A new method to produce categorical streamflow forecasts

Satish Kumar Regonda,[1,2] Balaji Rajagopalan,[1,2] and Martyn Clark[3]

Received 20 February 2006; revised 14 June 2006; accepted 30 June 2006; published 12 September 2006.

[1] Categorical forecasts of streamflow are important for effective water resources management. Typically, these are obtained by generating ensemble forecasts of streamflow and counting the proportion of ensembles in the desired category. Here we develop a simple and direct method to produce categorical streamflow forecasts at multiple sites. The method involves predicting the probability of the leading mode (or principal component) of the basin streamflows above a given threshold and subsequently translating the predicted probabilities to all the sites in the basin. The categorical probabilistic forecasts are obtained via logistic regression using a set of large-scale climate predictors. Application to categorical forecasts of the spring (April–June) streamflows at six locations in the Gunnison River Basin exhibited significant long-lead forecast skill.

Citation: Regonda, S. K., B. Rajagopalan, and M. Clark (2006), A new method to produce categorical streamflow forecasts, *Water Resour. Res.*, *42*, W09501, doi:10.1029/2006WR004984.

1. Introduction

[2] Many rivers in the western United States are heavily controlled by water storage structures to provide reliable water supply for a variety of societal needs. To effectively manage the scarce water resources, water managers require the following two elements: (1) skillful forecasts of streamflows at several lead times and (2) a decision support system that incorporates forecasted information and evaluates the impact of different management strategies. Here we analyze the first element, streamflow forecasts, which are provided in two forms according to the water manger's need: categorical (probabilities of wet or dry conditions) and volume (monthly or seasonal streamflows) flow forecasts. A categorical streamflow forecast provides the probability of occurrence of a particular event e.g., the chance of having higher streamflow. A volume flow forecast provides the amount of streamflows e.g., on daily, monthly, or seasonal timescales.

[3] Two approaches are used to forecast streamflow: physical and statistical models. In physical models, a hydrologic model is run using station data up to the start of the forecast to estimate the basin conditions (e.g., snowpack, soil moisture), and is then run into the future, with an "ensemble" of weather/climate forecasts, to produce ensemble forecasts of streamflow [e.g., *Day*, 1985; *Clark and Hay*, 2004]. The ensemble of weather/climate forecasts comprise a finite number of individual realizations of precipitation and temperature over the next several weeks or several seasons, which, when used as input to the hydrologic model, produce the same number of future realizations of streamflow. Categorical forecasts can be produced by counting the number of individual ensemble members that are above a predefined threshold. Statistical models on the other hand use empirical relations to forecast streamflow [*Garen*, 1992]. For example, snow water equivalent on 1 April may be used in a regression model to predict streamflow averaged over the months April through September. Uncertainty in the statistical models can be estimated easily (e.g., using the standard deviation of the regression residuals), and ensemble forecasts can be produced by sampling from the distribution of regression residuals [*Grantz et al.*, 2005; *Regonda et al.*, 2006]. Now, as with the physical models, categorical probabilities from the regression models can be computed from ensemble forecasts. Also, there are statistical methods (e.g., discriminant analysis) that directly estimate categorical forecasts for specific thresholds but require strong distributional assumptions, which when not satisfied, often the case with real data sets, require indirect resampling of errors.

[4] *Piechota et al.* [1998] developed a categorical forecast framework based on linear discriminant analysis (LDA). This approach has two main steps. First, probability density functions (PDFs) of a given predictor (e.g., the Southern Oscillation index) were estimated for three subsets of streamflow data (below normal, normal, and above normal), and the categorical forecast was estimated using Bayes rule

$$\Pr(Q_i|X) = \frac{p_i f_i(x)}{\sum_{i=1}^{k=3} p_i f_i(x)} \quad (1)$$

Here $\Pr(Q_i|X)$ is the probability of streamflow (Q) in the ith category, given the predictor X; p_i is the prior probability of the ith category (i.e., the fraction of observations in the ith category); and $f_i(x)$ is the probability of the predictor variable computed using data from only the ith category. *Piechota et al.* [1998] estimated the PDFs in equation (1) using nonparametric kernel density estimation [*Lall*, 1995].

