Data Assimilation in Ensemble Streamflow Prediction, at *Operational River Flow And Water Supply Forecasting*, Vancouver, British Columbia V6B 5K3, Canada , October 2011.
- Volumetric Streamflow Prediction; Comparing Historical Resample vs. Climate Model Forcing Data, AGU Fall Meeting, San Francisco, CA, Dec. 2011.

- Utilizing Data Assimilation Techniques to Improve the Characterization of Initial Condition for Ensemble Streamflow Prediction, *American Geophysical Union*, San Francisco, Dec. 2011.
- Examining the Ability of Sequential Data Assimilation Methods to Accurately Quantify the Uncertainty in Hydrologic Forecasting, *American Geophysical Union*, San Francisco, Dec. 2011**.**

d. Organizing Conference and Workshop Sessions

- Panel Discussion: State-of-the-Art of Uncertainty Analysis in Hydroclimate Modeling, World Water Congress, Florida, May 2016.
- From hydroclimate forecasting to water resources decision-making, *AGU Fall Meeting*, December 2016.
- Hydroclimatic Extremes: Drought, *AGU Fall Meeting*, December 2016
- Advances in Hydrometeorological Extremes Forecasting: Estimation, Integrated Risk Analysis, and Applications, *AGU Fall Meeting*, December 2015
- Hydroclimatic Extremes: Drought, *AGU Fall Meeting*, December 2015.
- Guest Editor: Special issue as guest editor in Journal of Hydrology on "Hydrologic Ensemble Prediction and Data Assimilation for Operational Hydrology and Water Resources Management
- Chaired the technical program committee of *AGU Chapman Conference*, July 2013, Portland State University, Portland
- Co-organized several sessions for *AGU fall* meeting, San-Francisco, December 2013. Topic: Advances in Hydrometeorological Predictions and Applications
- Organized a half-day meeting with project partners (mainly NWRFC and NRCS staff) to go over the progress made on the project and discuss the data and modeling limitations. June 2013.
- Co-organized and Chaired: Advances in Hydrometeorological Predictions and Applications, American Geophysical Union, December 2012.
- Co-Chaired- Land Data Assimilation Session, WATGLOBE Workshop, Chinese Academy of Science, Beijing, China, April 2013.

- Developed several communications with OHD about having CHPS training for incorporation of our methods. We would have liked to have the training at OHD but certain complications on the security issues did not let the training be conducted at OHD. Therefore, we have now planned the webinar training tentatively planned by CHPS staff in January.
- Co-organized the 2nd International Workshop on Data Assimilation for Operational Hydrologic Forecasting and Water Resources Management, *Incheon, Korea,* September 2012.

e. Theses and Dissertations

- Sepideh Khajehei (M.S. 2015), Multivariate Method for Generating Ensemble Climatologic Forcing Data for Hydrologic Applications, Portland State University.
- Caleb DeChant (Ph.D. 2014), *Assessing the Impacts of Physically-Based Land Surface Water Storage Estimation on Hydrological Droughts*, College of Engineering Dean's list/Outstanding Doctoral Student, Portland State University.
- Shahrbanou Madadgar (Ph.D. 2014), *Towards Improving Seasonal Drought Prediction under Hydroclimate Uncertainties*, First Female Doctoral Graduate Student in Civil & Environmental Engineering, Portland State University.
- Mohammad Reza Najafi (Ph.D 2013), Climate Change Impact on the Spatio-Temporal Variability of Hydro-Climate Extremes by Means of Bayesian Hierarchical Modeling, Outstanding Graduate Student of the Year, Portland State University.

J. Personnel

Student Researchers who were supported by CSTAR funds:

Sepideh Khajehei (Graduate Student, M.S. completed, Ph.D. student)

Ali Ahmadalipour (Ph.D. Student)

Hongxiang Yan (Ph.D. Student)

Mahkameh Zarekarizi (Ph.D. Student)

Shahrbanou Madadgar (Graduated, Ph.D.)

Caleb DeChant (Graduated, Ph.D.)
Mohammad Reza Najafi (Graduated, Ph.D.)
Golnaz Mirfendereski (Graduate Student, Ph.D.)

K. Reference

All the references are included in the “refereed publications” section above.