Multimodel Combination Techniques for Analysis of Hydrological Simulations: Application to Distributed Model Intercomparison Project Results

NEWSHA K. AJAMI

University of California, Irvine, Irvine, California

QINGYUN DUAN

Lawrence Livermore National Laboratory, Livermore, California

XIAOGANG GAO AND SOROOSH SOROOSHIAN

University of California, Irvine, Irvine, California

(Manuscript received 15 April 2005, in final form 17 November 2005)

ABSTRACT

This paper examines several multimodel combination techniques that are used for streamflow forecasting: the simple model average (SMA), the multimodel superensemble (MMSE), modified multimodel superensemble (M3SE), and the weighted average method (WAM). These model combination techniques were evaluated using the results from the Distributed Model Intercomparison Project (DMIP), an international project sponsored by the National Weather Service (NWS) Office of Hydrologic Development (OHD). All of the multimodel combination results were obtained using uncalibrated DMIP model simulations and were compared against the best-uncalibrated as well as the best-calibrated individual model results. The purpose of this study is to understand how different combination techniques affect the accuracy levels of the multimodel simulations. This study revealed that the multimodel simulations obtained from uncalibrated single-model simulations are generally better than any single-member model simulations, even the best-calibrated single-model simulations. Furthermore, more sophisticated multimodel combination techniques that incorporated bias correction step work better than simple multimodel average simulations or multimodel simulations without bias correction.

1. Introduction

Many hydrologists have been working to develop new hydrologic models or to try improving the existing ones. Consequently, a plethora of hydrologic models are in existence today, with many more likely to emerge in the future (Singh 1995; Singh and Frevert 2002a,b). With the advancement of the geographic information system (GIS), a class of models, known as distributed hydrologic models, has become popular (Russo et al. 1994; Vieux 2001; Ajami et al. 2004). These models explicitly account for spatial variations in topography, meteorological inputs, and water movement.

The National Weather Service Hydrology Laboratory (NWS-HL) has recently conducted the Distributed Model Intercomparison Project (DMIP; http://www.nws.noaa.gov/oh/hrl/dmip), which showcased state-of-the-art distributed hydrologic models from different modeling groups (Smith et al. 2004). It was found that there is a large disparity in the performance of the DMIP models (Reed et al. 2004). The more interesting findings were that multimodel ensemble averages perform better than any single-model simulations, including the best-calibrated single-model simulations, and the multimodel ensemble averages are more skillful and reliable than the single-model ensemble averages (Georgakakos et al. 2004). Georgakakos et al. (2004) attributed the superior skill of the multimodel ensembles to the fact that model structural uncertainty is accounted for in the multimodel approach. They went on to suggest that multimodel ensemble simulations should be considered as an operational forecasting tool. The fact that the simple multimodel averaging ap-

Corresponding author address: Newsha K. Ajami, Dept. of Civil and Environmental Engineering, University of California, Irvine, E/4130 Engineering Gateway, Irvine, CA 82697-2175.
E-mail: nkhodata@uci.edu

proach such as the one used by Georgakakos et al. (2004) has led to more skillful and reliable simulations motivated us to examine whether more sophisticated multimodel combination techniques can result in consensus simulations of even better skill.

Most hydrologists are used to the traditional contructionist approach, in which the goal of the modeler is to build a perfect model that can capture the real world processes as much as possible. The evolution of simple conceptual models to more physically based distributed models is a manifestation of this approach. The multimodel combination approach, on the other hand, works in essentially a different paradigm in which the modeler aims to extract as much information as possible from the existing models. The rationale behind multimodel combination lies in the fact that predictions from individual models invariably contain errors from various sources, including model input data, model state and parameter estimation, and model structural deficiencies (Beven and Freer 2001). With independently constructed models, these errors tend to be mutually independent in statistical property. Through model averaging, these errors would act to cancel each other out, resulting in better overall predictions.

The idea of combining predictions from multiple models was explored more than 30 years ago in econometrics and statistics (see Bates and Granger 1969; Dickinson 1973, 1975; Newbold and Granger 1974). Thompson (1976) applied the model combination concept in weather forecasting. He showed that the mean square error of forecasts generated by combining two independent model outputs is less than that of the individual predictions. Based on the study done by Clemen (1989), the concept of the combination forecasts from different models were applied in diverse fields ranging from management to weather prediction. Fraedrich and Smith (1989) presented a linear regression technique to combine two statistical forecast methods for long-range forecasting of the monthly tropical Pacific sea surface temperatures (SSTs). Krishnamurti et al. (1999) explored the model combination technique by using a number of forecasts from a selection of different weather and climate models. They called their technique multimodel superensemble (MMSE) and compared it to the simple model average (SMA) method. Krishnamurti and his group applied the MMSE technique to forecast various weather and climatological variables (e.g., precipitation, tropical cyclones, seasonal climate) and all of these studies agreed that consensus forecast outperforms any single-member model as well as the SMA technique (e.g., Krishnamurti et al. 1999, 2000a,b, 2001, 2002; Mayers et al. 2001; Yun et al. 2003). Kharin and Zwiers (2002) reported that for small sample size data the MMSE does not perform as well as simple ensemble mean or the regression-improved ensemble mean.

Shamseldin et al. (1997) first applied the model combination technique in the context of rainfall–runoff modeling. They studied three methods of combining model outputs, the SMA method, the weighted average method (WAM), and the artificial neural network (ANN) method. They applied these methods to combine outputs of five rainfall–runoff models for 11 watersheds. For all these cases they reported that the model combination simulation is superior to that of any single-model simulations. Later Shamseldin and O'Connor (1999) developed a real-time model output combination method (RTMOCM), based on the synthesis of the linear transfer function model (LTFM) and the WAM and tested it using three rainfall–runoff models on five watersheds. Their results indicated that the combined streamflow forecasts produced by RTMOCM were superior to those from the individual rainfall–runoff models. Xiong et al. (2001) refined the RTMOCM method by introducing the concept of Takagi–Sugeno fuzzy system as a new combination technique. Abrahart and See (2002) compared six different model combination techniques: the SMA; a probabilistic method in which the best model from the last time step is used to create the current forecast; two different neural network operations; and two different soft computing methodologies. They found that neural network combination techniques perform the best for a stable hydroclimate regime, while the fuzzy probabilistic mechanism generates superior outputs for a more volatile environment (flashier catchments with extreme events). Butts et al. (2004a,b) proposed a framework that allowed a variety of alternative distributed hydrological models (model structures) to be used to generate multimodel ensembles. They found exploring different model structures and using SMA or WMA multimodel combinations very beneficial as part of the overall modeling approach for operational hydrological prediction since it decreases model structural uncertainty.

This paper extends the work of Georgakakos et al. (2004) and Shamseldin et al. (1997) by examining several multimodel combination techniques, including SMA, MMSE, WAM, and modified multimodel average (M3SE) a variant of MMSE. As in Georgakakos et al. (2004), we will use the simulation results from the DMIP to evaluate various multimodel combination techniques. Through this study, we would like to answer this basic question, "Does it matter which multimodel combination techniques are used to obtain consensus simulation?" We will also investigate how the

accuracy of the multimodel simulations are influenced by different factors, including 1) the seasonal variations of hydrological processes, 2) number of independent models considered, and 3) accuracy levels of individual member models.

The paper is organized as follows: Section 2 overviews different model combination techniques. Section 3 describes the data used in this study. Section 4 presents the results and analysis. Section 5 provides a summary of major lessons and conclusions.

