

Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations

Naresh Devineni,¹ A. Sankarasubramanian,¹ and Sujit Ghosh²

Received 22 December 2006; revised 19 May 2008; accepted 6 June 2008; published XX Month 2008.

[1] A new approach for developing multimodel streamflow forecasts is presented. The methodology combines streamflow forecasts from individual models by evaluating their skill, represented by rank probability score (RPS), contingent on the predictor state. Using average RPS estimated over the chosen neighbors in the predictor state space, the methodology assigns higher weights for a model that has better predictability under similar predictor conditions. We assess the performance of the proposed algorithm by developing multimodel streamflow forecasts for Falls Lake Reservoir in Neuse River Basin, North Carolina (NC), through combining streamflow forecasts developed from two low-dimensional statistical models that use sea-surface temperature conditions as underlying predictors. To evaluate the proposed scheme thoroughly, we consider a total of seven multimodels that include existing multimodel combination techniques such as combining based on long-term predictability of individual models and by simple pooling of ensembles. Detailed nonparametric hypothesis tests comparing the performance of seven multimodels with two individual models show that the reduced RPS from multimodel forecasts developed based on the proposed algorithm is statistically significant from the RPSs of individual models and from the RPSs of existing multimodel techniques. The study also shows that adding climatological ensembles improves the multimodel performance resulting in reduced average RPS. Contingency analyses on categorical (tercile) forecasts show that the proposed multimodel combination technique reduces average Brier score and total number of false alarms, resulting in improved reliability of forecasts. However, adding multiple models with climatology also increases the number of missed targets (in comparison to individual models' forecasts) which primarily results from the reduction of increased resolution that is exhibited in the individual models' forecasts under various forecast probabilities.

Citation: Devineni, N., A. Sankarasubramanian, and S. Ghosh (2008), Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations, *Water Resour. Res.*, 44, XXXXXX, doi:10.1029/2006WR005855.

1. Introduction

[2] Seasonal to interannual (long-lead) streamflow forecasts based on climate information are essential for short-term water management and for setting up contingency measures during years of extreme climatic conditions. Interest in the development and application of long-lead streamflow forecasts has grown tremendously over the last decade primarily because of the improved monitoring of sea-surface temperature (SST) in the tropical Pacific, resulting in better understanding hydroclimatic teleconnections as well as because of the issuance of operational climate forecasts from General Circulation Models (GCMs) by various centers and research institutions on a monthly basis. Since GCM-predicted fields of precipitation and temperature are usually available at large spatial scales ($2.5^\circ \times$

2.5° - typically $275 \text{ Km} \times 275 \text{ Km}$ in the tropics), one needs to apply either dynamical or statistical downscaling to develop streamflow forecasts. Dynamical downscaling nests a regional climate model (RCM) with GCM outputs as boundary conditions to obtain precipitation and temperature at watershed scale ($60 \text{ Km} \times 60 \text{ Km}$). The downscaled precipitation and temperature at watershed scale could be used further as inputs into a watershed model to obtain seasonal streamflow forecasts [Leung et al., 1999; Roads et al., 2003; Seo et al., 2003, Carpenter and Georgakakos, 2001]. An alternative would be to use statistical downscaling, which maps the GCM precipitation and temperature forecasts to observed streamflow forecasts at a given point through a statistical relationship [Robertson et al., 2004a, 2004b; Landman and Goddard, 2002; Gangopadhyay et al., 2005]. One could also develop a low-dimensional statistical model without using GCM outputs by relating the observed streamflow to identified climatic precursors (e.g., El Nino Southern Oscillation (ENSO) indices) that influence the streamflow potential at the given site [Grantz et al., 2005; Hamlet and Lettenmaier, 1999; Souza and Lall, 2003; Sankarasubramanian and Lall, 2003]. Seasonal streamflow forecasts obtained using these approaches are better repre-

¹Department of Civil, Construction and Environmental Engineering, North Carolina State University, Raleigh, North Carolina, USA.

²Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA.

sented probabilistically in the form of ensembles to represent the uncertainty, particularly in quantifying the effects of both changing boundary conditions (SST) and initial conditions (atmospheric and land surface conditions). Apart from these uncertainties resulting from initial and boundary conditions, the model that is employed for developing streamflow forecasts could also introduce uncertainty in prediction. In other words, even if streamflow forecasts obtained by dynamical downscaling are forced with observed boundary and initial conditions (perfect forcings), it is inevitable that the simulated streamflows will have uncertainty in prediction, which is otherwise known as model error/uncertainty. A common approach to reduce model uncertainty is through the refinement of parameterizations and process representations of the model under consideration (e.g., GCMs or RCMs or hydrologic models). Given that developing and running GCMs is time consuming, recent efforts have focused on reducing the model error by combining multiple GCMs to issue operational climate forecasts [Rajagopalan et al., 2002; Robertson et al., 2004a, 2004b; Barnston et al., 2003; Doblas-Reyes et al., 2000; Krishnamurti et al., 1999]. Similarly, studies have shown that developing multimodel forecasts by combining different low-dimensional streamflow forecasting models results in considerable improvement over the performance of individual models [Regonda et al., 2006]. Thus combining streamflow forecasts from multiple models seems to be a good alternative in improving the overall predictability of seasonal streamflow forecasts and reducing the overall error in prediction.

[3] The main goal of this study is to develop and apply a new scheme for combining forecasts from multiple models, which could be either streamflow forecasts from low-dimensional models or GCM forecasts available at large spatial scales, by assessing the model's predictive skill conditioned on the predictor state. The basic reason leading to better performance of multimodel ensembles is due to the incorporation of realizations from various models, thereby increasing the number of ensembles to represent the conditional distribution of hydroclimatic attributes. Recent studies on improving seasonal climate forecasts using optimal multimodel combination techniques assign weights for a particular model based on its ability to predict the climatic variable over the entire period for which the GCM simulations are available [Rajagopalan et al., 2002; Robertson et al., 2004a, 2004b; Barnston et al., 2003]. Given that each model's predictive skill could also vary depending on the state of the predictor (SSTs for GCMs), we develop a new methodology for multimodel combination that assigns weights to each model by assessing the skill of the models contingent on the predictor state. The proposed methodology is employed upon two low-dimensional seasonal probabilistic streamflow forecasting models that primarily use tropical Pacific and Atlantic SST conditions to develop multimodel ensembles of streamflow forecasts. Though the proposed approach is demonstrated by combining two statistical streamflow forecasting models, we believe that the approach presented in section 3.1 could be extended to combine multiple GCMs. Since the streamflow forecasts are represented probabilistically, we use rank probability score (RPS) [Candille and Talagrand, 2005] as a measure to assess the model's predictive skill.

[4] Section 2 provides a brief background on multimodel combination techniques that is currently pursued in the literature for developing operational climate and seasonal streamflow forecasts. Section 3 presents the proposed multimodel combination scheme that assesses the skill of the model contingent on the predictor state. In section 4, we briefly discuss the two low-dimensional streamflow forecasting models that were employed for developing probabilistic streamflow forecasts for predicting the summer flows (July-August-September, JAS) into Falls Lake, Neuse River Basin, NC. Finally, in section 5, we illustrate the application of the proposed multimodel combination presented in section 3 to develop improved probabilistic streamflow forecasts for predicting JAS inflows into the Falls Lake, NC.

2. Background and Motivation

[5] Efforts to address model uncertainty through combining outputs from multiple models have been investigated in climate and weather forecasting [Doblas-Reyes et al., 2000; Rajagopalan et al., 2002; Krishnamurti et al., 1999] and in streamflow simulation through calibration [Boyle et al., 2000; Georgakakos et al., 2004; Marshall et al., 2006; Ajami et al., 2006, 2007]. Perhaps the simplest approach to develop multimodel forecasts is to pool the predicted values or the ensembles from all the models, thus giving equal weights for all the models [Palmer et al., 2000]. Recent research from PROVOST (PRediction Of climate Variations On Seasonal to interannual Time-scales) and from International Research Institute for Climate and Society (IRI) show that multimodel ensembles of climate forecasts provides improved reliability and resolution than the individual model forecasts [Palmer et al., 2000; Doblas-Reyes et al., 2000; Barnston et al., 2003]. Though the improved predictability of multimodel ensembles partly arise from increase in the sample size, studies have compared the performance of single models having the same number of ensembles as the pooled multimodel ensembles and have shown that the multimodel approach naturally offers better predictability because of the ability to incorporate outcomes from multiple models, thereby encompassing underlying different process parameterizations and schemes [Hagedorn et al., 2005]. Since the advantage gained through multimodel combination is a better representation of conditional distribution of hydroclimatic attributes, it is important to evaluate probabilistic forecasts developed from multimodel ensembles through various performance evaluation measures and to analyze the predictability for various geographic regions [Hagedorn et al., 2005]. Recent studies have also considered climatology as one of the candidate forecasting scheme in developing multimodel ensembles [Rajagopalan et al., 2002; Robertson et al., 2004a, 2004b].

[6] Another approach that is currently gaining attention is to develop a strategy for combining multimodel ensembles using either optimization methods [Rajagopalan et al., 2002; Robertson et al., 2004a, 2004b] or statistical techniques [Krishnamurti et al., 1999]. Under optimal combination approach, weights are obtained for each model as a fraction, such that the chosen skill/performance measure of the multi model ensembles obtained by using these fractions is maximized [Rajagopalan et al., 2002; Robertson et al., 2004a, 2004b; Regonda et al., 2006]. The easiest approach

to assign weights for multimodel ensembles is to give a higher weight for a model that has lower forecast error. Methods based on statistical techniques such as linear regression have also been employed so that the multimodel forecasts have better skill than single models [Krishnamurti *et al.*, 1999]. However, the application of optimal combination approach using either statistical or optimization techniques require observed climatic or streamflow attributes at a particular grid point or station. Studies have also used advanced statistical techniques such as canonical variate method [Mason and Mimmack, 2002] and Bayesian hierarchical method [Stephenson *et al.*, 2005] for developing multimodel combinations. Hoeting *et al.* [1999] show that the mean of the posterior distribution of the predictand obtained by averaging over all the models with its probability of occurrence provides better predictive ability (measured by logarithmic scoring rule) than the mean of the posterior distribution of the predictand obtained from a single model. For a detailed review of Bayesian averaging on various statistical models (e.g., generalized linear models and nonlinear regression models), [see Hoeting *et al.*, 1999].

[7] The multimodel combination method proposed here is motivated by the fact that the skill of the GCM forecasts or downscaled streamflow forecasts depends on the predictor conditions that determine the conditional distribution of the hydroclimatic attributes. Studies focusing on the skill of GCMs show that the overall predictability of GCMs is enhanced during ENSO years over North America [Brankovic and Palmer, 2000; Shukla *et al.*, 2000; Quan *et al.*, 2006]. Similarly, studies have also shown the importance of various oscillations or climatic conditions in influencing the predictability of GCMs over various parts of the globe. For instance, Giannini *et al.* [2004] show that tropical Atlantic variability plays a preconditioning state in the development of ENSO related teleconnection in determining GCM's ability to predict rainfall over North East Brazil (Nordeste), which is a region shown to have significant skill in seasonal climate prediction [Moura and Shukla, 1981; Ropelewski and Halpert, 1987 and references therein]. Giannini *et al.* [2004] show that the predictability of Nordeste rainfall using Community Climate Model Version 3 (CCM3) GCM [Kiehl *et al.*, 1998] is poor particularly if the North Atlantic SSTs exhibit opposite anomalous conditions to the tropical Pacific SSTs. More precisely, the predictability of Nordeste rainfall by CCM3 is negative with positive SST anomalies in tropical Pacific (warm) and negative SST anomalies (cold) in North Atlantic as well as with cold tropical Pacific (negative SST anomalies) and warm North Atlantic (positive SST anomalies) conditions. Naturally, under these predictor conditions, one would prefer to use climatology instead of climate forecasts, since they are negatively correlated with the observed rainfall. On the basis of this, we propose that for post-processing of individual model's forecasts to develop multimodel ensembles, one needs to assess the skill of the individual model ensembles based on the predictor state. By considering climatology as one of the candidate forecasts, we develop a multimodel combination scheme that formally assesses and compares the skill of the competing models under given predictor conditions so that lower weights are assigned for a model that has poor predictability under such conditions. The next section formally develops a multimodel combination scheme using

rank probability score (RPS) as the basic measure for evaluating the forecasting skill.

3. Multimodel Combination Based on Predictor State—Methodology

[8] Error resulting from climate forecasts is primarily of two types: (1) Uncertainty in initial and boundary conditions and (2) Model error [Hagedorn *et al.*, 2005]. The first source of error is typically resolved by representing the uncertainties in initial and boundary conditions in the form of ensembles. The second source of error arises from process representation, which could be reduced by combining forecasts from multiple models which incorporate various process representations and model physics to develop an array of possible scenarios of outcomes. Developing multimodel ensembles result in reducing both sources of error. However, even after developing multimodel ensembles, observations could lie outside the realm of these models [Hagedorn *et al.*, 2005]. Similarly, the performance of individual models and multimodel ensembles may be poor during certain boundary/SST conditions owing to limited relationship between SST conditions and precipitation/temperature over a particular location/grid [Goddard *et al.*, 2003]. Under these climatic conditions with all models having poor predictability, it may be useful to consider climatology as a forecast.

[9] Figure 1 demonstrates the motivation behind the proposed methodology by employing a mixture of regression models that depends on two predictors with the dominant predictor (X_1) influencing the predictand only if it crosses a certain threshold ($|X_1| > 1.0$). On the basis of the mixture model, a total of $n = 1000$ realizations is generated. The estimated correlation between y_1 and x_1 is 0.671 and y_1 and x_2 is 0.134, which would suggest one to give higher importance to predictor x_1 . Figure 1b shows the skill of the regression model, which is expressed as correlation between fitted y_1 and y_1 conditioned on x_1 , over the entire range of x_1 . To estimate this conditional correlation, we consider a bandwidth of 1 on the given x_1 such that the fitted values of y_1 obtained using the predictor x_1 within that bandwidth are only considered. We can clearly infer from Figure 1b, because of the limited influence of x_1 on y_1 between $x_{1L} = -1$ to 1, the skill of the model during those predictor conditions is very low. Developing a model based on the predictor x_1 alone would result in poor prediction particularly when $|X_1|$ is below the threshold value.

[10] Our approach of multimodel combination addresses this issue by assessing the model performance based on the boundary conditions-the predictor state. For instance, if the predictability of all models is really bad during a particular condition, then one would replace model forecasts with climatology by assigning higher weights for climatological ensembles. The following section formally describes the multimodel combination methodology that could be employed for a given set of forecasts from multiple models and the predictors that influence those forecasts.