[1]Department of Civil, Environmental and Architectural Engineering, University of Colorado, Boulder, Colorado, USA.
[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA.
[3]National Institute for Water and Atmospheric Research, Christchurch, New Zealand.

0043-1397/06/2006WR004984$09.00

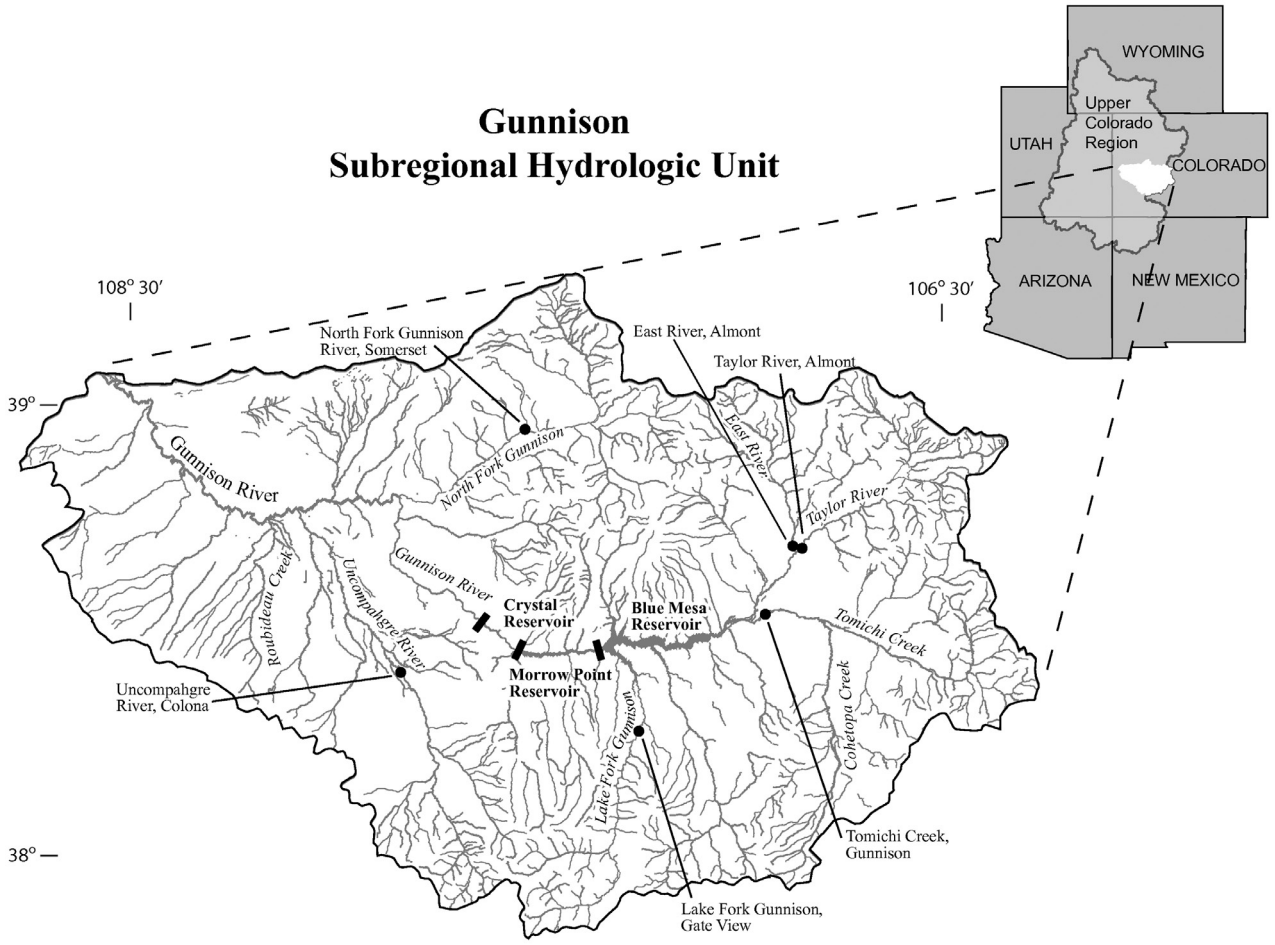

Figure 1. Map of the Gunnison River Basin and six key streamflow locations (shown as circles). Map was provided by James Pasquotto, University of Colorado, Boulder.