2. A brief description of the multimodel combination techniques

a. Simple model average

The SMA method is the multimodel ensemble technique used by Georgakakos et al. (2004). This is the simplest technique and is used as a benchmark for evaluating more sophisticated techniques in this work. SMA can be expressed by the following equation:

$$(Q_{\mathrm{SMA}})_t = \overline{Q}_{\mathrm{obs}} + \sum_{i=1}^{N} \frac{(Q_{\mathrm{sim}})_{i,t} - (\overline{Q}_{\mathrm{sim}})_i}{N}, \qquad (1)$$

where $(Q_{\mathrm{SMA}})_t$ is the multimodel streamflow simulation obtained through SMA at time t, $(Q_{\mathrm{sim}})_{i,t}$ is the ith model streamflow simulation for time t, $(\overline{Q}_{\mathrm{sim}})_i$ is the time average of the ith model streamflow simulation, $(\overline{Q}_{\mathrm{obs}})$ is the corresponding observed average streamflow, and N is the number of models under consideration.

b. Weighted average method

WAM is one of the model combination techniques specifically developed for rainfall–runoff modeling by Shamseldin et al. (1997). This method utilizes the multiple linear regression (MLR) technique to combine the model simulations. The model weights are constrained to be always positive and to sum to unity. If we have model simulations from N models, WAM can be expressed as

$$(Q_{\mathrm{WAM}})_t = \sum_{i=1}^{N} x_i (Q_{\mathrm{sim}})_{i,t} \qquad (2)$$

$$\text{S.t.} \begin{cases} x_i > 0 \\ \sum x_i = 1 \quad i = 1 \ldots N, \end{cases}$$

where $(Q_{\mathrm{WAM}})_t$ is the multimodel simulation obtained through WAM at time t. Equation (2) presents a simple multiple linear regression. The multiregression method is a tool for exploiting linear tendencies that may exist between dependent variable (here observed streamflow) and a set of independent variables (the simulated streamflow by various models contributing in multimodel ensemble). Shamseldin et al. (1997) used the constrained least squares technique to solve this multiple linear regression equation and estimate the weights. In the constrained least squares technique, the weights are restrained to be positive and to sum to unity. The available dataset is divided into two periods: training and validation. Over the training period the weights are estimated for the each model contributing in the multimodel combination. Subsequently the estimated weights are tested over the validation period. For more details about this method the reader should refer to Shamseldin et al. (1997).

c. Multimodel superensemble

The MMSE is a multimodel forecasting approach popular in meteorological forecasting. Here we apply this approach for hydrological forecasting. The MMSE uses the following logic (Krishnamurti et al. 2000b):

$$(Q_{\mathrm{MMSE}})_t = \overline{Q}_{\mathrm{obs}} + \sum_{i=1}^{N} x_i[(Q_{\mathrm{sim}})_{i,t} - (\overline{Q}_{\mathrm{sim}})_i], \qquad (3)$$

where $(Q_{\mathrm{MMSE}})_t$ is the multimodel streamflow simulation obtained through MMSE at time t, and $\{x_i, \mathrm{i} = 1, 2, \ldots, N\}$ are the regression coefficients (weights) computed over the training period. The weights (regression coefficients) are estimated through the unconstrained least squares technique where they can be assigned to any real numbers. As in the WAM multimodel combination technique, weights are estimated over the training period and validated over the forecast period.

Equation (3) comprises two main terms. The first term, $(\overline{Q}_{\mathrm{obs}})$, which replaces the MMSE simulation average streamflow with the observed average streamflow, serves to reduce the forecast bias. The second term, $\Sigma x_i[(Q_{\mathrm{sim}})_{i,t} - (\overline{Q}_{\mathrm{sim}})_i]$, reduces the variance of the combination of simulations using multiple regressions. Therefore, the logic behind this methodology is a simple idea of bias correction along with variance reduction. We also note that when a multimodel combination technique such as MMSE is used to predict hydrologic variables like streamflow, it is important that the average streamflow during the training period over which the model weights are computed is close to the average streamflow of the validation period (i.e., the stationarity assumption). In section 4, we will show that bias removal and stationarity assumption are important factors in multimodel simulation accuracy.

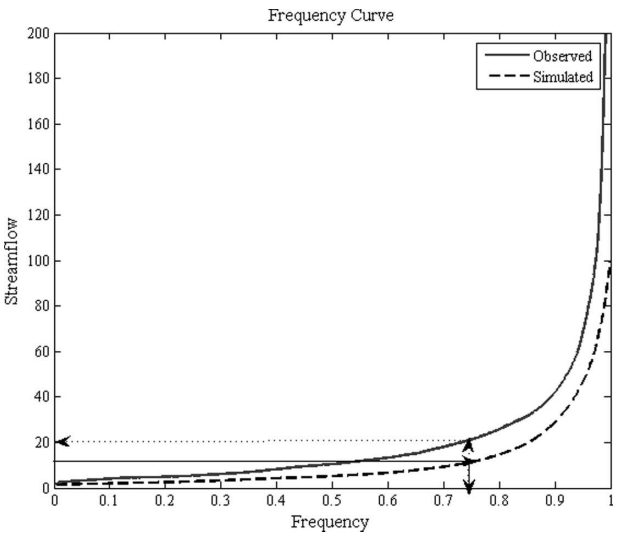

FIG. 1. Frequency curve that is being used for bias correction for the M3SE method.

d. Modified multimodel superensemble

The M3SE technique is a variant of the MMSE. This technique is similar to the MMSE except for the bias correction step. In the MMSE, model bias is removed by replacing the average of the simulations by the average of observed streamflow. In the M3SE, the bias is removed during the training period by mapping the model simulation at each time step to the observed streamflow with the same frequency as the simulated streamflow. The observed streamflow during the training period was also used later during the validation period in order to remove bias from multimodel ensemble flow simulations. Figure 1 illustrates how simulated streamflow is mapped into observed streamflow through frequency mapping. The solid arrow shows the original value of the simulation and the dashed arrow points to the corresponding observed value. For example, in the frequency curve presented in Fig. 1, a simulated streamflow value equal to 10 cms would map to a frequency of 0.74. However, the frequency value of 0.74 in the observed frequency curve maps to a value of 20 cms. Therefore, 10 cms in the simulated time series will be replaced by 20 cms, which was mapped from the observed streamflow frequency curve.

The frequency mapping bias correction method has been popular in hydrology because the bias-corrected hydrologic variables agree well statistically with the observations, while the bias correction procedure used in MMSE might lead to unrealistic values (i.e., negative values). This will be discussed later in the paper. After removing bias from each model forecast, the same solution procedure for MMSE is applied to M3SE.

e. Differences between the four multimodel combination techniques

The major differences between these multimodel combination methods are the model weighting schemes and the bias removal schemes. MMSE, M3SE, and WAM have variable model weights, while SMA has equal model weights. MMSE and M3SE compute the model weights through multiple linear regressions while WAM computes the model weights using a constrained least squares approach that ensures positive model weights and total weights equal to 1. With respect to bias correction, MMSE and SMA remove the bias by replacing the simulation mean with the observed mean, while WAM does not incorporate any bias correction. M3SE removes the bias by using frequency mapping method as illustrated in section 2d.