3.1. Multimodel Combination Based on Predictor State Space—Algorithm

[11] Let us suppose that we have streamflow forecasts, $Q_{i,t}^m$, where $m = 1, 2, \dots, M$ denotes the forecasts from M different models, $i = 1, 2, \dots, N_m$ represents ensembles of

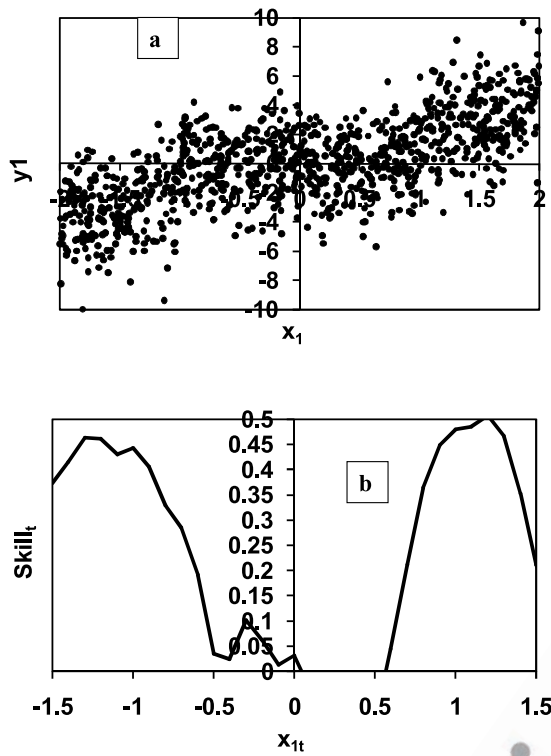


Figure 1. Importance of assessing the skill of the models from the predictor space. Realizations shown in Figure 1a are generated with the predictand y_1 depending on two predictors, x_1 and x_2 , with x_1 influencing the predictand only if the absolute value of the predictor x_1 is greater than the threshold value of 1. The underlying model is $y_{1t} = 2x_{1t} + 0.5x_{2t} + \varepsilon_t$ if $|x_{1t}| > 1$ and $y_{1t} = 0.25x_{2t} + \varepsilon_t$ if $|x_{1t}| \leq 1$. The noise term ε_t follows i.i.d, with normal distribution having zero mean and a standard deviation of 2. The predictors follow uniform distribution between -2 and 2 . A total of $n = 1000$ realizations is generated from this mixture model which could be analogously compared to two predictors as anomalous SST conditions influencing the local hydroclimatology.

mentioned before, developing multimodel ensembles based on optimal combination method requires the observed climatic/streamflow variables O_t , using which one could assess the skill of the probabilistic forecasts using rank probability score (RPS) [Murphy, 1970; Candille and Talagrand, 2005; Anderson, 1996] to obtain the weights w_t^m . It is important to note that RPS is evaluated for each year using the ensembles ($n = 10000$) representing the conditional distribution, which is quite different from correlation for which one needs a minimum of two years of forecasts. The main advantage of using RPS is that it quantifies the error in estimating the entire conditional distribution, whereas measures such as correlation and Root Mean Square Error (RMSE) consider only the error in predicting conditional mean. One could also employ continuous rank probability score, which compares the observed event with the forecasted probabilities using a parametric functional form [Candille and Talagrand, 2005; Wilks, 1995]. The rank probability skill score (RPSS) represents the level of improvement of the forecast RPS in comparison to the reference forecast RPS, which is usually assumed to be climatology. Appendix A provides details on obtaining RPS and RPSS for given probabilistic forecasts.

[12] Let us denote the RPS and RPSS of the probabilistic forecasts, $Q_{i,t}^m$, for each time step as RPS_t^m and $RPSS_t^m$, respectively. For the purpose of multimodel combination, we assess the skill of the model by analyzing its predictability under similar climatic conditions. One could identify similar predictor conditions in the predictor state space by choosing a distance metric that computes the distance between the current predictor state, X_t , and the historical predictor vector, X . For instance, if the current state of the predictors indicates El Nino conditions, then past El Nino conditions could be considered as similar conditions. The distance metric could be either Euclidean distance or a more generalized distance measure such as the Mahalanobis distance, which is more useful if the predictors exhibit correlation among them. Compute the distances d_{it} between the current conditioning state X_t , and the historical predictor vector X_i using Mahalanobis distance:

$$d_{it} = \sqrt{(X_t - X_i)^T \hat{\Sigma}^{-1} (X_t - X_i)} \quad (1)$$

where $\hat{\Sigma}$ denotes the variance-covariance matrix of the historical predictor vector X . One can note that if $l = t$, the distance metric, d_{it} , reduces to zero. Similarly, if X_t is the principal components of the original SST fields, then the Mahalanobis distance boils down to the Euclidean distance. Using the distance vector d , the ordered set of nearest neighbor indices J can be identified. Thus the j th element in the distance vector metric provides the j th closest state X_i to the current state X_t . Using this information, we assess the performance of each model in the predictor state space as

$$\lambda_{i,k}^m = \frac{1}{K} \sum_{j=1}^K RPS_{(j)}^m \quad (2)$$

where $RPS_{(j)}^m$ denotes the skill of the forecasting model, m , for the year that represents the j th closest condition (obtained from J) to the current condition X_t . In other

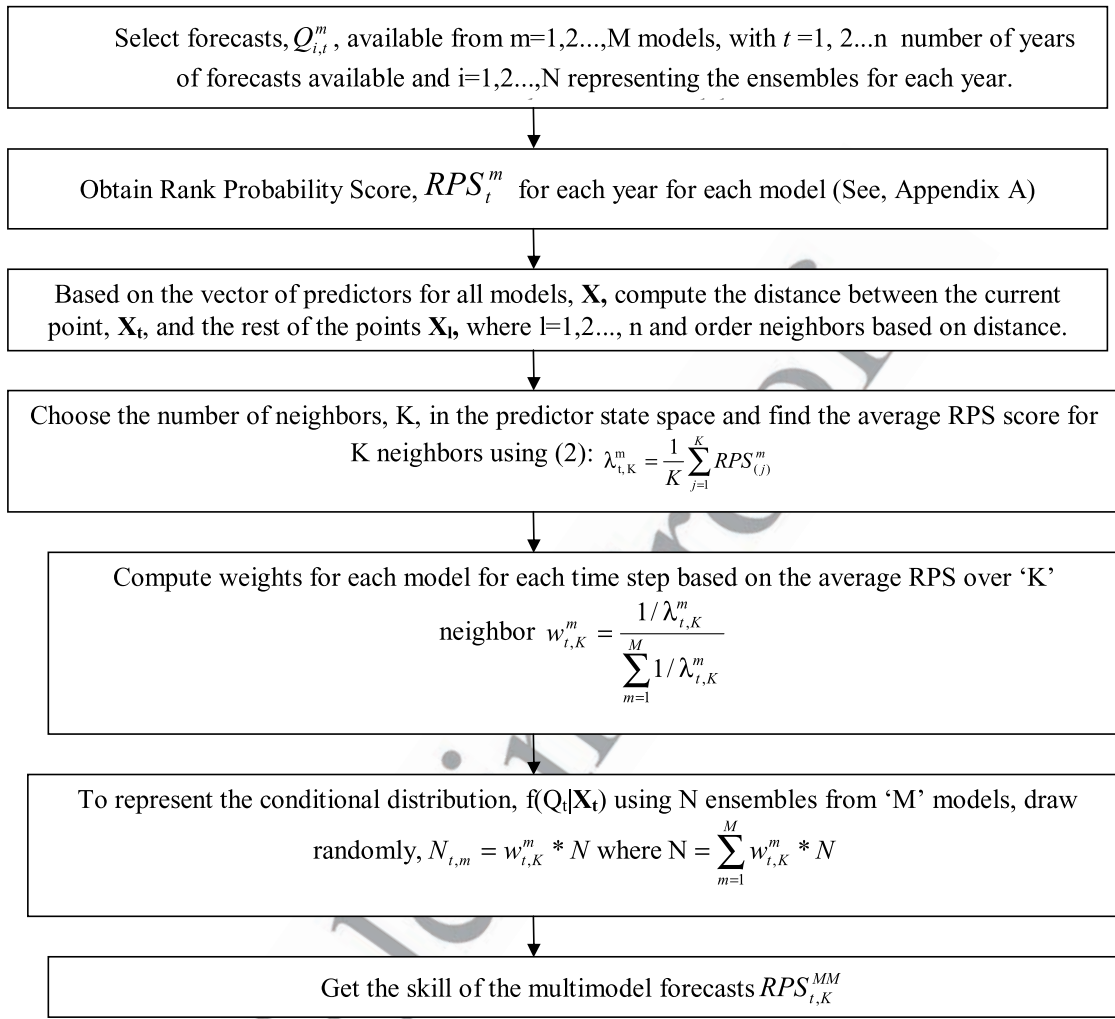


Figure 2. Flowchart of the multimodel combination algorithm described in section 3.1 for fixed number of neighbors “ K ” in evaluating the model skill from the predictor state space. To apply the same algorithm for K_t that gives the minimum RPS from the multimodel ensembles, compute $RPS_{t,k}^{MM}$ for $k = 1, 2, \dots, n - 1$ and choose K_t that corresponds to minimum $RPS_{t,k}^{MM}$.

words, $\lambda_{t,K}^m$ summarizes the average skill of the forecasting model, m , by choosing K years that resemble conditions very similar to the current condition, $\mathbf{X}_t$. Using $\lambda_{t,K}^m$ obtained for each model at each time step, we obtain the weights for the multimodel combination so that a model with better performance during particular climatic conditions needs to be represented with more number of ensembles in comparison to a model with lower predictability under those conditions. It is important to note that RPS is a measure of error in predicting the probabilities and it is evaluated based on the entire ensembles that represent the conditional distribution of streamflows. We can use the average skill of the model, $\lambda_{t,K}^m$, to obtain the weights for each model.

$$w_{t,K}^m = \frac{1/\lambda_{t,K}^m}{\sum_{m=1}^M 1/\lambda_{t,K}^m} \quad (3)$$

Equation (3) basically gives higher weights for a model that has smaller average RPS in the predictor state space. If $\lambda_{t,K}^m$

is zero for a subset of models $M_1 \leq M$, then the weights $w_{t,K}^m$ are distributed equally ($1.0/M_1$) between the models for which $\lambda_{t,K}^m$ is zero and the remaining models ($M - M_1$) are assigned zero weight. For instance, if two models out of five, have an average RPS ($\lambda_{t,K}^m$) of zero taken over K neighbors, then these two models will be assigned a weight of 0.5 and the remaining three models will be assigned a weight of zero. The multimodel forecasts for each time step could be developed by drawing $w_{t,K}^m * N$ ensembles from each model to constitute the multimodel ensembles. Thus one has to specify the number of neighbors K to implement this approach. It is also important to note that choosing fewer K does not imply that the multimodel forecasts are developed using the observed predictand and predictors based in the identified K similar conditions. In fact, $Q_{i,t}^m$ are forecasts developed based on the observed values of the predictand and predictor over a particular training period (for leave-one-out cross-validated forecasts, we use $n - 1$ years of record as training period; for adaptive forecasts, we develop the model using n_1 years and the remaining $(n - n_1)$ are considered for validation). Thus we use the weights,

431 $w_{i,K}^m$, only to draw the ensembles from $Q_{i,K}^m$, which is in fact
 432 obtained based on the chosen training period for developing
 433 the forecasts. The simplest approach for selecting the
 434 number of neighbors is to find a fixed K that corresponds to
 435 the lowest-average RPS from the multimodel forecast. The
 436 other approach would be to choose a different number of
 437 neighbors K_i for each year. Under this approach, “Varying
 438 K_i ”, the chosen K_i corresponds to the minimum RPS that
 439 could be obtained from the multimodel ensembles for that
 440 year. Obviously, *Varying* K_i will ensure the lowest-average
 441 RPS for the entire forecast, but it is computationally
 442 intensive. The performance of multimodel ensembles is
 443 compared with individual model’s predictive skill using
 444 various verification measures such as average RPS ($\overline{RPS}$),
 445 average RPSS ($\overline{RPSS}$), and anomaly correlation ($\hat{\rho}$).

447 4. Seasonal Streamflow Forecasts Development 448 for the Neuse Basin

449 [13] For the purpose of demonstrating the multimodel
 450 approach proposed in section 3, we first develop probabi-
 451 listic seasonal streamflow forecasts from two different
 452 models based on the climate information for the Falls Lake,
 453 Neuse River Basin in North Carolina (NC). We develop
 454 streamflow forecasts based on two low-dimensional statis-
 455 tical models, one based on parametric regression approach
 456 and another using a nonparametric approach based on
 457 resampling [Souza and Lall, 2003]. We first provide brief
 458 baseline information about the Neuse Basin and its
 459 importance to the water management of the Research
 460 Triangle Park area of NC.

462 4.1. Baseline Information for the Neuse Basin

463 [14] Falls Lake (location shown in Figure 3a) is a multi-
 464 purpose reservoir authorized for flood control, water supply,
 465 water quality, recreation and for fish/wildlife protection.
 466 Given that the water demand in the Triangle area has been
 467 growing rapidly in the last decade, multi-year droughts
 468 (1998–2005) and ensued restrictions has increased the
 469 importance of long-lead forecasts toward better management
 470 of water supply systems. Observed streamflow information
 471 at Falls Lake is available from 1928 to 2002 from the United
 472 States Army Corps of Engineers (USACE) (<http://epec.saw.usace.army.mil/fall05.htm>). Figure 3a provides the season-
 474 ality of inflow into Falls Lake. Typically, 46% of the annual
 475 inflow occurs during January–February–March (JFM), and
 476 the low flows during July–August–September (JAS) contrib-
 477 ute 14% of the annual inflows. From a water management
 478 perspective, developing streamflow forecasting models for
 479 the low flow season is important since maintaining the
 480 operational rule curve of 251.5’ is challenging during those
 481 months resulting in mandatory restrictions [Weaver, 2005],
 482 which arises because of increased demand for water quality
 483 and water supply releases in the summer (in comparison to
 484 the winter demand). (http://epec.saw.usace.army.mil/Falls_WC_Plan.pdf). In addition to this, the demand in the Wake
 486 County, NC has also grown by about 20%–62% from
 487 1995–2000 [Weaver, 2005] resulting in three severe
 488 droughts (summers of 2002, 2005 and 2007) in the past
 489 five years. Thus managing the reservoirs during the summer
 490 is very important from the perspective of invoking
 491 restrictions to meet the end of season target storage
 492 conditions without compromising the flood risk arising

from hurricanes (K. Golembesky et al., Improved drought
 management of falls lake reservoir: Role of multimodel
 streamflow forecasts in setting up restrictions, manuscript in
 preparation, 2008). The following section provides a brief
 overview of climate and streamflow teleconnection in the
 US.

4.2. Climate and Streamflow Teleconnection in the US—Brief Overview

[15] Climatic variability, at interannual and interdecadal
 time scales, resulting from ocean-atmosphere interactions
 modulate the moisture delivery pathways and has significant
 projections on continental-scale rainfall patterns [Trenberth
 and Guillemot, 1996; Cayan et al., 1999] and streamflow
 patterns at both global and hemispheric scales [Dettinger et
 al., 2000b] as well as at regional scales [e.g., Guetter and
 Georgakakos, 1996; Piechota and Dracup, 1996]. Efforts in
 understanding the linkages between exogenous climatic
 conditions such as tropical sea-surface temperature (SST)
 anomalies and regional hydroclimatology over the U.S.
 have offered the scope of predicting the rainfall/streamflow
 potential on season-ahead and long-lead (12 to 18 months)
 basis [Hamlet and Lettenmaier, 1999; Georgakakos, 2003;
 Wood et al., 2002, 2005]. Interannual modes such as the El
 Nino-Southern Oscillation (ENSO) resulting from anom-
 alous SST conditions in the tropical Pacific Ocean
 influences the interannual variability of precipitation and
 temperature over many regions of the globe [Rasmusson
 and Carpenter, 1982; Ropelewski and Halpert, 1987]. Most
 of the studies focusing on climatic variability over the South
 Eastern US have shown that warm tropical Pacific
 conditions during October–December lead to above-normal
 precipitation during winter and below-normal precipitation
 during summer if warm pool conditions prevail in the
 tropical Pacific during the spring [Schmidt et al., 2001;
 Lecce, 2000; Hansen et al., 1998; Zorn and Waylen, 1997].
 Studies have also reported ENSO related teleconnection
 between precipitation and temperature over NC during both
 winter and summer seasons [Roswintarti et al., 1998;
 Rhome et al., 2000]. We basically develop a low-
 dimensional model by identifying SST conditions that
 influence the seasonal streamflow forecasts into Falls Lake
 during July–September (JAS).