3. The study basins and data

We have chosen to evaluate the multimodel combination methods using model simulation outputs collected from the DMIP (Smith et al. 2004). The DMIP was conducted over several river basins within the Arkansas–Red River basin. Five of the DMIP basins are included in this study: Illinois River basin at Watts, Oklahoma; Illinois River basin at Eldon, Oklahoma; Illinois River basin at Tahlequah, Oklahoma; Blue River basin at Blue, Oklahoma; and Elk River basin at Tiff City, Missouri. Figure 2 shows the location of the basins while Table 1 lists the basin topographic and climate information. Silty clay is the dominant soil texture type of those basins, except for the Blue River, where the dominant soil texture is clay. The land cover of these basins is mostly dominated by broadleaf forest and agriculture crops (Smith et al. 2004).

The average maximum and minimum surface air temperature in the region are approximately 22° and 9°C, respectively. Summer maximum temperatures can get as high as 38°C, and freezing temperatures occur generally in December through February. The climatological annual average precipitation of the five basins in the region is between 1010 and 1160 mm yr^{-1} (Smith et al. 2004).

Seven different modeling groups contributed to the DMIP by producing hourly streamflow simulations for the DMIP basins using their distributed models, driven by meteorological forcing data provided by the NWSHL. The hourly precipitation data, available at 4×4 km^2 spatial resolution, was generated from the NWS Next-Generation Weather Radar (NEXRAD). Other meteorological forcing data such as air temperature, downward solar radiation, humidity, and wind speed

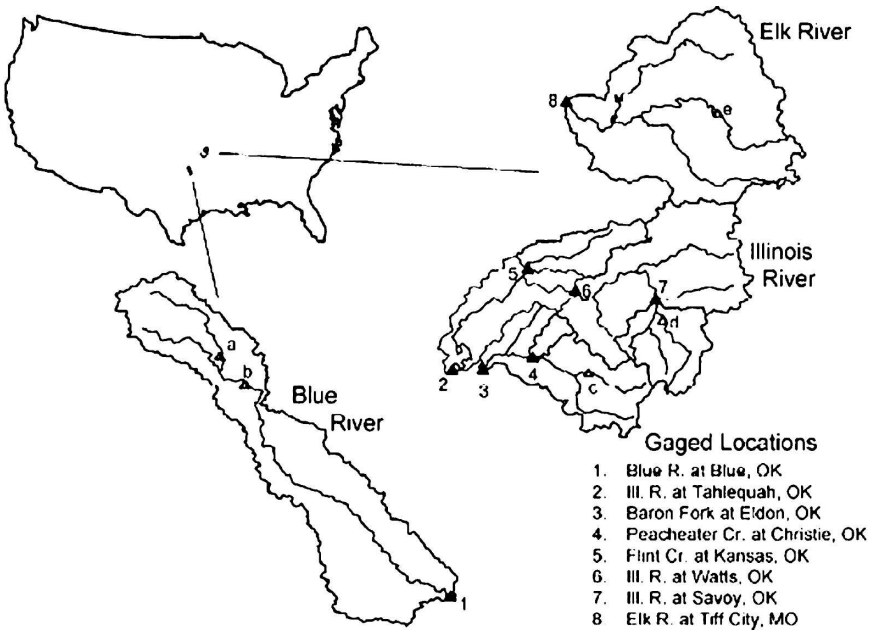

FIG. 2. DMIP test basins (Smith et al. 2004).

were obtained from the University of Washington (Maurer et al. 2001). Table 2 lists the participating groups and models. For more details on model description and simulation results, readers should refer to Reed et al. (2004).

For this study, we obtained the hourly streamflow simulations from all participating models for the entire DMIP study period: 1993–99. All the participating models were run in hourly time steps. The uncalibrated streamflow simulation results were used for multimodel combination study. Hourly observed streamflow data, along with the best-calibrated single-model streamflow simulations from the DMIP, were used as the benchmarks for comparing accuracy levels of the different multimodel simulations. The data period from 1993 to 1996 was used to train the model weights, while the rest of the data period (1997–99) was used for validating the consistency of the multimodel simulations using these weights.

4. Multimodel combination results and analysis

a. Model evaluation criteria

Before we present the results, it should be noted that two different statistical criteria are used to compare the results of this work: hourly root-mean-square error (HRMS) and the Pearson correlation coefficient (R). These two statistical criteria are commonly used in the field of hydrology to compare accuracy of different time series in matching observed variables (Hogue et al. 2000; Ajami et al. 2004). These criteria are defined as follows:

$$\mathrm{HRMS} = \sqrt{\left\{\frac{1}{n}\sum_{t=1}^{n}[(Q_{\mathrm{sim}})_t - (Q_{\mathrm{obs}})_t]^2\right\}}, \tag{4}$$

$$R = \frac{\sum_{t=1}^{n}[(Q_{obs})_t(Q_{\mathrm{sim}})_t] - [n\overline{Q}_{\mathrm{obs}}\overline{Q}_{\mathrm{sim}}]}{\sqrt{\left[\sum_{t=1}^{n}(Q_{\mathrm{obs}})_t^2 - n(\overline{Q}_{\mathrm{obs}})^2\right]\left[\sum_{t=1}^{n}(Q_{\mathrm{sim}})_t^2 - n(\overline{Q}_{\mathrm{sim}})^2\right]}}, \tag{5}$$

where n is the number of data points.

TABLE 1. Basin information.

Basin name	U.S. Geological Survey (USGS) gauge location		Area (km^2)	Annual rainfall (mm)	Annual runoff (mm)	Dominant soil texture	Vegetation cover
	Lat	Lon					
Illinois River basin at Eldon	35°55′16″	94°50′18″	795	1175	340	Silty clay	Broadleaf forest
Blue River basin at Blue	33°59′49″	96°14′27″	1233	1036	176	Clay	Woody savannah
Illinois River basin at Watts	36°07′48″	94°34′19″	1645	1160	302	Silty clay	Broadleaf forest
Elk River basin at Tiff City	36°37′53″	94°35′12″	2251	1120	286	Silty clay	Broadleaf forest
Illinois River basin at Tahlequah	35°55′22″	94°55′24″	2484	1157	300	Silty clay	Broadleaf forest

b. Comparison of the multimodel consensus predictions and the uncalibrated individual model predictions

In the first set of numerical experiments, the multimodel simulations were computed from the uncalibrated individual model simulations using the different multimodel combination techniques described in section 2. Figures 3a–j present the scatterplots of the HRMS versus R values of the individual model simulations and those of the SMA simulations. The horizontal axis in these figures denotes the Pearson coefficient from the individual models and SMA, while the vertical axis denotes HRMS of these simulations. Note that the most desired accuracy value set in matching observed streamflow is located at the lower-right corner of the figures. Figures 3a, 3c, 3e, 3g, and 3i show the results for the training period, while Figs. 3b, 3d, 3f, 3h, and 3j show the results over validation period. These figures clearly show that the statistics from the individual model simulations are almost always worse than those of the SMA simulations. These results confirm the fact that just simply averaging the individual model simulations would lead to improved accuracy levels, which is consistent with the conclusions from the paper by Georgakakos et al. (2004).

Figures 4a–j show the scatterplots of the HRMS and R for all multimodel combination techniques as well as for the best-uncalibrated and the best-calibrated individual model simulations during the training and validation periods. Clearly shown in these figures is that all multimodel simulations have superior performance statistics compared to the best-uncalibrated individual model simulation (best-uncal). More interestingly, the multimodel simulations generated by MMSE and M3SE show noticeably better performance statistics than those by SMA. This implies that there are benefits in using more sophisticated multimodel combination techniques. The simulations generated by WAM show worse performance statistics than the simulations generated by other multimodel combination techniques. This suggests that the bias removal step incorporated by other multimodel combination techniques is important in improving simulation accuracy especially during the validation period. It was found that reducing the variance solely improves the *R* and HRMS compared to the best-performing uncalibrated member model between 3%–12% and 13%–30% over the training period and 0%–3% and 8%–16% over the validation period, respectively. Adding the bias removal step to the procedure improves the *R* and HRMS compared to the combination methods with just variance reduction step (WAM combination technique), between 2%–4% and 10%–16% over the training period and 0%–5% and 0%–10% over the validation period. These results highlight two interesting observations. First, the bias removal step improves the HRMS statistics more significantly than R. The second observation is that the major progress during model combination methods happens over the variance reduction step, even though it is hard to disregard the improvement gained during the bias removal step. It is noteworthy that adding the bias removal step to the multimodel combination technique does not significantly increase the complexity and computation time of the combination process. Figure 5 depicts an excerpt of streamflow simulation results from M3SE and MMSE during the forecast period. The advantage of the bias removal technique in the M3SE over that of the MMSE is indicated by the fact that negative streamflow values were generated by the MMSE for some parts of the hydrograph (over the low flow periods) while the M3SE does not suffer from this problem.

The advantage of multimodel simulations from the training period carries into the validation period in almost all cases except for Blue River basin, where the performance statistics of the multimodel simulations are equal to or slightly worse than the best-uncalibrated individual model simulation. The reason for the relative poor performance in Blue River basin is that a noticeable change in streamflow characteristics is observed from the training period to the validation period (i.e., the average streamflow changes from 10.8 cms in the

TABLE 2. DMIP participant modeling groups and characteristics of their distributed hydrological models (Reed et al. 2004).

Participant	Model	Primary application	Spatial unit for rainfall–runoff calculation	Rainfall–runoff scheme	Channel routing scheme
Agricultural Research Services (ARS)	SWAT	Land management/agricultural	Hydrologic Response Unit (HRU)	Multilayer soil water balance	Muskingum or variable storage
University of Arizona (ARZ)	SAC-SMA	Streamflow forecasting	Subbasins	SAC-SMA	Kinematic wave
Environmental Modeling Center (EMC)	Noah land surface model	Land–atmosphere interactions	1/8° grids	Multilayer soil water and energy balance	—
Hydrologic Research Center (HRC)	HRCDHM	Streamflow forecasting	Subbasins	SAC-SMA	Kinematic wave
Office of Hydrologic Development (OHD)	HL-RMS	Streamflow forecasting	16 km^2 grid cells	SAC-SMA	Kinematic wave
Utah State University (UTS)	TOPNET	Streamflow forecasting	Subbasins	TOPMODEL	—
University of Waterloo (UWO)	WATFLOOD	Streamflow forecasting	1-km grid		Linear storage routing

training period to 7.17 cms in the validation period; standard deviation from 27.6 to 16.8 cms). This might be the indication that the stationarity assumption for streamflow was violated. Consequently, the accuracy levels of the simulations during validation period were adversely affected. Future work could include a more diagnostic analysis of the data to identify the causes for the poor validation results. If the stationarity assumption holds, the mean and variance of the streamflow from one period to another should be similar or very close (the closeness of these values is a subjective judgment made by the modeler or forecaster). Therefore the accuracy level during the validation period will not deteriorate significantly. To use this technique in the operational mode so as to decrease the deterioration in the forecast, the forecaster (modeler) should constantly compare the mean and variance of recently available real-time observations against the mean of the historical observations for the same period. Statistical measures could be included in the procedure to identify when the mean and variance of the new observations are such that the condition of data stationarity is violated. In some cases the modeler may decide to use just some specific years to remove the bias and train the multimodel scheme to facilitate a more accurate real-time forecast (e.g., if the current year seems to be a wet year, the modeler could use historical data from other wet years).

To get a measure of how multimodel simulations fare against the best-calibrated single-model simulations, we also included them in Figs. 4a–j. As revealed in these figures, MMSE and M3SE outperform the best-calibrated models (best-cal) for all the basins except Blue River basin during the training period. During validation period, however, the best-calibrated single-model simulations have shown a slight advantage in performance statistics over the multimodel simulations. MMSE and M3SE are shown to be the best-performing combination technique during validation period and have statistics comparable to those of the best-calibrated case, while WAM and SMA have worse performance statistics.

c. Application of multimodel combination techniques to river flow predictions from individual months

Hydrological variables such as streamflow are known to have a distinct annual cycle. The simulation accuracy of hydrologic models for different months often mimic this annual cycle, as shown in Fig. 6, which displays the performance statistics of the individual model simulations for Illinois River basin at Eldon during the train-