4.3. Seasonal Streamflow Forecasts Development—Individual Models

[16] Our objective is to estimate the conditional dis-
 tribution of streamflows, $f(Q_i|X_i)$, which would occur in
 the upcoming season based on the climatic conditions X_i
 using the chosen statistical model. The next two sections
 (sections 4.3.1 and 4.3.2) discuss in detail about predictor
 selection and the performance of forecasts developed from
 the two statistical models.

4.3.1. Diagnostic Analyses, Predictor Identification, and Dimension Reduction

[17] To identify predictors that influence the streamflow
 into Falls Lake during JAS, we consider SST conditions
 during April–June (AMJ) which could be obtained from
 International Research Institute for Climate and Society
 (IRI) data library (<http://iridl.ldeo.columbia.edu/expert/SOURCES/KAPLAN/EXTENDED/.sst>). Figure 3b
 shows the Spearman rank correlation between the observed

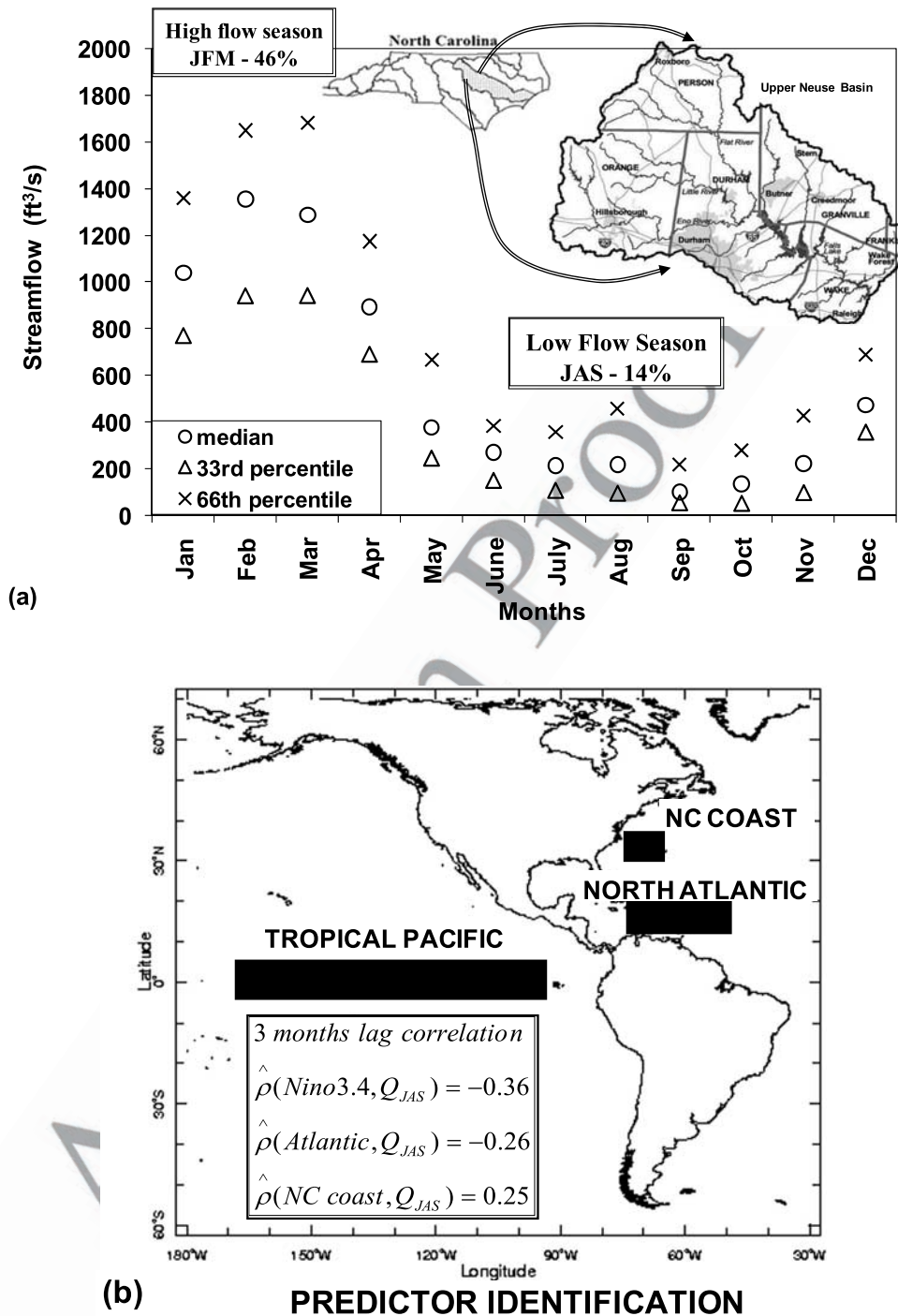


Figure 3. Hydroclimatology of Neuse Basin and predictors identification: (a) seasonality of Neuse Basin and (b) climatic predictors that influence the streamflow. Figure 3a shows the seasonality of Neuse Basin with the flow during JAS, accounting for 14% of the annual streamflow. Insert in Figure 3a shows the location of the Falls Lake in NC. Figure 3b shows the SST regions that influence the streamflow into the Falls Lake. SST regions that have significant correlation at 95% confidence interval (>0.21 or <-0.21) are only considered for model development.

streamflow during JAS at the Falls Lake and the SST conditions during AMJ. From Figure 3b, we see clearly that SST girds over ENSO region (170E–90W and 5S–5N), the North Atlantic region (80W–40W and 10N–20N), and the NC Coast region (75W–65W and 22.5N–32.5N) influence the summer flows into Falls Lake. This is in line

with earlier findings [Roswintarti *et al.*, 1998; Rhome *et al.*, 2000] suggesting that warm conditions in the tropical Pacific and tropical North Atlantic result in above-normal inflow conditions in Falls Lake. It is important to note that we consider SST regions whose correlations are significant and greater than the threshold value of $\pm 1.96/\sqrt{n-3}$ where

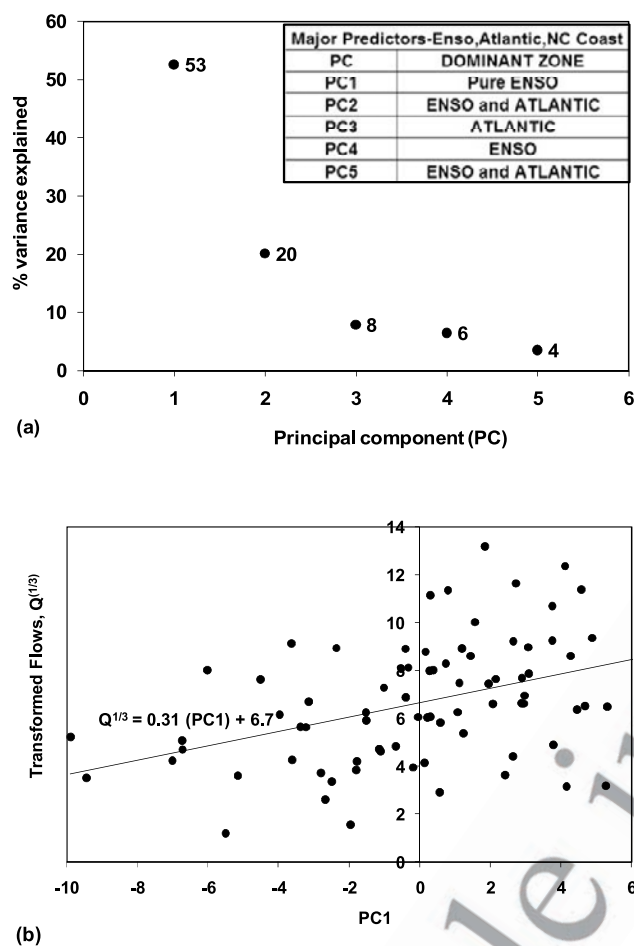


Figure 4. Selection of principal components and their relationship with the flows. Figure 4a shows the scree plot of the principal components of the SSTs in the three regions (in Figure 3b) indicating the percentage variance explained by each component. The inset in Figure 4a indicates the dominant zone of each PC based on eigenvectors analysis. Figure 4b shows the relationship between the first PC and the cube root transformed flows.

n is the total number of years ($n = 78$ years for Falls Lake) of observed records used for computing the correlation. We didn't consider atmospheric conditions such as mean sea level pressure, geo-potential height for developing the streamflow forecasting model, since the National center for Environmental Prediction (NCEP) reanalysis data are available only from 1950. Further, adding atmospheric predictors for predicting post-1950 flows also did not result in any significant improvement in the skill (results not shown).

[18] Given that the SST fields are correlated to each other, we apply Principal Components Analysis (PCA) to identify the dominant modes in the SST grids. PCA, also known as empirical orthogonal function (EOF) analysis, transform the correlated predictors (e.g., SST grids) to uncorrelated principal components by applying singular value decomposition (SVD) on the covariance/correlation matrix of the predictors. Importance of each principal component is quantified, which is usually summarized by the scree plot, based on the fraction of the variance the principal component represents

with reference to the original predictor variance. Mathematics of PCA and the issues in selecting the number of principal components using scree plot could be found in the work of Wilks [1995].

[19] We applied PCA by performing SVD on the covariance matrix of the grid points of SSTs over the three different regions. We did not spatially average the SSTs over the three regions before performing PCA. Instead, we pooled SST grids that were significantly correlated from three regions and then performed PCA. Figure 4a shows the percentage of variance explained by each principal component and the first two components accounts for 73% of the total variance shown in the predictor field in Figure 4a. On the basis of the eigenvectors obtained from PCA (figure not shown), the first component representing the ENSO region has correlation of 0.36 with observed streamflow and the second component representing both the Pacific and the Atlantic has a correlation of -0.23 (significance level ± 0.21 for 78 years of record) with the inflows at Falls Lake. We employ these two principal components to develop seasonal streamflow forecasts for JAS for the Falls Lake.

4.3.2. Performance of Individual Model Forecasts—Resampling and Regression Approach

[20] We consider two non-linear models, parametric regression (with the predictand being cube root of the flows) and semiparametric resampling models [Souza and Lall, 2003], in developing multimodel forecasts. The linear relationship between the cube root of the flows and the dominant principal component (Figure 4b) suggests regression as a favorable candidate for developing streamflow forecasts for the Falls Lake. However, the conditional distribution of the parametric regression model (cube root transformed flows) follows normal distribution which is purely dependent on the conditional mean and its point-forecast error. Hence we consider the semiparametric resampling model, which inherently addresses this issue (being not dependent on two moments alone) by having the entire conditional distribution resampled from yesteryear flows. On the basis of the observed streamflow Q_t and the predictors X_t , with the first two principal components from PCA, the conditional distribution of streamflows is estimated based on the parametric regression and semiparametric resampling methods.

[21] Given the observed summer inflows had skewness of 1.9, we applied cube root transform to bring the data to normal distribution. After transformation, the skewness of the transformed flows reduced to 0.25. Figure 4b suggests that the relationship between the first principal component (PC1) and the transformed flows (cube root) is linear with the exception of few outlying observations, which were primarily contributed by the summer hurricanes ($\sqrt[3]{Q} \geq 10$). Thus, for the parametric regression model, we consider the cube root transformed flows as the predictand and the first two principal components of SSTs as the predictors. Residual analyses of regression based on quantile plots and skewness ($=0.08$) tests on the residuals showed that the normal assumption of the transformed flows is valid. The estimate of the conditional mean and conditional standard deviation of the transformed flows are obtained from the regression estimate and from the point forecast error respectively. Using the conditional mean and conditional standard deviation of the transformed flows, we generate

Table 1. Performance of Individual Model Forecasts and Various Multimodel Schemes Under Leave-One-Out Cross-Validated Forecasts and Adaptive Forecasts for Two Different Strategies of Choosing the Number of Neighbors K (Varying K_t)^{a,b}

Model/Multimodel	Cross-validated (1928–2005)			Adaptive Forecasts (1976–2005)		
	$\hat{\rho}$	$\overline{RPS}$	$\overline{RPSS}$	$\hat{\rho}$	$\overline{RPS}$	$\overline{RPSS}$
Resampling (Res)	0.38	0.429	−0.020	0.44	0.517	−0.069
Regression (Reg)	0.35	0.409	0.050	0.49	0.424	0.011
MM1 (Res + Clim)	0.45	0.389	0.097	0.47	0.403	0.035
MM2 (Reg + Clim)	0.43	0.386	0.110	0.48	0.371	0.064
MM3 (MM1 + MM2)	0.47	0.381	0.121	0.51	0.374	0.061
MM4 (Res + Reg + Clim)	0.44	0.384	0.110	0.50	0.383	0.052
MM5 (Res + Reg)	0.39	0.392	0.082	0.49	0.396	0.037
MM6 (equal weights)	0.39	0.397	0.072	0.49	0.399	0.037
MM7 (long-term skill)	0.39	0.397	0.072	0.49	0.400	0.035

^aAll the performance evaluation measures are calculated based on 78 years of data for leave-one-out cross-validated forecasts and 30 years for the adaptive forecasts from 1976 to 2005.

^b $\hat{\rho}$, Pearson correlation between the observed flows and the condition mean of forecasts; $\overline{RPS}$, average rank probability score; $\overline{RPSS}$, average rank probability skill score; MM, multimodel ensembles; Clim, climatological ensembles.

ensembles from the normal distribution and transform it back into the original space, by taking cube, to obtain the conditional distribution of flows, $Q_{i,t}^n$.

[22] The second approach, the semiparametric resampling algorithm developed by Souza and Lall [2003], is data-driven and estimates the conditional distribution by resampling the observed flow values that were under climatic conditions similar to the current predictor conditions. To identify similar climatic conditions or the neighbors, the semiparametric method employs regression coefficients (estimated between cube root transformed flows and the predictors for Falls Lake inflow forecasts) as weights to compute the distance between the current conditions and the past predictor conditions. For additional details, see Souza and Lall [2003]. A total of thirty-three number of neighbors was considered for obtaining the forecasts from the semiparametric model. This was chosen by choosing the neighbor that gave the highest correlation between observed flows and the conditional mean of the leave-one-out cross-validated forecasts, whose procedure is detailed below.

[23] Leave-one-out cross-validation is a rigorous model validation procedure that is carried out by leaving out the predictand and predictors from the observed data set (Q_t, X_t , $t = 1, 2, \dots, n$) for the validating year and the model is developed using the remaining $n - 1$ observations [Craven and Whaba, 1979]. For instance, to develop retrospective leave-one-out forecasts from parametric regression, a total of n regression models are developed by leaving out the observation in each validating year. By employing the developed forecasting model with $n - 1$ observations, the left out observation (Q_{-t} , with $-t$ denoting the validating year) is predicted by using the state of the predictor/principal components (X_{-t}) in that validating year. By utilizing the two principal components from PCA, we develop both leave-one-out cross-validated retrospective streamflow forecasts and adaptive streamflow forecasts for the season, JAS using the two statistical models. The conditional distribution of streamflows forecasts from each model is represented in the form of ensembles.

[24] To obtain adaptive streamflow forecasts, we develop the forecasting models based on the observed streamflow and the two dominant principal components from 1928–

1975 and employ the developed model to predict the streamflow for a 30-year period from 1976–2005. Table 1 gives various performance measures of the cross-validated forecasts and adaptive forecasts for both models. Figure 5 shows the adaptive streamflow forecasts for both parametric regression and the semiparametric model. The correlation between the observed streamflows and the ensemble mean of the cross-validated forecasts for resampling and regression approach is 0.38 and 0.35 respectively, which is significant for the 78 years of observations. From Table 1, we infer that the correlation between the observed streamflows and the ensemble mean of the adaptive forecasts is 0.44 and 0.49 for resampling (Figure 5a) and regression (Figure 5b) approach, respectively. Table 1 also shows other performance evaluation measures such as $\overline{RPS}$, and $\overline{RPSS}$ for adaptive and leave-one-out cross-validated forecasts for both models. On the basis of $\overline{RPS}$ and $\overline{RPSS}$, we find that regression model seems to perform better than the resampling model under both cross-validated and adaptive forecasts. Since the correlation between the observed and the ensemble mean is significant for both models under leave-one-out cross-validated forecasts and adaptive forecasts, we combine the forecasts from both these models for developing multimodel ensembles for the Falls Lake system.