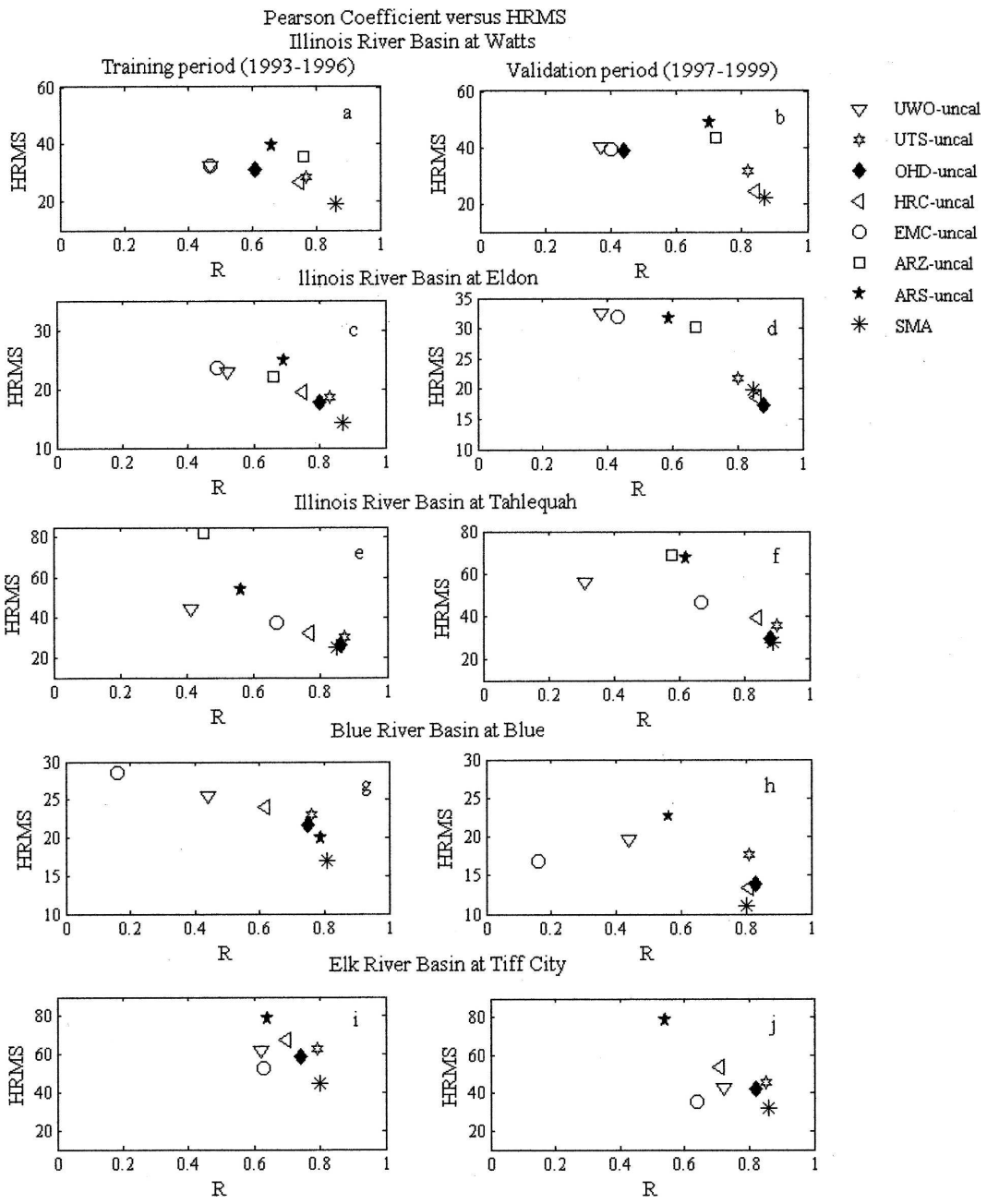

FIG. 3. Hourly root-mean-square error vs Pearson coefficient for SMA and uncalibrated member models for all of the basins.

ing period. Figure 5 reveals that a model might perform well in some months, but poorly in other months, when compared to other models. This led us to hypothesize that the weights for different months should take on different sets of values to obtain consistently skillful simulations for all months. To test this hypothesis, model weights for each calendar month were computed separately for all basins and all multimodel combination techniques.

Figures 7a–j show the scatterplots of the HRMS values when a single set of model weights were computed for overall training period versus the HRMS values when monthly weights were computed. Figures 7a,c,e,g,i were for the training period and Figs. 7b,d,f,h,j were for the validation period. From these figures, it is clear that the performance of MMSE and M3SE with monthly weights is generally better than that with single sets of weights for the entire training period. Applying

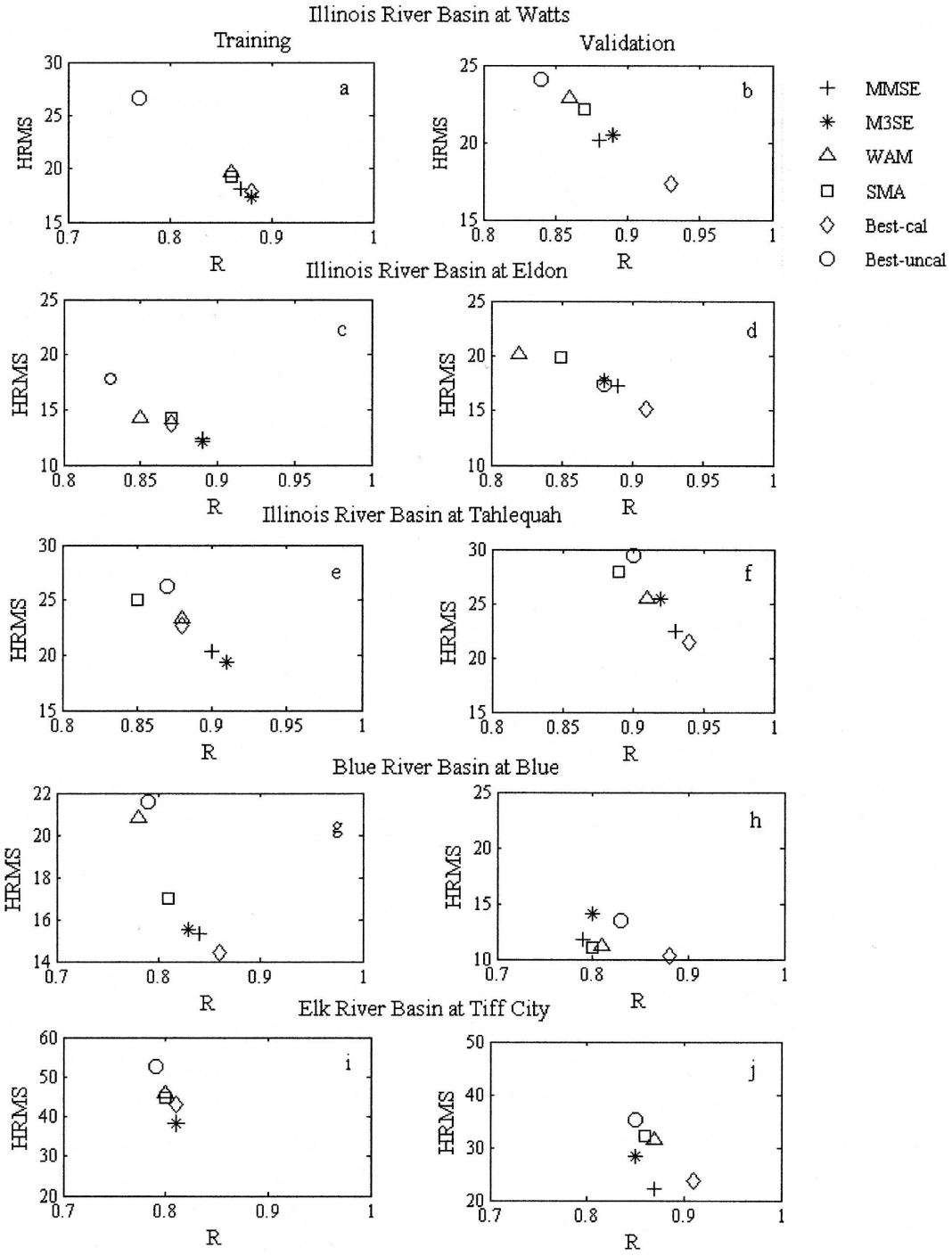

FIG. 4. Hourly root-mean-square error vs Pearson coefficient for all model combinations (MMSE, M3SE, WAM, and SMA) against the best-performing uncalibrated and calibrated model for all the basins (the closer to the bottom-right corner, the better the model).

monthly weights for WAM does not improve the results, and in some cases the results worsen over the training period. During the validation period, however, the performance statistics using single sets of weights are generally better than those using monthly weights. This is because the stationarity assumptions are more easily violated when the multimodel techniques are applied monthly.

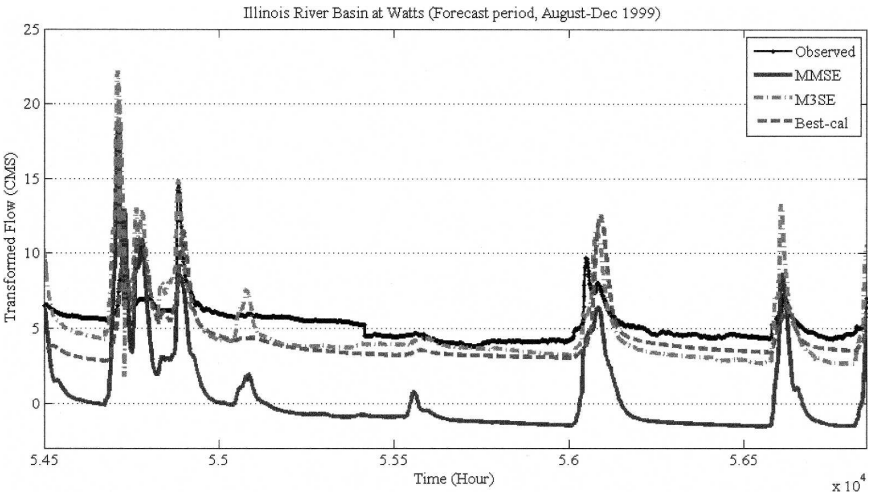

FIG. 5. An excerpt of streamflow simulation results for the Illinois River basin at Watts during the forecast period, illustrating the performance of MMSE and M3SE combination techniques against the observed and best-calibrated model. As can be seen M3SE has feasible streamflow values when MMSE produces negative streamflow values.

d. The effect of different number of models used for model combination on the skill levels of multimodel predictions

The use of multimodel simulations leads to the question of how many models are needed to ensure high accuracy from multimodel simulations. To address this question, we performed a series of experiments by sequentially removing a different number of models from consideration. Figure 8 displays the test results for MMSE. Shown in the figure are the average HRMS and R values when a different number of models was included in model combination. The figure suggests that the inclusion of at least four models is necessary for the MMSE to obtain consistently good skillful results. The figure also shows that including over five models would actually slightly deteriorate the results. This indicates that the accuracy levels of the individual member models may affect the overall accuracy levels of the combination results. To illustrate how important the accuracy of individual models is on the accuracy of the multimodel simulations, we experimented with removing the best-performing models and the worst-performing models from consideration. The effects of removing the best and worst models on the HRMS and *R* values are shown in Figs. 9a–d. The immediate left point from the center in the figures corresponds to the case in which the worst-performing model (w1) was removed and the next point with the two worst models (w1 and w2) removed. The immediate right point from the center in the figure corresponds to the case in which the best-performing model (b1) was removed and the next point with two best models (b1 and b2) removed. The results presented in Fig. 9 highlight two interesting observa-

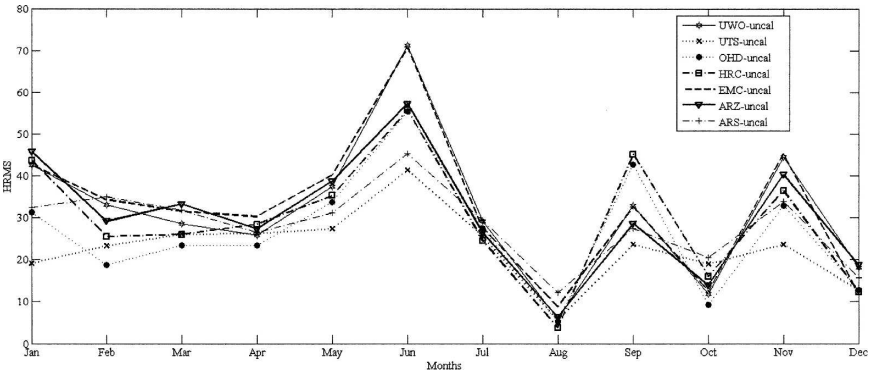

FIG. 6. Monthly HRMS of uncalibrated member models for Illinois River basin at Eldon.

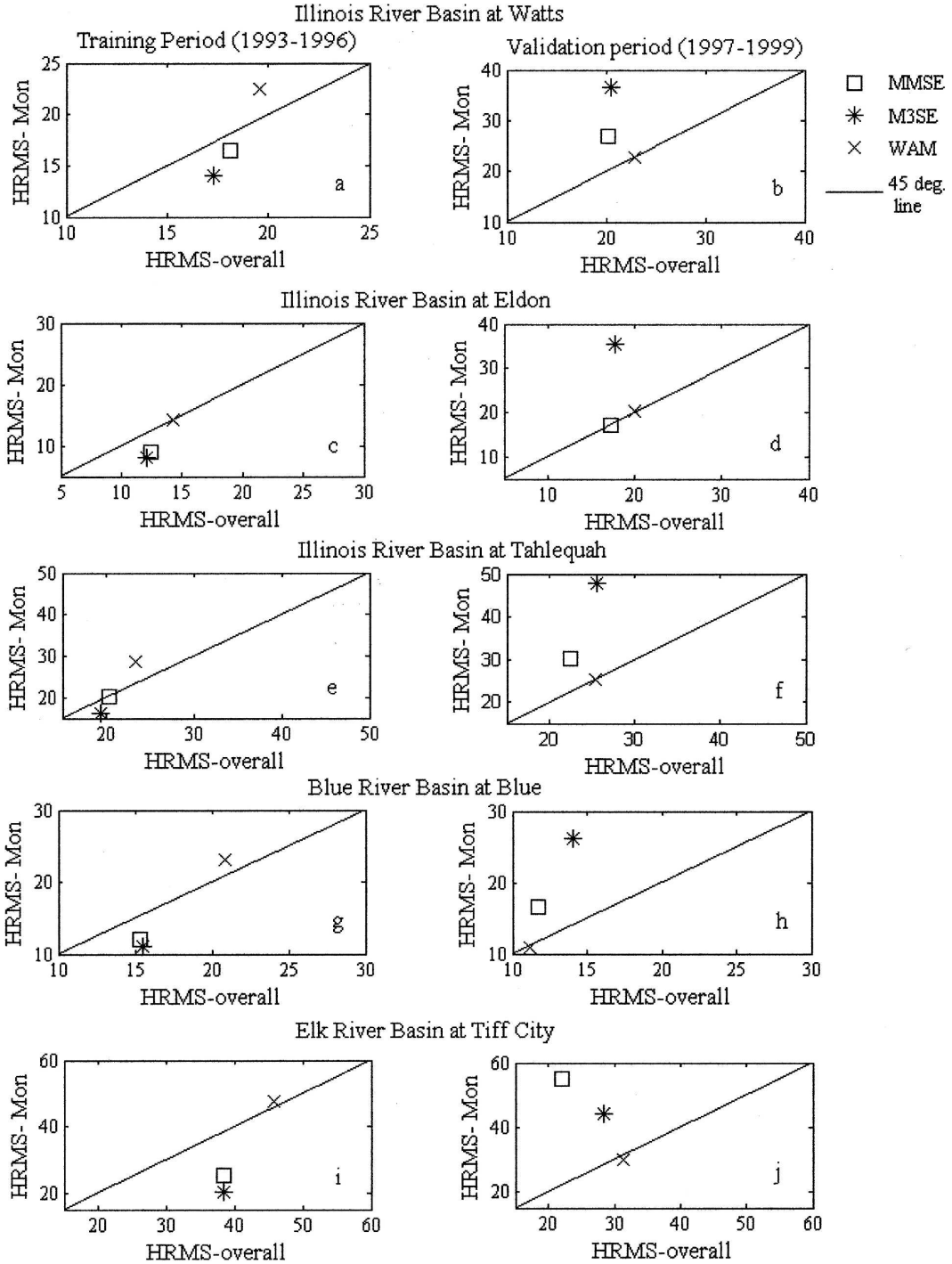

FIG. 7. Hourly root-mean-square error of overall combination methods (HRMS-overall) vs monthly combination methods (HRMS-Mon) for all the basins.