5. Multimodel Combination Based on Predictor State Space—Results and Analyses

[25] In this section, we apply the multimodel combination algorithm discussed in section 3.1 to combine the forecasts from individual models – parametric regression and semiparametric resampling - with climatological ensembles. For this purpose, we consider seven different multimodels by combining: (1) resampling and climatology ensembles (MM1), (2) regression and climatology ensembles (MM2), (3) MM1 and MM2 ensembles (MM3), (4) resampling, regression and climatology ensembles at one step (MM4), (5) resampling and regression ensembles (MM5), (6) resampling and regression ensembles with equal weights (MM6), and (7) resampling and regression ensembles based on their long-term predictability (MM7). Multimodel MM6, which is analogous to pooling of ensembles without optimal combination and multimodel MM7 provide the baseline

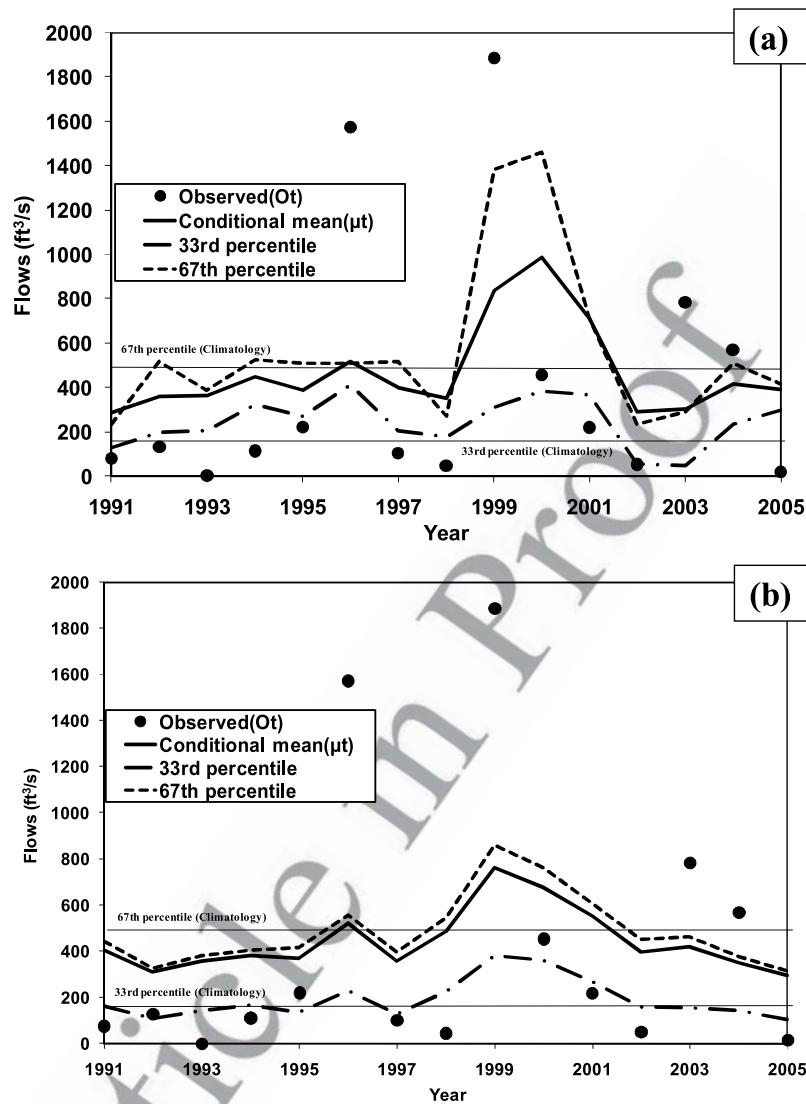


Figure 5. Performance of individual models in predicting observed streamflows during 1991–2005 for the Falls Lake. (a) Semiparametric resampling model of Souza and Lall [2003]. (b) Parametric regression. Forecasts from both models were obtained by using the observed streamflows during JAS and predictors (PC1 and PC2 in Figure 3) for the period 1928–1990. The horizontal lines denote the 33rd percentile ($Q < \text{below-normal category}$) and 67th percentile ($Q \geq \text{above-normal category}$) of observed flows.

734 comparison with the current state of the art in multi-
 735 model ensembles development. For MM6, we give equal
 736 weights by enforcing $w_i^{\text{Re } g} = w_i^{\text{Re } s} = 0.5$ for each year in
 737 equation (3). To develop ensembles from MM7, we assign
 738 $\lambda_i^{\text{Re } s} = 1/\overline{RPS}^{\text{Re } s}$, $\lambda_i^{\text{Re } g} = \overline{RPS}^{\text{Re } g}$, where $\overline{RPS}^{\text{Re } s} = 0.429$
 739 and $\overline{RPS}^{\text{Re } g} = 0.409$ denote the average RPS for resampling
 740 and regression models (from Table 1) indicating their long-
 741 term predictability.

742 [26] The motivation in considering climatology as one of
 743 the candidates in multimodel combinations (except MM5,
 744 MM6, and MM7) is based on the presumption that if the
 745 observation falls outside the predictive ability of all the
 746 models under certain predictor conditions, then climatology
 747 should be preferred over individual model forecasts. Fur-
 748 ther, recent studies have shown that a two step procedure of
 749 combining first each individual model forecasts separately
 750 with climatology and then combining the resulting “M”
 751 combinations at the second step improves the skill of

multimodel ensembles [Robertson et al., 2003; Goddard 752
 et al., 2003]. Since implementing the proposed multimodel 753
 algorithm requires selection of neighbors, we primarily 754
 employ the “Varying K_i ” approach for comparing the 755
 performance of seven multimodels with two candidate 756
 models. In the next section, however, we employ the fixed 757
 number of neighbors just for demonstrating the proposed 758
 multimodel combination methodology. 759

5.1. Skill of Individual Models From Predictor State Space 761

[27] The primary motivation in the proposed approach for 763
 multimodel combination is to evaluate competing models’ 764
 predictability in the neighborhood of the predictor state and 765
 give appropriate weights based on equation (3) for all the 766
 models to develop multimodel ensembles. Figure 6 analyzes 767
 the performance of regression and resampling streamflow 768
 forecasts shown earlier (Figure 5) from the predictor state 769

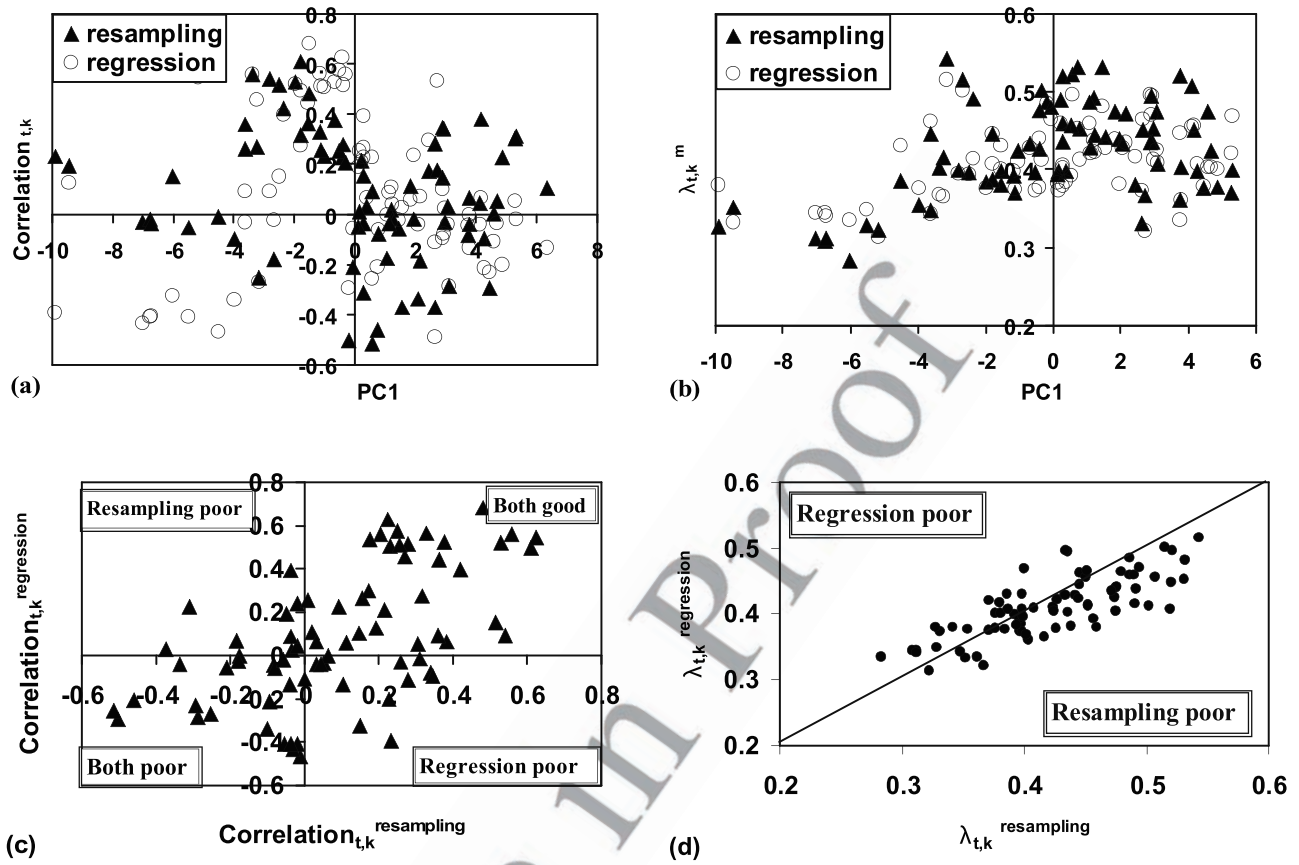


Figure 6. Performance of individual models from the predictor state space by considering $K = 15$ neighbors. (a) Correlation vs. PC1. (b) $\lambda_{t,K}^m$ vs. PC1. (c) Correlation of regression vs. correlation of resampling. (d) $\lambda_{t,K}$ of regression vs. $\lambda_{t,K}$ of resampling. $\lambda_{t,K}^m$ is computed from the leave-one-out cross-validated forecasts given in Table 2 for both candidate models and by assuming $K = 15$ in equation (2). Correlation is computed between the observed streamflows and the ensemble mean of the leave-one-out cross-validated forecasts by considering 15 neighbors from the current state. Note the consistent poor performance of regression model in Figure 6c as well as the high negative values of PC1.

space by considering a fixed $K = 15$. The performance of the streamflow forecasts are evaluated using average RPS and correlation over the identified neighbors in the predictor state space. From Figure 6a, one may prefer to choose forecasts from resampling model instead of forecasts from parametric regression since the correlation of the regression is negative if the dominant/first principal component (PC1) < -4 . This could also be seen from Figure 6b with the RPS of resampling being less than the RPS of regression. Figures 6c and 6d show the relative performance of both models against each other. From Figure 6c, we can see that one would prefer climatological ensembles over forecasts from candidate models, particularly when correlations estimated from the neighborhood on both models are negative. From Figure 6d, we can also identify conditions (above the diagonal line) with the RPS of regression model being higher than that of RPS of resampling indicating the poor performance of the regression ensembles. Thus the multimodel combination algorithm in section 3.1 primarily identifies these conditions based on RPS using equation (2) and develops a general procedure for multimodel combination. The next section compares the performance of seven different multimodels in developing cross-validated and adaptive forecasts using “Varying K_t ” approach.

5.2. Performance of Multimodel Forecasts

[28] The performance of seven different multimodel combinations is compared in Table 1 with the performance of two individual models under leave-one-out cross-validated forecasts and under adaptive forecasts for the period 1976–2005. For developing climatological ensembles for multimodel combinations, we simply bootstrap the observed JAS streamflows into Falls Lake assuming each year has equal probability of occurrence. This is a reasonable assumption since there is no year to year correlation between the summer flows.

5.2.1. Leave-One-Out Cross-Validated Streamflow Forecasts

[29] Under leave-one-out cross-validated forecasts, we can clearly see that forecasts from MM3 perform better than the regression and resampling models’ forecasts based on all the three statistics considered (Table 1). MM3 also performs better than the rest of the six multimodels. To ensure the skill exhibited by MM3 is statistically significant from the rest of the six multimodels and two statistical models, we perform detailed hypothesis tests based on $\overline{RPS}$. The hypothesis tests employed is based on the nonparametric approach under which we resample the computed RPS_t^m

t2.1 **Table 2.** Hypothesis Testing of $\overline{RPS}$ for Various Models Listed in Table 1 and This Table^a

t2.2	Model A								
t2.3	Model B	Res	Reg	MM1	MM2	MM3	MM4	MM5	MM6
t2.4	Regression (Reg)	0.85							
t2.5	MM1 (Res + Clim)	>0.99	0.86						
t2.6	MM2 (Reg + Clim)	0.98	0.96	0.63					
t2.7	MM3 (MM1 + MM2)	0.99	0.97	0.96	0.83				
t2.8	MM4 (Res + Reg + Clim)	>0.99	0.98	0.80	0.59	0.18			
t2.9	MM5 (Res + Reg)	>0.99	0.95	0.40	0.32	0.16	0.16		
t2.10	MM6 (equal weights)	>0.99	0.90	0.27	0.21	0.09	0.05	<0.01	
t2.11	MM7 (long-term skill)	>0.99	0.90	0.27	0.20	0.08	0.05	<0.01	0.33

^aEntries in the table provide the percentile value of the test statistic $\overline{RPS}^A - \overline{RPS}^B$ (from Table 1 for the chosen Model A and B) from the resampled null distribution. For a given pair, Models A and B are indicated by the topmost row and leftmost column, respectively.

t2.12

817 from each model for each year to develop the null
818 distribution [Hamill, 1999]. Details on hypothesis test and
819 the test statistic suggested by Hamill [1999] are briefly
820 discussed in Appendix B.

821 [30] Table 2 provides the percentile values of the test
822 statistic $\overline{RPS}^A - \overline{RPS}^B$ based on the constructed null
823 distribution using (B4) and (B5) for each pair of models A
824 and B. Thus, if $\overline{RPS}^B < \overline{RPS}^A$ ($\overline{RPS}^B > \overline{RPS}^A$), then we
825 would like the observed test statistic, $\overline{RPS}^A - \overline{RPS}^B$, to fall
826 under the right (left) tail of the null distribution to reject the
827 null hypothesis. On the basis of $\overline{RPS}$ given in Table 1 for
828 the chosen models A and B, one could reject or accept the
829 null hypothesis based on the chosen level of significance, α .
830 For instance, to test whether $\overline{RPS}^{MM1} = \overline{RPS}^{Res}$ at $\alpha = 10\%$
831 significance level, from Table 2, we can see that the
832 percentile value of the test statistic is >0.99 implying a
833 p-value < 0.01. Hence we can reject the null hypothesis
834 $\overline{RPS}^{MM1} = \overline{RPS}^{Res}$ and accept the alternate hypothesis
835 $\overline{RPS}^{MM1} < \overline{RPS}^{Res}$, which implies that the skill exhibited by
836 MM1 is significantly greater than that of the skill of the
837 resampling model. Thus we utilize the information given in
838 Table 2 to compare Model A (in columns) with Model B (in
839 rows) to test whether the increased skill exhibited by the
840 multimodel forecasts are statistically significant compared
841 to the skill of individual model forecasts. We also discuss
842 various related issues that include utility of climatological
843 ensembles in improving the skill and on the effectiveness of
844 the two-step multimodel combination approach.

845 5.2.1.1. Comparison Between Individual Models and 846 Multimodels

847 [31] From Table 2, we can clearly see that the reduced
848 $\overline{RPS}$ of the multimodels (in Table 1) are statistically
849 significant from the $\overline{RPS}$ of the resampling and regression
850 models. This could be inferred from the percentiles of the
851 test statistic $\overline{RPS}^A - \overline{RPS}^B$ from Table 1 (Model A:
852 Resampling or Regression) falling on the right tails of the
853 constructed null distribution with the p-value being less than
854 0.10 for all the multimodels (Model B: MM3-MM7).
855 Similarly, we also see that there is no significant difference
856 between $\overline{RPS}$ of the regression model and $\overline{RPS}$ of the
857 resampling ($\alpha = 0.10$) forecasts. Further, we also understand
858 that combining individual models with climatological
859 ensembles alone (Model B: MM1 and MM2) leads to an
860 improved representation of the conditional distribution.

5.2.1.2. Performance of Multimodel Forecasts Based on Equal Weights and Long-Term Predictability

861 [32] To compare the performance of the proposed multi-
862 model combination algorithm in section 3.1 with existing
863 multimodel techniques, we have included multimodel fore-
864 casts based on equal weights (MM6: pooling of ensembles)
865 and based on each model's long-term predictability (MM7).
866 From Table 2, we infer that $\overline{RPS}$ of MM5 (Resampling and
867 Regression alone) is significantly lesser than $\overline{RPS}$ s of MM6
868 and MM7. Similarly, comparing the performance of MM6
869 and MM7 with MM3 and MM4, we observe that $\overline{RPS}$ of
870 MM3 and MM4 are statistically smaller than $\overline{RPS}$ s of MM6
871 and MM7. This shows that combining multiple models
872 based on predictor conditions performs better than combin-
873 ing multiple models by pooling and based on long-term
874 predictability.

5.2.1.3. Utility of Climatological Ensembles

875 [33] Multi-models (MM1, MM2, MM3 and MM4) use
876 climatological ensembles to develop multimodel forecasts.
877 An interesting fact that we can observe from Table 2 is that
878 adding climatological ensembles with individual models to
879 develop multimodel forecasts (Model B: MM1, MM2) leads
880 to a $\overline{RPS}$ that is statistically smaller than the individual
881 models' $\overline{RPS}$. Similarly, comparing the performance of
882 MM1 and MM2 (Model A) with MM5, MM6 and MM7
883 (Model B), we see that the percentiles of the test statistic are
884 around 0.2–0.4 implying no statistical significance, which
885 suggests that combining multiple models without climatol-
886 ogy may not lead to significant reduction in $\overline{RPS}$. However,
887 adding climatology to multiple models which is represented
888 by MM4 (Model B) reduces $\overline{RPS}$, but it is not statistically
889 significant compared to $\overline{RPS}$ s of MM1 and MM2. Further,
890 comparing the performance of MM4 (Model B) with MM5
891 (Model A) shows that there is certainly an improvement in
892 adding climatology ensembles (p -value = 0.16). To
893 summarize, adding climatological ensembles with indivi-
894 dual model forecasts certainly leads to improvement in the
895 conditional distribution representation, but improvements
896 are much more substantial upon combining climatological
897 ensembles with forecasts from multiple candidate models.

5.2.1.4. Effectiveness of Two-Step Multimodel Combination

901 [34] To understand the effectiveness of the two-step
902 procedure (individual models are combined first with cli-
903 matological ensembles and the resulting “M” forecasts are
904 905

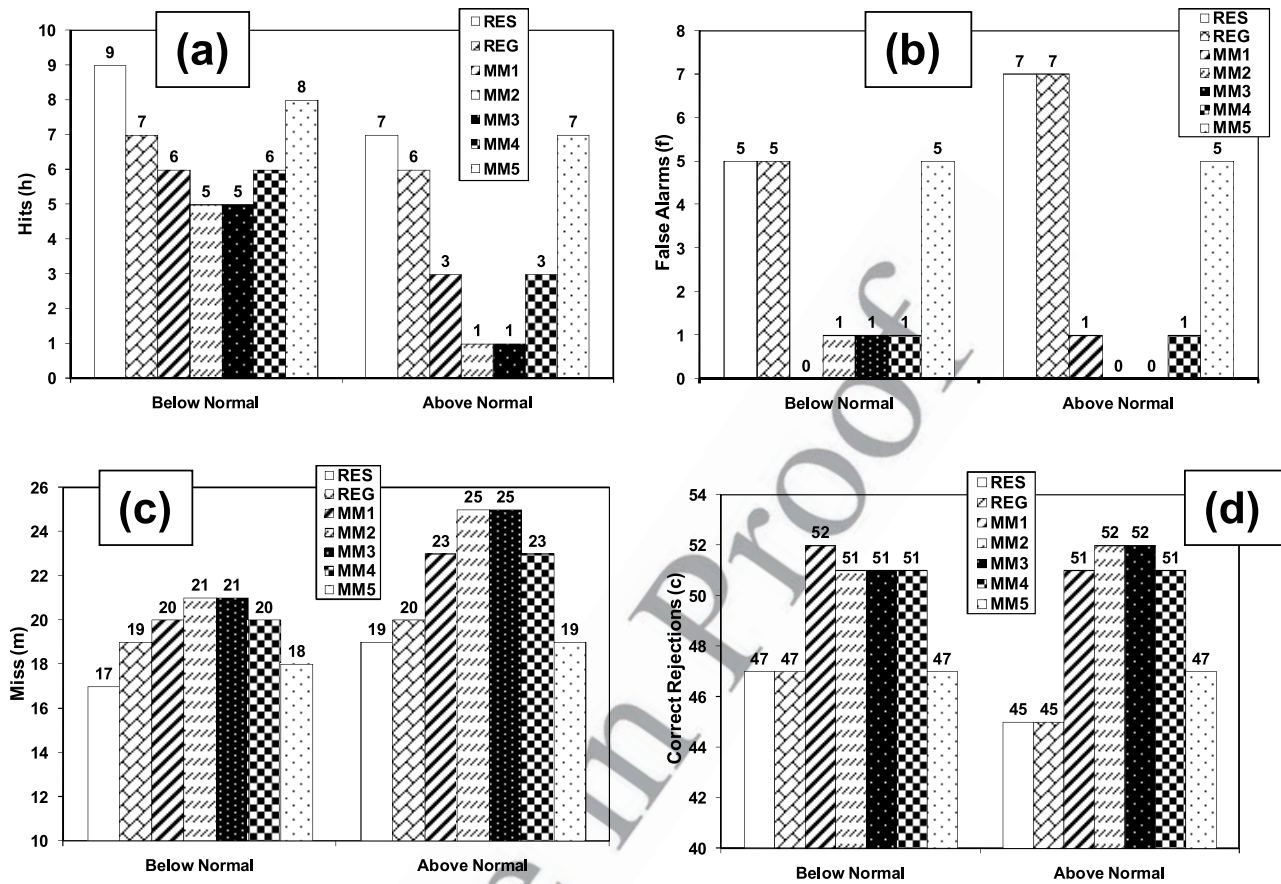


Figure 7. Performance evaluation of multimodel forecasts based on contingency table. Numbers of (a) hit, (b) false alarms, (c) misses, and (d) correct rejections are given for below-normal and above-normal categories for a threshold probability of 0.5.

combined at the second step) of multimodel combination, we compare the performance of MM3 (Model A – two step combination) with MM4 (Model B – One step combination). From Table 2, the computed test statistic $RPS^{MM4} - RPS^{MM3}$ has a percentile value of 0.18 under the constructed null distribution. This indicates that the null hypothesis $RPS^{MM4} = RPS^{MM3}$ can be rejected if one chooses a significance level of $\alpha = 0.10$.

[35] Thus we summarize that though the difference in RPS between MM3 and MM4 is not significant (p-value is 0.18), the two-step procedure of developing multimodel forecasts with climatology overall improves the conditional distribution estimation, as indicated by smaller average RPS , than one-step combination. To recapitulate, for a significance level of $\alpha = 0.10$, average RPS of MM3 is significantly less than the average RPS of MM1, MM6 and MM7. Upon comparing MM3's performance with the rest of the multimodels (MM2, MM4 and MM5), we find that average RPS of MM3 is weakly significant against average RPS s of MM2 (p-value = 0.17), MM4 (p-value = 0.18) and MM5 (p-value = 0.17). This could possibly be due to the slightly better performance of regression model in comparison to the resampling model.

5.2.1.5. Performance of Multimodel Under Different Flow Regimes

[36] We also compared the performance of resampling, regression models with the best performing multimodel

MM3 under different flow regimes. On the basis of this, the average RPS of resampling, regression and MM3 are 0.524, 0.480 and 0.460 respectively under below-normal flow conditions. Under above-normal flow conditions, the average RPS of resampling, regression and MM3 are 0.276, 0.237 and 0.214 respectively, whereas the average RPS of resampling, regression and MM3 are 0.487, 0.509 and 0.480 respectively under above-normal flow conditions. This indicates clearly that the regression (resampling) performs better than resampling (regression) during below-normal (above-normal) flow conditions. But, multimodel (MM3), since it constitutes more number of ensembles from the best performing model contingent on the predictor state, performs better than the two individual models under all flow conditions.

5.2.1.6. False Alarms and Missed Targets

[37] One main purpose of performing multimodel combination based on predictor state is to reduce false alarms and missed targets in the issued forecasts. Figure 7 summarizes the comparison between two candidate models, regression and resampling, and five multimodels MM1, MM2, MM3, MM4 and MM5, based on number of hits, false alarms, missed targets and correct rejections for below-normal and above-normal tercile categories. We are not reporting the performance of MM6 and MM7, since they are exactly the same as that of MM5 under both tercile categories. We choose a threshold probability of 0.5 to issue

t3.1 **Table 3.** Average Brier Score, Average Reliability, Average Resolution, and Average Ignorance for Below-Normal and Above-Normal Categorical Forecasts Obtained From the Leave-One out Cross-Validated Forecasts^a

t3.2	Model/Multimodel	Below Normal				Above Normal			
		$\overline{BS}$	$\overline{REL}$	$\overline{RES}$	$\overline{IGN}$	$\overline{BS}$	$\overline{REL}$	$\overline{RES}$	$\overline{IGN}$
t3.3									
t3.4	Resampling (Res)	0.217	0.048	0.052	0.886	0.212	0.019	0.029	0.933
t3.5	Regression (Reg)	0.199	0.015	0.037	0.883	0.210	0.018	0.029	0.847
t3.6	MM1 (Res + Clim)	0.199	0.019	0.041	0.857	0.204	0.005	0.023	0.847
t3.7	MM2 (Reg + Clim)	0.196	0.013	0.037	0.866	0.207	0.005	0.019	0.838
t3.8	MM3 (MM1 + MM2)	0.197	0.011	0.035	0.861	0.205	0.005	0.021	0.840
t3.9	MM4 (Res + Reg + Clim)	0.197	0.009	0.033	0.853	0.203	0.006	0.024	0.841
t3.10	MM5 (Res + Reg)	0.203	0.032	0.051	0.864	0.204	0.016	0.033	0.868
t3.11	MM6 (equal weights)	0.204	0.020	0.036	0.870	0.207	0.018	0.033	0.874
t3.12	MM7 (long-term skill)	0.204	0.013	0.029	0.875	0.208	0.020	0.033	0.873

t3.13 ^aThe average ignorance score of climatological forecasts is $\log_2 n_c$, where n_c is the number of categories of forecasts. For tercile forecasts, $\overline{IGN}_{\text{climatology}} = 1.585$.

960 a forecast alarm under a particular tercile category. It is
 961 important to note that contingency analysis of forecasts (in
 962 Figure 7) is very sensitive to the chosen threshold and the
 963 number of forecasts [Wilks, 1995]. From Figure 7b, we
 964 clearly see that the false alarms are clearly dropping with all
 965 the three multimodel forecasts. However, multimodel
 966 forecasts tend to increase the missed targets and reduce
 967 the number of hits. In the case of MM3, the drop in the
 968 number of hits is 4 and 6 for below-normal and above-
 969 normal categories, but the number of false alarms reduces
 970 considerably. This suggests that adding climatology to
 971 individual models certainly reduces the number of hits and
 972 tends to produce multimodel forecasts (MM1, MM2, MM3
 973 and MM4) with reduced resolution (ability of the model to
 974 distinguish from climatology). This could also be inferred
 975 from the plot for MM1 and MM2 from Figure 7. Reasons
 976 for reduced false alarms and increased missed targets are
 977 discussed further in the next Section.

978 5.2.1.7. Reliability and Resolution of Forecasts

979 [38] A better measure to evaluate categorical probabilistic
 980 forecasts would be to quantify the average Brier score
 981 (Table 3). Table 3 also provides average reliability ($\overline{REL}$)
 982 and average resolution ($\overline{RES}$) along with average ignorance
 983 ($\overline{IGN}$) for regression, resampling and seven multimodel
 984 forecasts for below-normal and above-normal categories.
 985 Brier score is similar to RPS, but more appropriate for
 986 dichotomous events (e.g., whether the observed event is
 987 below-normal or not). Brier score, specifying the squared
 988 error in categorical forecasts, could be split into reliability,
 989 resolution and uncertainty. For additional details, see Mason
 990 [2001]. For $\overline{BS}$ to be close to zero, it is important that the
 991 reliability term should be close to zero and resolution should
 992 be large [Wilks, 1995]. Average ignorance, otherwise known
 993 as log-scoring rule, is a double-valued function of average
 994 Brier score and is a better measure in evaluating the
 995 probabilities forecasts, since it generalizes the categorical
 996 forecasts beyond the binary case [Roulston and Smith,
 997 2003]. The forecasts are considered to be useful, if $\overline{IGN}$ of
 998 forecasts is less than the $\overline{IGN}$ of climatology.