tions. First, notice that excluding the best model(s) would deteriorate the simulation accuracy more significantly compared to eliminating the weakest model(s). Therefore including more skillful models in the multimodel ensemble set led to more accurate simulations, since they are the major source of skill in the multimodel combination. The second interesting observation is that excluding the first worst-performing model (W1) caused deterioration in accuracy of multimodel simulation (HRMS increases and R decreases) while we

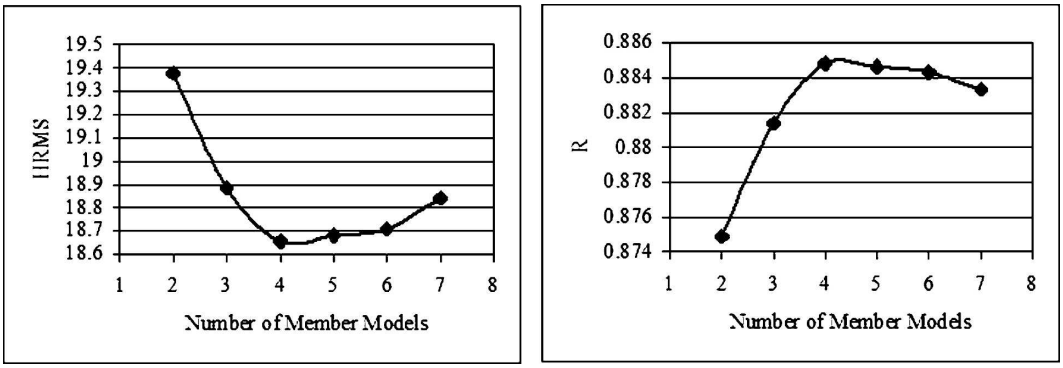

FIG. 8. Average HRMS and *R* statistics for MMSE when a different number of models was included in model combination.

would expect a monotonic improvement of statistic (not deterioration). This reveals that even the worst model(s) can capture some processes within the watershed that has been ignored by other models. This characteristic can make them relatively useful in the multimodel combination strategy.

5. Conclusions and future direction

We have tested four different multimodel combination techniques to the streamflow simulation results from the DMIP, an international project sponsored by the NWS Office of Hydrologic Development, to inter-

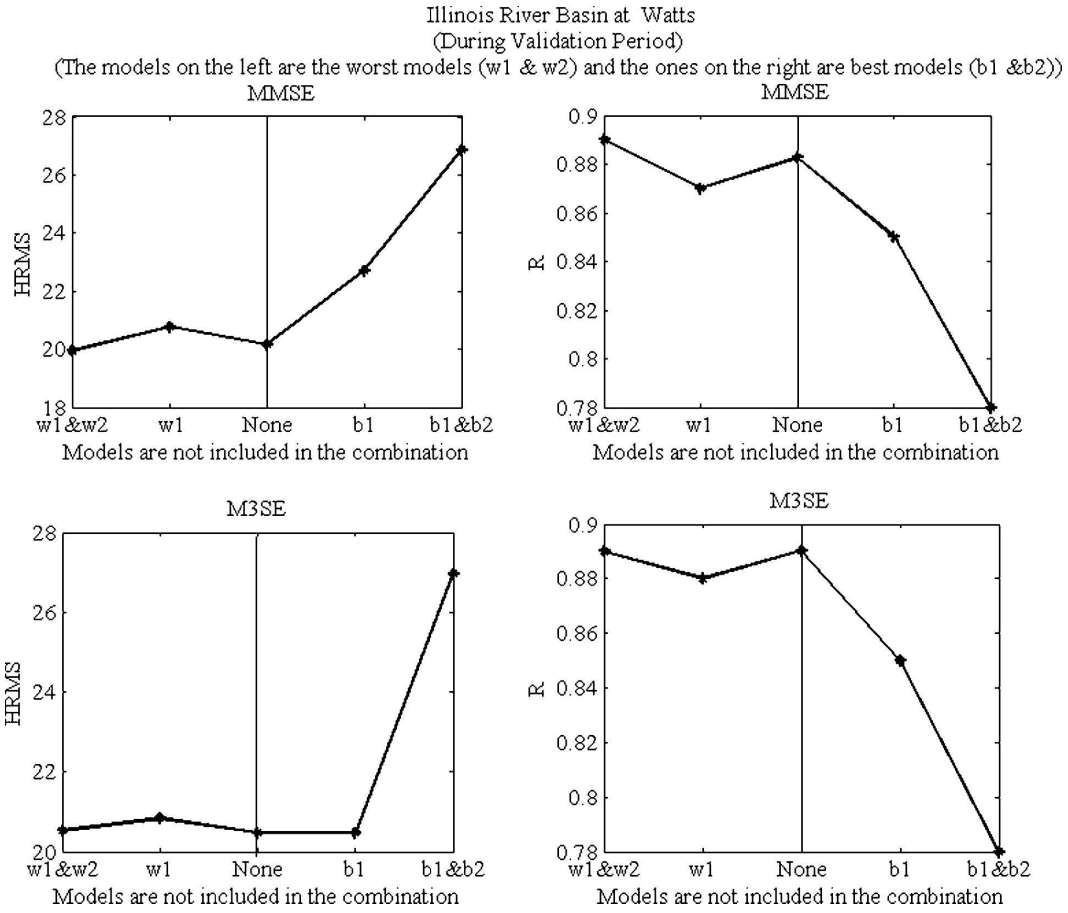

FIG. 9. Goodness-of-fit statistics for various model combinations.

compare seven state-of-the-art distributed hydrologic models in use today (Smith et al. 2004). The DMIP results show that there is a large disparity in the performance of the participating models in representing the streamflow. While developing more sophisticated models may lead to more agreement among models in the future, this work has been motivated by the premise that the accuracy of the existing models is not fully realized. Multimodel combination techniques are viable tools to extract the strengths from different models while avoiding the weaknesses.

Through a series of numerical experiments, we have learned several valuable lessons. First, simply averaging the individual model simulations would result in consensus multimodel simulations that are superior to any single-member model simulations. More sophisticated multimodel combination approaches such as MMSE and M3SE can improve the simulation accuracy even further. The multimodel simulations generated by the MMSE and M3SE can be better than or at least comparable to the best-calibrated single-model simulations. This suggests that future operational hydrologic simulations should incorporate a multimodel combination strategy.

Second, in examining the different multimodel combination techniques, it was found that bias removal is an important step in improving the simulation accuracy of the multimodel simulations. MMSE and M3SE simulations, which incorporated bias correction steps, perform noticeably better than WAM simulations, which did not incorporate bias removal. The M3SE has the advantage of generating consistent streamflow results over the MMSE because its bias removal technique is more compatible with hydrologic variables such as streamflow. Also important is the stationarity assumption when using multimodel combination techniques for simulating streamflow. In the Blue River basin where the average streamflow values are significantly different between the training and validation periods, the advantages of multimodel simulations were lost during the validation period. This finding was also confirmed when the multimodel combination techniques were applied to streamflow from individual months.

Third, we attempted to address how many models are needed to ensure the good accuracy of multimodel simulations. We found that at least four models are required to obtain consistent multimodel simulations. We also found that the multimodel simulation accuracy is related to the accuracy of the individual member models. If the simulation accuracy from an individual model is poor in matching observations, removing that model from consideration does not affect the accuracy of the multimodel simulations very much. On the other hand, removing the best-performing model from consideration does adversely affect the multimodel simulation accuracy. This conclusion supports the need for a better understanding of hydrological processes and to develop well-performing hydrological models that will be included in the multimodel ensemble set. These models are a major source of skill and their contribution in the multimodel combination can advance accuracy and skill of the final results.