999 [39] From Table 3, we clearly understand that the average
 1000 Brier scores of MM3 and MM4 are smaller than the average
 1001 Brier scores of resampling and regression models. Since
 1002 Brier scores represent squared error in probabilities, even a
 1003 small difference could help in reducing the error in estimat-
 1004 ing the probability under a particular category. For instance,

the difference between regression and MM3 average Brier 1005
 scores is only 0.0024 in identifying the below-normal 1006
 category, but this amounts to an average error reduction 1007
 of 5% ($\sqrt{0.0024}$) in quantifying the below-normal category. 1008
 Thus, based on average Brier scores, we infer that multi- 1009
 model forecasts (MM3) improve the probability of predict- 1010
 ing the appropriate tercile category of occurrence. 1011

[40] To ensure the difference between $\overline{BS}$ of MM3 and 1012
 $\overline{BS}$ s of individual models are statistically significant from 1013
 the $\overline{BS}$ of resampling and regression models, we performed 1014
 a hypothesis testing on Brier scores similar to the hypothesis 1015
 testing on $\overline{RPS}$ as suggested by Hamill [1999]. The p-values 1016
 of hypothesis tests on comparing $\overline{BS}$ s of regression and 1017
 resampling models (Model A) with $\overline{BS}$ of MM3 (Model B) 1018
 are 0.02 and 0.25 for resampling and regression respectively 1019
 under below-normal category. Similarly, the p-values for 1020
 above-normal category upon comparing $\overline{BS}$ s of two 1021
 candidate models' forecasts with MM3 forecasts are 0.01 1022
 and 0.02 for resampling and regression respectively. We did 1023
 not perform separate hypothesis tests for comparing the 1024
 Brier score of MM4, MM5, MM6 and MM7 with the 1025
 individual model forecasts. Thus we infer that the $\overline{BS}$ of 1026
 MM3 forecasts is significant in comparison to the two 1027
 candidate models in identifying the above-normal category. 1028

[41] Figures 8a and 8b compare the reliability of MM3 1029
 forecasts with the reliability of individual model forecasts 1030
 for below-normal and above-normal categories. Reliability 1031
 diagrams provide information on the correspondence be- 1032
 tween the forecasted probabilities for a particular category 1033
 (e.g., above-normal, normal and below-normal) and how 1034
 often (frequency) that category is being observed under that 1035
 forecasted probability [Wilks, 1995]. For instance, if we 1036
 forecast the occurrence of below-normal category as 0.9 1037
 over n_1 years ($n_1 \leq n$), then we expect the actual outcome to 1038
 fall under below-normal category for $0.9 * n_1$ times over the 1039
 long-term. To construct Figure 8, we utilized leave-one-out 1040
 cross-validated forecasts and divided the forecasted prob- 1041
 ability for each category into percentiles. Table 4 provides 1042
 the total number of forecasts that were issued with different 1043
 forecast probabilities under below-normal and above- 1044
 normal categories. 1045

[42] Though MM3 and MM4 show overconfidence 1046
 (increased observed relative frequency) for high forecast 1047
 probabilities (0.6 for below-normal and 0.6 for above- 1048
 normal) (Figure 8), we can infer from Table 3 that the 1049

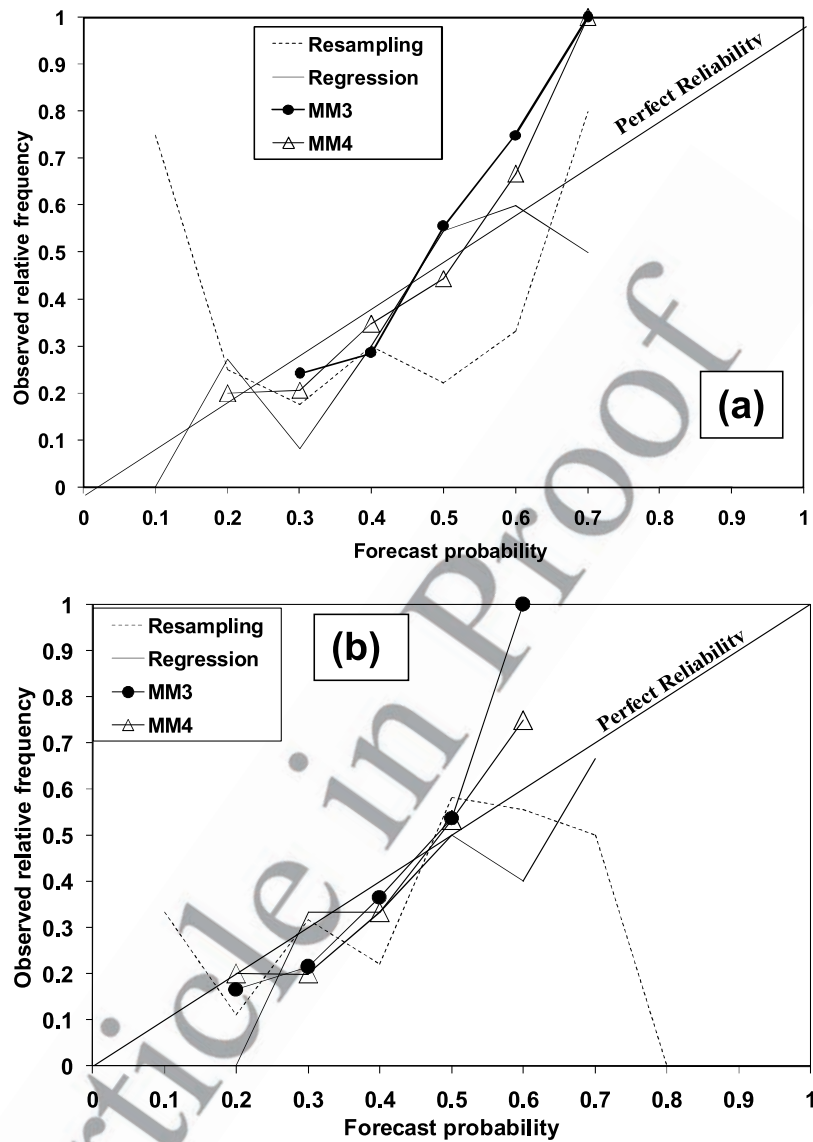


Figure 8. Comparison of reliability of cross-validated forecasts from regression, resampling, and MM3 forecasts for (a) below-normal and (b) above-normal categories.

average reliability scores of MM3 and MM4 are smaller than that of the rest of the models. The increased observed relative frequency could also be due to smaller number of forecasts issued under those forecast probabilities. For instance, we can infer from Table 4 that the number of forecasts issued with a forecast probability of 0.6 for MM3 under above-normal category is just 1. Similarly, under below-normal category, we see that the number of forecasts issued for MM3 and MM4 are 4(2) and 3(3) for a forecast probability of 0.6 (0.7) respectively. For predicting above-normal category, MM1, MM2 and MM3 have same reliability, but MM4, combining multiple models with climatology, has the lowest Brier score. This is primarily because MM4 shows higher resolution than MM1, MM2 and MM3. [43] An interesting observation from Table 3 is that combining multiple models with climatology always reduces the resolution, but increases the reliability of forecasts. For instance, resampling is over-confident (high observed relative frequency) for a forecast probability of

Table 4. Number of Forecasts Issued With Various Forecast Probabilities Under Below-Normal and Above-Normal Categories^a

Forecast Probability	Below Normal				Above Normal				
	Res	Reg	MM3	MM4	Res	Reg	MM3	MM4	
0.0	0	0	0	0	0	0	0	0	t4.4
0.1	4	1	0	0	3	5	0	0	t4.5
0.2	24	22	2	10	18	11	6	10	t4.6
0.3	17	12	33	29	22	27	28	25	t4.7
0.4	10	20	28	23	9	12	30	24	t4.8
0.5	9	11	9	9	12	10	13	15	t4.9
0.6	3	5	4	3	9	10	1	4	t4.10
0.7	10	6	2	3	4	3	0	0	t4.11
0.8	1	0	0	1	1	0	0	0	t4.12
0.9	0	1	0	0	0	0	0	0	t4.13
1.0	0	0	0	0	0	0	0	0	t4.14

^aRes, resampling; Reg, regression.

t4.15

Table 5. Sensitivity of the Multimodel Algorithm in Figure 2 to Two Different Performance Evaluation Metrics, Squared Error Between the Conditional Mean (SECM), and Observed Flows and $L = 1$ Norm Instead of RPS in Equation (2)^a

	Model/Multimodel	SECM			$L = 1$ Norm		
		$\bar{p}$	$\overline{RPS}$	$\overline{RPSS}$	$\bar{p}$	$\overline{RPS}$	$\overline{RPSS}$
Resampling (Res)		0.377	0.429	-0.020	0.377	0.429	-0.020
Regression (Reg)		0.353	0.409	0.050	0.353	0.409	0.050
MM1 (Res + Clim)		0.469	0.381	0.115	0.402	0.413	0.046
MM2 (Reg + Clim)		0.405	0.384	0.119	0.370	0.407	0.069
MM3 (MM1 + MM2)		0.464	0.376	0.134	0.408	0.408	0.063
MM4 (Res + Reg + Clim)		0.446	0.382	0.117	0.391	0.407	0.060
MM5 (Res + Reg)		0.415	0.388	0.092	0.386	0.410	0.042
MM6 (equal weights)		0.386	0.397	0.072	0.383	0.397	0.072
MM7 (long-term skill)		0.389	0.397	0.072	0.389	0.397	0.072

^a $L = 1$ norm is calculated using equation (A2).

0.1 under below-normal category, which leads to high resolution. Similarly, under above-normal category, we see that individual models have good resolution for a forecast probability of 0.7, whereas multimodels do not have any forecasts issued under that forecast probability. Reduced resolution from multimodel forecasts will naturally lead to reduced hits. This is in line with some of the earlier studies on multi-model combination [Barnston et al., 2003; Robertson et al., 2004a, 2004b]. This reduced resolution leads to increased missed targets under multimodel combinations with climatology, but reduces the false alarms which results from the overconfidence of individual models (Figure 7).

5.2.1.8. Ignorance Score and Utility of Multimodel Ensembles in Falls Lake Operation

[44] On the basis of average ignorance, we understand that all multimodels and individual models have reduced average ignorance in comparison to the ignorance of climatology. Further, under below-normal category, multimodels (MM1, MM2, MM3, MM4 and MM5) developed based on the algorithm in section 3.2 have lower ignorance score in comparison to MM6 and MM7 with MM3 having the smallest ignorance score of all the models. Under above-normal category, MM2 has the smallest average ignorance score with MM3 and MM4 performing almost similarly. Average ignorance score also suggests that adding climatology is important in improving the performance of multimodels, since MM5 has a higher ignorance score than other multimodels (MM1, MM2, MM3 and MM4).

[45] Utilizing the probabilistic streamflow forecasts from these three models, resampling, regression and multimodels, for invoking restrictions to improve the end of season target storage show that multimodels perform consistently in identifying appropriate (above-normal or below-normal) storage conditions in September and thus reducing false alarms in invoking restrictions (K. Golembesky et al., manuscript in preparation, 2008). To relate how the reduced RPS from multimodels would result in improving Falls Lake operation, (K. Golembesky et al., manuscript in preparation, 2008) also show that a 5% reduction in the risk of not meeting the end of season target storage (corresponding to 251.5 ft) during below-normal inflow years would require more than 10% reduction in water supply releases during the season. To summarize the dis-

cussion, the proposed multimodel algorithm in section 3.2 overall improves the predictability of tercile forecasts in comparison to the individual model forecasts and multimodel forecasts developed by pooling (MM6) and based on long-term predictability (MM7).

5.2.1.9. Sensitivity of the Algorithm to Performance Evaluation Metric

[46] The proposed algorithm in section 3.2 employs average RPS over the chosen K neighbors to evaluate the performance of given forecasting model. In this section, we evaluate the sensitivity of the algorithm to two different performance metrics, squared error between the conditional mean (SECM) and observed flows and $L = 1$ norm, in developing multimodel forecasts. Given ensembles of streamflow forecasts from an individual model, we estimate SECM from the conditional mean and $L = 1$ norm (using equation (A2)) for each year of forecasts. Thus we replace RPS in equation (2) with estimates of SECM and $L = 1$ norm to quantify the average performance of the given model over the considered “ K ” neighbors.

[47] Table 5 summarizes the average RPS and average RPSS for seven multimodels under the two performance evaluation metrics. From Table 5, we understand that multimodels, MM2, MM3 and MM4, developed from this study show reduced RPS than the RPS of individual models and over existing techniques on multimodel combination (MM6 and MM7). However, comparing the performance of multimodel forecasts from SECM and $L = 1$ norm (Table 5) with the performance of multimodel forecasts obtained using RPS (in Table 1), we clearly see that multimodel forecasts developed using SECM and RPS (as performance evaluation metric) perform similarly in developing improved multimodel forecasts. Under $L = 1$ norm, the reduction in RPS is slightly less indicating only a marginal improvement in developing multimodel combination. One possible reason for improved performance under SECM and under RPS as performance metric is in penalizing poor forecasts more severely by considering the squared error in flows and cumulative probabilities respectively. This suggests that the algorithm is sensitive to the choice of the performance evaluation metric of individual forecasts which therefore needs to be selected carefully to ensure improved multimodel forecasts. In the next section, we further evaluate the proposed algorithm by developing adaptive streamflow forecasts for the period 1991–2005.

5.2.2. Adaptive Forecasts

[48] Figure 9 shows the adaptive streamflow forecasts from MM3 for the period 1991–2005 by training the individual models based on the data available during 1928–1976. Table 1 provides various performance measures of adaptive forecasts for all the models based on the 30-year validation period. From Table 1, we can clearly see that the skill of MM3 and MM4 forecasts is much higher than the skill of individual models and the rest of the multimodels. We did not perform any hypothesis tests on the reported skill measures, since the skill of multimodel (MM3 and MM4) are shown to be significant than that of individual model forecasts based on cross-validated forecasts. From Figure 9a, we can clearly see that MM3 forecasts perform better than individual model forecasts shown in Figure 5. This could be understood by focusing on the forecasts from three models, resampling (Figure 5a),

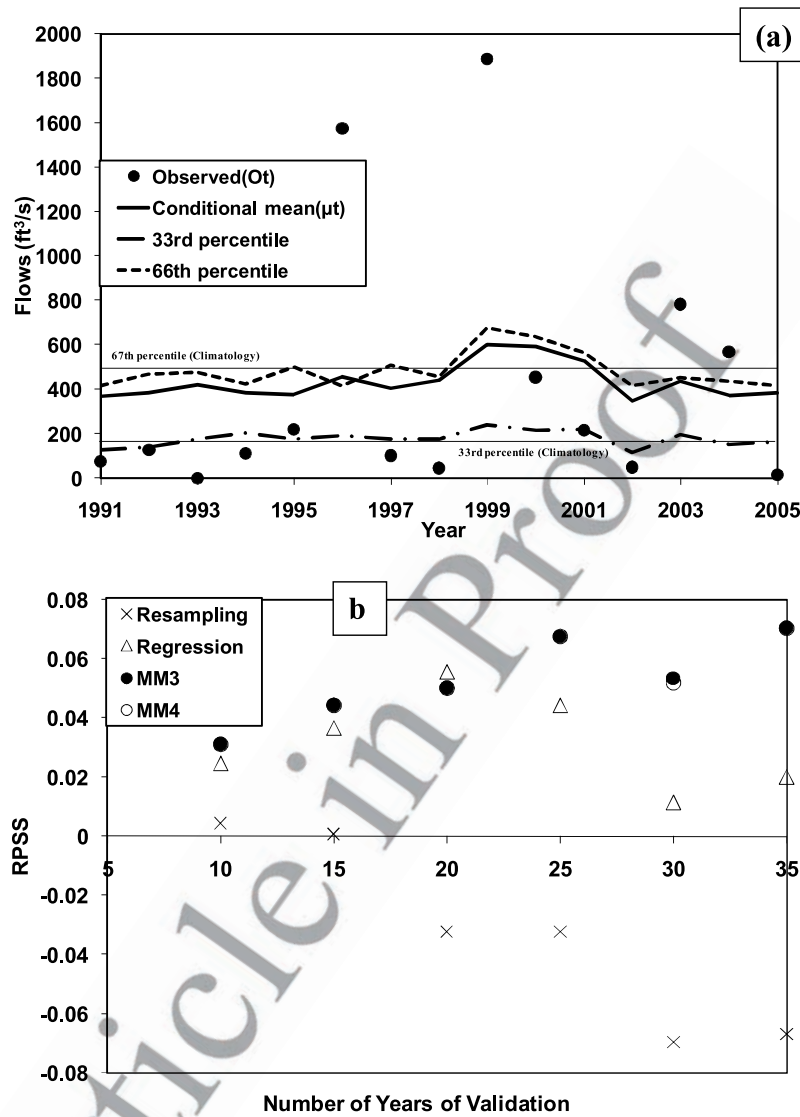


Figure 9. Performance of adaptive forecasts using the multimodel combination algorithm detailed in section 3.1. (a) Fifteen years of adaptive forecasts. (b) Average RPSs versus number of years of validation for evaluating adaptive forecasts.

regression (Figure 5b) and multimodel (Figure 9a), for years 2004 and 2002. For instance in 2004, an above-normal inflow year, regression incorrectly suggests it as a normal year (forecasted probabilities for below-normal (BN), normal (N) and above-normal (AN) are 0.36, 0.39 and 0.35 respectively), whereas resampling (forecasted BN, N, AN probabilities are 0.18, 0.38 and 0.44 respectively) and multimodel (forecasted BN, N, AN probabilities are 0.25, 0.36 and 0.39 respectively) correctly identifies it as an above-normal year. On the other hand in year 2002, a below-normal inflow year, regression (forecasted BN, N, AN probabilities are 0.34, 0.36 and 0.30 respectively) incorrectly suggests it as a normal inflow year, whereas resampling (forecasted BN, N, AN probabilities are 0.44, 0.32 and 0.24 respectively) and multimodel (forecasted BN, N, AN probabilities are 0.40, 0.31 and 0.29 respectively) correctly identifies it as a below-normal inflow year.

[49] The model did not predict accurately the high summer flows in years 1996 and 1999, mainly because these

flows were primarily due to hurricanes (1996: Hurricane Frank; 1999; Hurricane Floyd). Further, in both these years, more than 60% of seasonal flows occurred during the last 20 days of the season because of hurricanes. Thus, to better predict the flows in this year, one needs to develop updated forecasts all through the season [Sankarasubramanian *et al.*, 2007]. Analyses of the weights (figure not shown) showed that, as expected, weights of regression and resampling models were higher than the weights of climatology during below-normal and above-normal inflow years. However, weights of climatological ensembles were almost similar to the weights of resampling and resampling models particularly when flow values are in the normal category. Figure 9b compares the average RPSS between the individual model forecasts and MM3 and MM4 forecasts based on the number of years of validation of the adaptive forecasts. For each validation period (e.g., 30 years), we used the remaining data from 1928 for developing the forecasts. From Figure 9b, we can also see

that as the length of validation period increases, the performance of multimodel forecasts improves.

[50] Since we obtain the number of neighbors by “Varying K_i ”, which is going to vary depending on the predictor state, we also plot the distance between the chosen neighbors and the predictor state $PC1_i$. We also plot the distance of the chosen neighbor from the predictor state represented by each model. It is important to note that $PC1$ primarily denotes ENSO conditions (correlation between $PC1$ and $Nino3.4 = 0.36$), thus positive (negative) $PC1$ denoting the El Nino (La Nina) conditions. From Figure 10, we can clearly see that during normal conditions, the distance between the current predictor state and the chosen neighbor’s predictor state is small. During extreme predictor conditions, the distance is large since the nearest neighbor that results in reduced RPS could be far away from the conditioning state of the predictor. This shows that the multimodel algorithm developed in this study identifies similar predictor conditions in improving the performance of the individual model forecasts.

[51] To understand how the prediction intervals of multimodel forecasts compare with the prediction intervals of individual model forecasts, we show the ratio of interquartile range of the conditional distribution to the median of the conditional distribution (IQRM) in Figure 10b (analogous to coefficient of variation) for streamflow forecasts from three models, resampling, regression and multimodel (MM3), contingent on the principal component $PC1$. The reason to express this as a ratio, instead of a measure of the spread of the distribution, is that it incorporates the shift in the conditional distribution (conditional bias) across the models. On the basis of this, we observe the following based on the 78 years of forecasts: The IQRM of multimodel forecasts (MM3) were in between (greater than) the IQRM of regression in 45 years (31 years). In general, the IQRM of multimodel forecasts were slightly higher during below-normal years (low $PC1$ values), which could also arise since the median flow values are much smaller during below-normal inflow years.

6. Summary and Conclusions

[52] A new methodology for developing optimal multimodel combinations is presented and demonstrated by combining probabilistic streamflow forecasts from two low-dimensional streamflow forecasting models developed for the Falls Lake reservoir of the Neuse River Basin, NC. The proposed approach develops multimodel combinations by evaluating the skill of the candidate forecasting models’ contingent on the predictor(s) state. By identifying “ K_i ” similar conditions (to the current predictor state) using Mahalanobis distance, the algorithm evaluates the skill of the model by computing average RPS under “ K_i ” neighboring conditions. Using the average RPS, we obtain weights for drawing ensembles from each model to develop multimodel ensembles. This will lead to increased representation of ensembles from a candidate model if its performance is relatively better than other candidate models under particular predictor conditions. To evaluate the proposed scheme, we consider a total of seven different multimodels that includes multimodels with no optimal combination (pooling of ensembles) and multimodel combination based on long-term predictability. The study

also evaluates the ability of multimodels in improving tercile forecasts using Brier scores and using reliability plots.

[53] By comparing the performance of these seven multimodels with individual statistical (regression and resampling) models based on various measures such as correlation, average RPS and average rank probability skill score, we show that the forecasts from the proposed methodology have improved predictability compared to: (1) forecasts from individual models, (2) multimodel ensembles with no optimal combination (pooling), and (3) over multimodel forecasts based on long-term predictability. To ensure that the improved skill exhibited by the multimodel scheme is statistically significant over the skill of individual models, we perform detailed nonparametric hypothesis tests as suggested by Hamill [1999] by comparing the average RPS of the proposed scheme with other schemes. Results (p-value) from hypothesis tests show that the reduced RPSs shown by the seven multimodels are statistically significant than the RPSs of individual models. Comparing the performance of the proposed methodology with multimodel forecasts based on long-term predictability show that developing optimal model combinations contingent on the predictor certainly leads to improved predictability. Further the study also shows that adding climatological ensembles with individual statistical models also resulted in significant reduction in $\overline{RPS}$. However, depending on the threshold probability chosen, adding climatology could also potentially reduce the number of hits and increase missed targets, which primarily arises because of reduced resolution of multimodel forecasts. The two-step procedure of multimodel combination (MM3) – combining individual models first with climatology and then the resulting combinations are combined at the second step – and one step combination (MM4) that includes climatological ensembles, reduce the $\overline{RPS}$ considerably and perform better than the rest of the multimodels and low-dimensional statistical models. Both MM3 and MM4 also improve the reliability of forecasts and reduce the error, in terms of average Brier score, in developing tercile forecasts. The study also shows that adaptive forecasts from the proposed multimodel approach perform better than the adaptive multimodel forecasts obtained based on long-term predictability.

[54] The proposed multimodel algorithm aims at combining multiple models by analyzing the skill of the models contingent on the predictor state. As shown in Figure 6, if the predictability of all the models is poor under a particular condition, then our approach will eventually replace the multimodel ensembles with only climatological ensembles. Though the proposed approach is demonstrated by combining two low-dimensional streamflow forecasting models, it could be extended even to combine outputs from multiple GCMs. One difficulty in implementing this for multiple GCMs is on the necessity of having at least one dominant climatic mode that influences all the considered models. An obvious candidate for such a common predictor is to consider the leading principal components of global SSTs. However, as pointed out by Hagedorn *et al.* [2005], it is important to develop multimodel ensembles based on the region as well as by identifying SSTs or climatic modes that influence the streamflow/precipitation potential for the

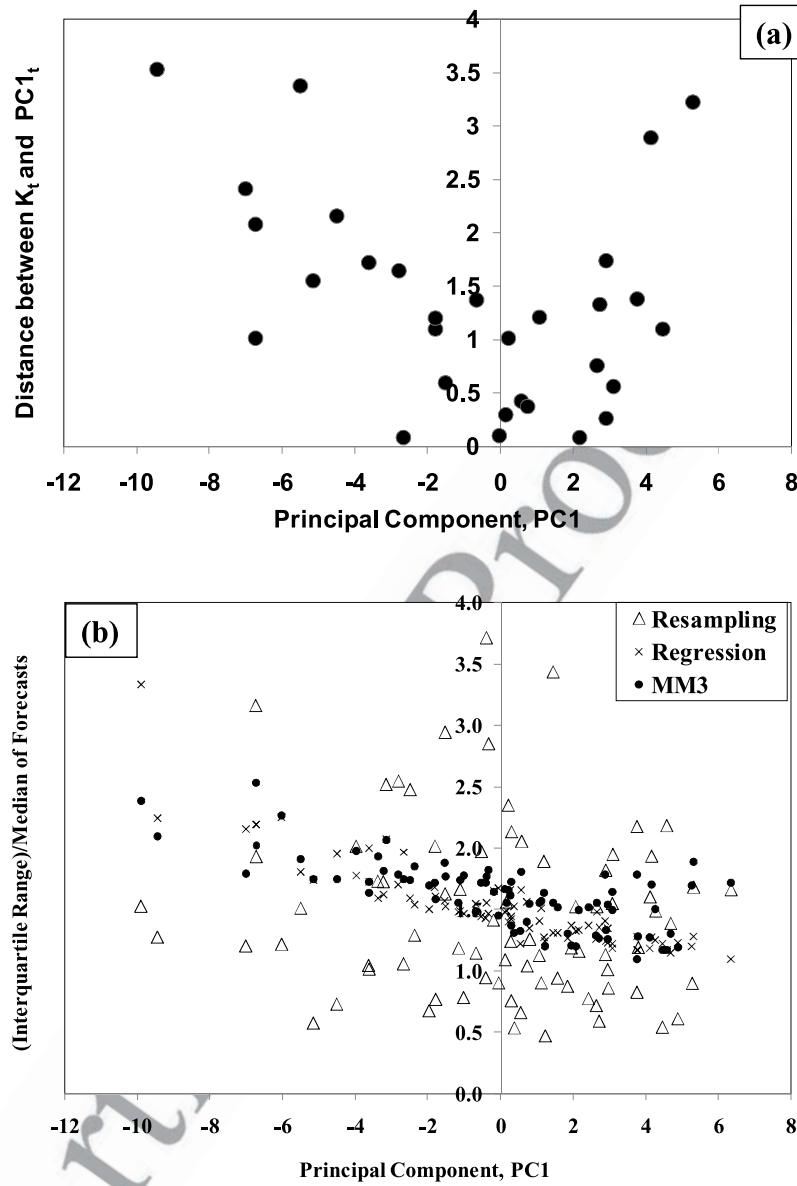


Figure 10. Role of predictor conditions in (a) choosing the neighbors and (b) in influencing the prediction interval of multimodel forecasts. The prediction intervals are expressed as the ratio of the interquartile range of the forecasts to the median of the forecasts.

1337 region. Our future studies will investigate the scope of
 1338 combining multiple GCMs to develop multimodel ensem-
 1339 bles of precipitation forecasts and in combining streamflow
 1340 forecasts obtained by different downscaling schemes.

1341 **Appendix A: Rank Probability Score and Rank** 1342 **Probability Skill Score**

1343 [55] Given that seasonal forecasts are better represented
 1344 probabilistically using ensembles, expressing the skill of the
 1345 forecasts using correlation requires summarizing the fore-
 1346 casts using some measures of central tendency such as mean
 1347 or median of the conditional distribution, which does not
 1348 give any credit to the probabilistic information in the
 1349 forecast. Rank Probabilistic Skill Score (RPSS) computes
 1350 the cumulative squared error between the categorical fore-
 1351 cast probabilities and the observed category in relevance to

a reference forecast [Wilks, 1995]. Here category represents
 1352 dividing the climatological/observed streamflow, Q , into $d =$
 1353 $1, 2, \dots, D$ divisions and expressing the marginal probabilities
 1354 as $P_d(Q)$. Typically, the divisions are made equal probabil-
 1355 itically with $O = 3$ categories known as terciles with each
 1356 category having $1/3$ probability of occurrence. These three
 1357 categories are known as below-normal, normal and above-
 1358 normal whose end points provide streamflow values
 1359 corresponding to the particular category. Thus, for a total of
 1360 D categories, the end points based on climatological
 1361 observations for d th category could be written as Q_d, Q_{d+1}
 1362 (For $d = 1, Q_1 = 0$; $d = D; Q_{D+1} = Q_{\max}$). Given streamflow
 1363 forecasts at time “ t ” from m th model with $i = 1, 2, \dots, N$
 1364 ensembles, $Q_{i,t}^m$, then the forecast probabilities for the d th
 1365 category could be expressed as $FP_{d,t}^m(Q) = n_{d,t}^m/N$ by
 1366 computing the number of ensembles between $Q_d \leq Q_{i,t}^m \leq$
 1367 Q_{d+1} . To compute RPSS, the first step is to compute rank
 1368

1369 probability score (RPS). Given D categories and $FP_{d,t}^m(Q)$
 1370 for a forecast, we can express the RPS, which is otherwise
 1371 known as $L = 2$ norm, for a particular year “ t ” from m th
 1372 model as

$$RPS_t^m = \sum_{d=1}^D [CF_{d,t}^m - CO_d]^2 \quad (A1)$$

1374 where $CF_{d,t}^m = \sum_{q=1}^d FP_{d,t}^m$ is the cumulative probabilities of
 1375 forecasts up to category d and CO_d is the cumulative
 1376 probability of the observed event up to category d . Similar
 1377 measures for evaluating the entire conditional distribution of
 1378 forecasts have also been proposed [Muller et al., 2005]. For
 1379 instance, the $L = 1$ norm (equation (A2)) could be written as
 1380 the absolute sum of deviation in cumulative probabilities of
 1381 forecasts and the observed event.

$$RPS - L1_t^m = \sum_{d=1}^D |CF_{d,t}^m - CO_d| \quad (A2)$$

1383 Thus if Q_t , the observed streamflow falls in the d th category,
 1384 $CO_q = 0$ for $1 \leq q \leq d-1$ and $CO_q = 1$ for $d \leq q \leq D$.
 1385 Given RPS, we can compute RPSS in relation to a reference
 1386 forecast, which is usually climatological forecasts having
 1387 equal probability of occurrence under each category $FP_{d,t}^{c \lim}$
 1388 (Q) = $1/D$.

$$RPSS_t^m = 1 - \frac{RPS_t^m}{RPS_t^{c \lim}} \quad (A3)$$

1390 Low RPS indicates high skill and vice versa. The relative
 1391 measure RPSS, if it is positive, then the forecast skill
 1392 exceeds that of the climatological probabilities. RPSS could
 1393 give an overly pessimistic view of the performance of the
 1394 forecasts and it is a tough metric for evaluating probabilistic
 1395 forecasts [Goddard et al., 2001]. For a detailed example on
 1396 how to compute RPS and RPSS for given forecast, see
 1397 Goddard et al. [2003]. In this study, we have computed RPS
 1398 and RPSS for each year and both regression and resampling
 1399 ensembles by assuming $D = 3$.

1400 Appendix B: Hypothesis Tests for Evaluating 1401 Streamflow Forecasts

1402 [56] This appendix briefly summarizes the hypothesis
 1403 tests employed for finding whether the skill of the stream-
 1404 flow forecasts, indicated by average RPS, from any of the
 1405 two different models given in Table 1 are statistically
 1406 significant. For testing the null hypothesis on the probabi-
 1407 listic forecasts skill measure, average RPS, we employ the
 1408 nonparametric hypothesis tests based on resampling ap-
 1409 proach. For a detailed discussion of the methodology
 1410 employed, see Hamill [1999]. The null hypothesis for
 1411 testing the average RPS could be written as:

$$H_0 : \overline{RPS}^A - \overline{RPS}^B = 0 \quad (B1)$$

$$H_A : \overline{RPS}^A - \overline{RPS}^B \neq 0 \quad (B2)$$

Given that RPS_t^A and RPS_t^B are calculated each year for the
 candidate models A and B using equation (A1), we can get
 the test statistic: $\overline{RPS}^A - \overline{RPS}^B$ where the average RPS for
 any model “ m ” could be calculated using (B6):

$$\overline{RPS}^m = n^{-1} \sum_{t=1}^n RPS_t^m \quad (B3)$$

Using RPS_t^A and RPS_t^B , we now generate the null
 distribution. To develop this, we generate an indicator
 variable I_t , taking on a value of 1 (Model A) or 2 (Model B)
 with equal probability. The indicator is used to select either
 Model A or B every year using which the resampled statistic
 is developed as: $\overline{RPS}^{1,*} - \overline{RPS}^{2,*}$ where

$$\overline{RPS}^{1,*} = n^{-1} \sum_{t=1}^n RPS_t^{I_t} \quad (B4)$$

$$\overline{RPS}^{2,*} = n^{-1} \sum_{t=1}^n RPS_t^{(3-I_t)} \quad (B5)$$

Basically, equations (B4) and (B5) ensure that the
 resampled statistic is computed using both Models A and
 B with equal probability. By obtaining 10000 such
 resampled statistic estimates, we construct the null distribu-
 tion and use that to calculate the percentile value of the
 estimated test statistic, $\overline{RPS}^A - \overline{RPS}^B$, for the two candidate
 models A and B. On the basis of the reported percentile
 value, one could get the p-value to accept or reject the null
 hypothesis for the chosen significance level, α .