This work was based on a limited dataset. There are only seven models and a total of seven years of hourly streamflow data. The findings are necessarily subject to these limitations. Longer dataset and more models (especially skillful models) might improve the multimodel combination results especially during the verification period; however, this needs to be investigated. Further, the regression-based techniques used here (i.e., MMSE, M3SE, and WAM) are vulnerable to a multicolinearity problem, which may result in unstable or unreasonable estimates of the weights (Winkler 1989). This, in turn, would reduce the substantial advantages achieved employing these combination strategies. There are remedies available to deal with a colinearity problem (Shamseldin et al. 1997; Yun et al. 2003). This may entail more independent models to be included in the model combination. It is recommended that a set of hydrologic forecast experiments be conducted using forecasted input forcings (such as forecasted precipitation and temperature) to evaluate performance of multimodel combinations as a real-time forecasting tool.

The multilinear-regression-based approach presented here is only one type of the multimodel combination approach. Over recent years, other model combination approaches have been developed in fields other than hydrology, such as the Bayesian model average (BMA) method, in which model weights are proportional to the individual model accuracy and can be computed recursively as more observation information becomes available (Hoeting et al. 1999). Model combination techniques are still young in hydrology. The results presented in this paper and others (e.g., Butts et al. 2004a,b; Georgakakos et al. 2004) show promise that multimodel simulations will be a superior alternative to current single-model simulation.

Acknowledgments. This work was supported by NSF Sustainability of Semi-Arid Hydrology and Riparian Areas (SAHRA) Science and Technology Center (NSF EAR-9876800) and HyDIS project (NASA Grant NAG5-8503). The work of the second author was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory, under Contract W-7405-

Eng-48. The authors greatly acknowledge detailed and conclusive comments by Dr. Michael Smith and two anonymous reviewers on the original manuscript.

REFERENCES

Abrahart, R. J., and L. See, 2002: Multi-model data fusion for river flow forecasting: An evaluation of six alternative methods based on two contrasting catchments. *Hydrol. Earth Syst. Sci.,* **6,** 655–670.

Ajami, N. K., H. Gupta, T. Wagener, and S. Sorooshian, 2004: Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. *J. Hydrol.,* **298,** 112–135.

Bates, J. M., and C. W. J. Granger, 1969: The combination of forecasts. *Oper. Res. Quart.,* **20,** 451–468.

Beven, K. J., and J. Freer, 2001: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.,* **249,** 11–29.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen, 2004a: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow prediction. *J. Hydrol.,* **298,** 242–266.

——, ——, and J. Overgaard, 2004b: Improving streamflow simulations and flood forecasts with multi-model ensembles. *Proceedings of the 6th International Conference on Hydroinformatics,* S. Y. Liong, K. K. Phoon, and V. Babovic, Eds., World Scientific, 1189–1196.

Clemen, R. T., 1989: Combining forecasts: A review and annotated bibliography. *Int. J. Forecasting,* **5,** 559–583.

Dickinson, J. P., 1973: Some statistical results in the combination of forecast. *Oper. Res. Quart.,* **24,** 253–260.

——, 1975: Some comments on the combination of forecasts. *Oper. Res. Quart.,* **26,** 205–210.

Fraedrich, K., and N. R. Smith, 1989: Combining predictive schemes in long-range forecasting. *J. Climate,* **2,** 291–294.

Georgakakos, K. P., D. J. Seo, H. Gupta, J. Schake, and M. B. Butts, 2004: Characterizing streamflow simulation uncertainty through multimodel ensembles. *J. Hydrol.,* **298,** 222–241.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.,* **14,** 382–417.

Hogue, T. S., S. Sorooshian, V. K. Gupta, A. Holz, and D. Braatz, 2000: A multistep automatic calibration scheme for river forecasting models. *J. Hydrometeor.,* **1,** 524–542.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate,* **15,** 793–799.

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved skill of weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550.

——, ——, D. W. Shin, and C. E. Williford, 2000a: Improving tropical precipitation forecasts from a multianalysis superensemble. *J. Climate,* **13,** 4217–4227.

——, ——, Z. Zhang, T. LaRow, D. Bachiochi, and C. E. Williford, 2000b: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

——, and Coauthors, 2001: Real-time multianalysis-multimodel superensemble forecasts of precipitation using TRMM and SMM/I products. *Mon. Wea. Rev.,* **129,** 2861–2883.

——, and Coauthors, 2002: Superensemble forecasts for weather and climate. *Proc. ECMWF Seminar on Predictability of Weather and Climate,* Reading, United Kingdom, ECMWF.

Maurer, E. P., G. M. O'Donnell, D. P. Lettenmaier, and J. O. Roads, 2001: Evaluation of the land surface water budget in NCEP/NCAR and NCEP/DOE reanalyses using an off-line hydrologic model. *J. Geophys. Res.,* **106** (D16), 17 841–17 862.

Mayers, M., T. N. Krishnamurti, C. Depradine, and L. Moseley, 2001: Numerical weather prediction over the eastern Caribbean using Florida State University (FSU) global and regional spectral models and multi-model/multi-analysis superensemble. *Meteor. Atmos. Phys.,* **78,** 75–88.

Newbold, P., and C. W. J. Granger, 1974: Experience with forecasting univariate time series and the combination of forecasts. *J. Roy. Stat. Soc.,* **137A,** 131–146.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, D. J. Seo, and DMIP Participants, 2004: Overall distributed model intercomparison project results. *J. Hydrol.,* **298,** 27–60.

Russo, R., A. Peano, I. Becchi, and G. A. Bemporad, Eds., 1994: *Advances in Distributed Hydrology*. Water Resources Publications, 416 pp.

Shamseldin, A. Y., and K. M. O'Connor, 1999: A real-time combination method for the outputs of different rainfall–runoff models. *Hydrol. Sci. J.,* **44,** 895–912.

——, ——, and G. C. Liang, 1997: Methods for combining the outputs of different rainfall–runoff models. *J. Hydrol.,* **197,** 203–229.

Singh, V. P., Ed., 1995: *Computer Models of Watershed Hydrology*. Water Resources Publications, 1144 pp.

——, and D. K. Frevert, Eds., 2002a: *Mathematical Models of Large Watershed Hydrology*. Water Resources Publications, 914 pp.

——, and ——, Eds., 2002b: *Mathematical Modeling of Small Watershed Hydrology and Applications*. Water Resources Publications, 972 pp.

Smith, M. B., D.-J. Seo, V. I. Koren, S. Reed, Z. Zhang, Q. Duan, F. Moreda, and S. Cong, 2004: The distributed model intercomparison project (DMIP): Motivation and experiment design. *J. Hydrol.,* **298,** 4–26.

Thompson, P. D., 1976: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.,* **105,** 228–229.

Vieux, B. E., 2001: *Distributed Hydrologic Modeling Using GIS*. Kluwer Academic, 294 pp.

Winkler, R. L., 1989: Combining forecasts: A philosophical basis and some current issues. *Int. J. Forecasting,* **5,** 605–609.

Xiong, L. H., A. Y. Shamseldin, and K. M. O'Connor, 2001: A non-linear combination of the forecasts of rainfall–runoff models by the first-order Takagi–Sugeno fuzzy system. *J. Hydrol.,* **245,** 196–217.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Climate,* **16,** 3834–3840.