[57] **Acknowledgments.** This study was supported by the North
 Carolina Water Resources Research Institute. We also would like to thank
 the three anonymous reviewers and the associate editor whose valuable
 comments led to significant improvements in our manuscript. Thanks to
 Tom Fransen, NC DENR, and Terry Brown, USACE, for sharing Falls
 Lake inflow data to this study.

References

- Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic
 Bayesian multimodel combination framework: Confronting input, param-
 eter and model structural uncertainty in hydrologic prediction, *Water
 Resour. Res.*, 43, W01403, doi:10.1029/2005WR004745.
 Ajami, N. K., Q. Duan, X. Gao, and S. Sorooshian (2006), Multimodel
 combination techniques for hydrological forecasting: Application to dis-
 tributed model intercomparison project results, *J. Hydrometeorol.*, 7(4),
 755–768.
 Anderson, J. L. (1996), A method for producing and evaluating probabil-
 istic forecasts from ensemble model integrations, *J. Clim.*, 9, 1518–
 1530.
 Barnston, A. G., S. J. Mason, L. Goddard, D. G. DeWitt, and S. E. Zebiak
 (2003), Multimodel combination in seasonal climate forecasting at IRI,
Bull. Am. Meteorol. Soc., 84(12), 1783–1796.
 Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved
 calibration of hydrologic models: Combining the strengths of manual and
 automatic methods, *Water Resour. Res.*, 36(12), 3663–3674.
 Brankovic, C., and T. N. Palmer (2000), Seasonal skill and predictability of
 ECMWF PROVOST ensembles, *Q. J. R. Meteorol. Soc.*, 126, 2035–
 2067.

- Candille, G., and O. Talagrand (2005), Evaluation of probabilistic prediction systems for a scalar variable, *Q. J. R. Meteorol. Soc.*, *131*(609), 2131–2150.
- Carpenter, T. M., and K. P. Georgakakos (2001), Assessment of Folsom lake response to historical and potential future climate scenarios. 1: Forecasting, *J. Hydrol.*, *249*, 148–175.
- Cayan, D. R., K. T. Redmond, and L. G. Riddle (1999), ENSO and hydrologic extremes in the western United States, *J. Clim.*, *12*(9), 2881–2893.
- Craven, P., and G. Whaba (1979), Optimal smoothing of noisy data with spline functions, *Numer. Math.*, *31*, 377–403.
- Dettinger, M. D., and H. F. Diaz (2000), Global characteristics of stream flow seasonality and variability, *J. Hydrometeorol.*, *1*(4), 289–310.
- Doblas-Reyes, F. J., M. Deque, and J. P. Piedelievre (2000), Multimodel spread and probabilistic seasonal forecasts in PROVOST, *Q. J. R. Meteorol. Soc.*, *126*(567), 2069–2087.
- Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer (2005), The rationale behind the success of multimodel ensembles in seasonal forecasting. II: Calibration and combination, *Tellus, Ser. A: Dyn. Meteorol. Oceanogr.*, *57*(3), 234–252.
- Gangopadhyay, S., M. Clark, and B. Rajagopalan (2005), Statistical downscaling using K-nearest neighbors, *Water Resour. Res.*, *41*, W02024, doi:10.1029/2004WR003444.
- Georgakakos, K. P. (2003), Probabilistic climate-model diagnostics for hydrologic and water resources impact studies, *J. Hydrometeorol.*, *4*(1), 92–105.
- Giannini, A., R. Saravanan, and P. Chang (2004), The preconditioning role of tropical Atlantic variability in the development of the ENSO teleconnection: Implications for the prediction of Nordeste rainfall, *Clim. Dyn.*, *22*, 839–855.
- Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane (2001), Current approaches to seasonal-to-interannual climate predictions, *Int. J. Climatol.*, 1111–1152.
- Goddard, L., A. G. Barnston, and S. J. Mason (2003), Evaluation of the IRI's "net assessment" seasonal climate forecasts 1997–2001, *Bull. Am. Meteorol. Soc.*, *84*, 1761–1781.
- Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona (2005), A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts, *Water Resour. Res.*, *41*, W10410, doi:10.1029/2004WR003467.
- Guetter, A. K., and K. P. Georgakakos (1996), Are the El Nino and La Nina predictors of the Iowa River seasonal flow?, *J. Appl. Meteorol.*, *35*(5), 690–705.
- Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer (2005), The rationale behind the success of multimodel ensembles in seasonal forecasting. I: Basic concept, *Tellus, Ser. A: Dyn. Meteorol. Oceanogr.*, *57*(3), 219–233.
- Hamill, T. M. (1999), Hypothesis tests for evaluating numerical precipitation forecasts, *Weather Forecast.*, *14*(2), 155–167.
- Hamlet, A. F., and D. P. Lettenmaier (1999), Columbia River streamflow forecasting based on ENSO and PDO climate signals, *J. Water Resour. Plann. Manage.*, *125*, 333–341.
- Hansen, J. W., A. W. Hodges, and J. W. Jones (1998), ENSO influences on agriculture in the southeastern United States, *J. Clim.*, *11*, 404–411.
- Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*, 382–401.
- Kiehl, J. T., J. J. Hack, G. B. Bonan, B. A. Boville, D. L. Williamson, and P. J. Rasch (1998), The National Center for Atmospheric Research Community Climate Model: CCM3, *J. Clim.*, *11*(6), 1131–1149.
- Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran (1999), Improved weather and seasonal climate forecasts from multimodel superensemble, *Science*, *285*, 1548–1550.
- Landman, W. A., and L. Goddard (2002), Statistical recalibration of GCM forecasts over southern Africa using model output statistics, *J. Clim.*, *15*, 2038–2055.
- Lecce, S. A. (2000), Spatial variations in the timing of annual floods in the southeastern United States, *J. Hydrol.*, *235*(3–4), 151–169.
- Leung, L. R., A. F. Hamlet, D. P. Lettenmaier, and A. Kumar (1999), Simulations of the ENSO hydroclimate signals in the Pacific Northwest Columbia River basin, *Bull. Am. Meteorol. Soc.*, *80*, 2313–2329.
- Marshall, L., A. Sharma, and D. Nott (2006), Modeling the catchment via mixtures: Issues of model specification and validation, *Water Resour. Res.*, *42*, W11409, doi:10.1029/2005WR004613.
- Mason, S. J., and G. M. Mimmack (2002), Comparison of some statistical methods of probabilistic forecasting of ENSO, *J. Clim.*, *15*(1), 8–29.
- Moura, A. D., and J. Shukla (1981), On the dynamics of droughts in north-east Brazil—observations, theory and numerical experiments with a general-circulation model, *J. Atmos. Sci.*, *38*(12), 2653–2675.
- Muller, W. A., C. Appenzeller, F. J. Doblas-Reyes, and M. A. Liniger (2005), A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes, *J. Clim.*, *18*, 1513–1523.
- Nobre, P., A. D. Moura, and L. Q. Sun (2001), Dynamical downscaling of seasonal climate prediction over nordeste Brazil with ECHAM3 and NCEP's regional spectral models at IRI, *Bull. Am. Meteorol. Soc.*, *82*, 2787–2796.
- Murphy, A. H. (1970), Ranked probability score and probability score—a comparison, *Mon. Weather Rev.*, *98*, 917.
- Palmer, T. N., C. Brankovic, and D. S. Richardson (2000), A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations, *Q. J. R. Meteorol. Soc.*, *126*, 2013–2033.
- Piechota, T. C., and J. A. Dracup (1996), Drought and regional hydrologic variation in the United States: Associations with the El Nino Southern Oscillation, *Water Resour. Res.*, *32*(5), 1359–1373.
- Quan, X., M. Hoerling, J. Whitaker, G. Bates, and T. Xu (2006), Diagnosing sources of US seasonal forecast skill, *J. Clim.*, *19*, 3279–3293.
- Rajagopalan, B., U. Lall, and S. E. Zebiak (2002), Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles, *Mon. Weather Rev.*, *130*, 1792–1811.
- Rasmusson, E. M., and T. H. Carpenter (1982), Variations in tropical sea-surface temperature and surface wind fields associated with the Southern Oscillation El-Nino, *Mon. Weather Rev.*, *110*(5), 354–384.
- Regonda, S. K., B. Rajagopalan, M. Clark, and E. Zagona (2006), A multimodel ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin, *Water Resour. Res.*, *42*, W09404, doi:10.1029/2005WR004653.
- Rhime, J. R., D. S. Niyogi, and S. Raman (2000), Mesoclimatic analysis of severe weather and ENSO interactions in North Carolina, *Geophys. Res. Lett.*, *27*(15), 2269–2272.
- Roads, J., S. Chen, S. Cocke, L. Druyan, M. Fulakeza, T. LaRow, P. Lonergan, J. H. Qian, and S. Zebiak (2003), International Research Institute/Applied Research Centers (IRI/ARCs) regional model intercomparison over South America, *J. Geophys. Res.*, *108*(D14), 4425, doi:10.1029/2002JD003201.
- Robertson, A. W., S. Kirshner, and P. Smyth (2004a), Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model, *J. Clim.*, *17*, 4407–4424.
- Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard (2004b), Improved combination of multiple atmospheric GCM ensembles for seasonal prediction, *Mon. Weather Rev.*, *132*, 2732–2744.
- Ropelewski, C. F., and M. S. Halpert (1987), Global and regional scale precipitation patterns associated with the El-Nino Southern Oscillation, *Mon. Weather Rev.*, *115*(8), 1606–1626.
- Roswintarti, O., D. S. Niyogi, and S. Raman (1998), Teleconnections between tropical Pacific sea surface temperature anomalies and North Carolina precipitation anomalies during El Nino events, *Geophys. Res. Lett.*, *25*(22), 4201–4204.
- Roulston, M. S., and L. A. Smith (2003), Evaluating probabilistic forecasts using information theory, *Mon. Weather Rev.*, *130*, 1653–1660.
- Sankarasubramanian, A., and U. Lall (2003), Flood quantiles in a changing climate: Seasonal forecasts and causal relations, *Water Resour. Res.*, *39*(5), 1134, doi:10.1029/2002WR001593.
- Sankarasubramanian, A., U. Lall, and S. Espuneva (2007), Role of retrospective forecasts of GCM forced with persisted SST anomalies in operational streamflow forecasts development, *J. Hydrometeorol.*, in press.
- Schmidt, N., E. K. Lipp, J. B. Rose, and M. E. Luther (2001), ENSO influences on seasonal rainfall and river discharge in Florida, *J. Clim.*, *14*, 615–628.
- Seo, D. J., V. Koren, and N. Cajina (2003), Real-time variational assimilation of hydrologic and hydrometeorological data into operational hydrologic forecasting, *J. Hydrometeorol.*, *4*, 627–641.
- Shukla, J., et al. (2000), Dynamical seasonal prediction, *Bull. Am. Meteorol. Soc.*, *81*, 2593–2606.
- Souza, F. A., and U. Lall (2003), Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm, *Water Resour. Res.*, *39*(11), 1307, doi:10.1029/2002WR001373.
- Stephenson, D. B., C. A. S. Coelho, F. J. Doblas-Reyes, and M. Balmaseda (2005), Forecast assimilation: A unified framework for the combination

- 1615 of multimodel weather and climate predictions, *Tellus, Ser. A: Dyn.*
 1616 *Meteorol. Oceanogr.*, 57, 253–264.
 1617 Trenberth, K. E., and C. J. Guillemot (1996), Physical processes involved in
 1618 the 1988 drought and 1993 floods in North America, *J. Clim.*, 9(6),
 1619 1288–1298.
 1620 Weaver, C. J. (2005), The drought of 1998–2002 in North Carolina—
 1621 precipitation and hydrologic conditions, in *USGS Scientific Investiga-*
 1622 *tions Report*.
 1623 Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*,
 1624 Academic Press, New York.
 1625 Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier (2002), Long-
 1626 range experimental hydrologic forecasting for the eastern United States,
 1627 *J. Geophys. Res.*, 107(D20), 4429, doi:10.1029/2001JD000659.
 1628 Wood, A. W., A. Kumar, and D. P. Lettenmaier (2005), A retrospective
 1629 assessment of National Centers for Environmental Prediction climate
 model-based ensemble hydrologic forecasting in the western United States, *J. Geophys. Res.*, 110, D04105, doi:10.1029/2004JD004508.
 Zorn, M. R., and P. R. Waylen (1997), Seasonal response of mean monthly
 streamflow to El Nino/Southern Oscillation in north central Florida, *Prof.*
Geogr., 49(1), 51–62.
-
- N. Devineni and A. Sankarasubramanian, Department of Civil,
 Construction and Environmental Engineering, North Carolina State
 University, 2501 Stinson Drive, Box 7908, Raleigh, NC 27695-7908,
 USA. (sankar_arumugam@ncsu.edu)
 S. Ghosh, Department of Statistics, North Carolina State University,
 Raleigh, NC 27695-8203, USA.

American Geophysical Union
Author Query Form

Journal: **Water Resources Research**
Article Name: **Sankarasubramanian(2006WR005855)**

Please answer all author queries.

1. There must be at least two heads per level. Please provide another head under section 3 with the same level as section 3.1 or section 3.1 head will be deleted.
2. Section 3.2 was cited in the text but there was no such section provided. Please check.
3. "Equation (B6)" was cited in appendix but no such equation is provided. Please check.
4. Citation of "*Roulston and Smith, 2000*" was changed to ""***Roulston and Smith, 2003***" to match the reference list. Please check if appropriate.
5. According to AGU reference format, when two or more works published in the same year and with same first author, they are differentiated by a lowercase letter (a, b, c, etc.) immediately following the year. "*Robertson et al., 2004*" and "*Robertson et al., 2004*" was changed to "*Robertson et al., 2004a*" and "*Robertson et al., 2004b*" respectively in the reference list. In addition, citations of "*Robertson et al., 2004*" in the text was changed to "*Robertson et al., 2004a, 2004b*", please advise if this is appropriate.
6. The following references are cited in the text but are not present in the reference list. Please provide the appropriate information for the reference or delete the citation from the text.
 - "*Georgakakos et al., 2004*"
 - "*Dettinger et al., 2000b*"
 - "*Robertson et al., 2003*"
 - "*Mason [2001]*"
7. "*K. Golembesky et al., manuscript in preparation, 2008*" was deleted from the reference list and was cited fully in the text. Please provide an update for the publication status of this reference.
8. Please provide an update for the publication status of the reference "*Sankarasubramanian et al. (2007), in press*".
9. Please provide complete bibliographic reference in "*Weaver (2005)*".
10. The following references were present in the list of references but were not cited in your paper. Kindly provide a citation for the said reference or remove it from the reference list.
 - "*Dettinger and Diaz (2000)*"
 - "*Doblas-Reyes et al. (2005)*"
 - "*Nobre et al. (2001)*"
11. Please provide complete mailing address (building/street address and postal code) of S. Ghosh.