This first step is hence very similar to contingency table analysis (e.g., estimating the joint probability of X and Q for different categories of X and Q), except the probability of X, $f_i(x)$, is allowed to vary within each of the k categories of Q. The second step in the *Piechota et al.* [1998] method is extension to multiple predictor variables. Multiple variables (X_1, ..., X_n) were included by assigning weights to each potential predictor variable, hence

$$\Pr(Q_i|X_1,\ldots,X_n) = \sum_{j=1}^{n} w_j \Pr\left(Q_i|X_j\right), \quad 0 \le w_j \le 1, \quad \sum_{j=1}^{n} w_j = 1 \tag{2}$$

in which w_i are the weights assigned to each predictor variable. The weights were determined by minimizing the error in historical forecasts.

[5] The *Piechota et al.* [1998] approach is attractive in that it provides a direct method to produce categorical streamflow forecasts. However, the method is somewhat cumbersome in that it requires both kernel density methods for estimating the PDF for each forecast category for use in equation (1), as well as optimization methods for identifying the weights for use in equation (2). Together equations (1) and (2) are similar to logistic regression

$$\Pr(Q_i|X_1,\ldots,X_n) = 1 - \frac{1}{1+\exp\left(\beta_o + \sum_{j=1}^{n}\beta_j X_j\right)} \tag{3}$$

where (β_0, ..., β_η) are regression coefficients obtained by a maximum likelihood procedure. Logistic regression offers potential improvements over LDA as it fits a function throughout the data and hence does not rely on the likely noisy ratio of probabilities in the LDA approach. Further, logistic regression is included in many statistical software packages.

[6] Apart from the *Piechota et al.* [1998] studies, almost all of the current methods used to generate categorical streamflow forecasts are somewhat indirect. The purpose of this study is twofold. First, we develop a simple and direct method to produce categorical streamflow forecasts at multiple locations. Second, we compare the direct forecasting methods against indirect ensemble-based methods. We intend our new method to be complementary to the multimodel ensemble forecasting technique developed by *Regonda et al.* [2006] and also an attractive alternative to the methods of *Piechota et al.* [1998].

[7] The proposed framework and its application to the six key streamflow locations (Figure 1) in the Gunnison River Basin (GRB) are described in the following sections. These locations are along the main reservoir system in the basin, and also points of release of water that satisfy various basin needs (e.g., agriculture, municipal, and transbasin diversions).

2. Methodology

[8] The proposed integrated framework has the following three components: (1) Use principal component analysis (PCA) to identify the leading modes of spatial variability in regional streamflow; these modes are also known as the leading principal components (PCs). (2) Assess relationships between the leading streamflow PCs and global ocean and atmospheric variables to identify potential predictors. (3) Use logistic regression to issue categorical streamflow forecasts. Results from *Regonda et al.* [2006] are used for components 1 and 2, while the logistic regression framework is developed in this study. Categorical forecasts are evaluated using the Brier skill score.

[9] This method was applied to six streamflow locations in the GRB and, categorical spring streamflow forecasts were issued on the first of each month from December (i.e., 4 month lead time) through 1 April in a cross-validated mode, in which all data from a given year is dropped from the forecasting framework while predicting in that year. We describe below the methodology along with the data and the application.

2.1. Data

[10] The daily discharges at the six streamflow locations for the period 1949–2004 were obtained from the U.S. Geological Survey (USGS) (http://water.usgs.gov/). As expected, the streamflows exhibited a snow driven annual hydrograph (not shown), which has spring snowmelt (April–July) resulting in the major contribution (greater than 70%) of annual flows. Hence we considered spring seasonal flows (averaged flow during April–July) in this analysis, and developed spring streamflow time series for each of the six locations.

[11] Snow water equivalent (SWE) data were obtained from Natural Resource Conservation Service (NRCS). The SWE measurements are taken around the beginning of each month, and records of February, March and April are considered, respectively, at 10, 14, and 14 locations in the basin.

[12] Monthly global ocean and atmospheric variables such as sea surface temperatures, geopotential height, zonal wind, and meridional wind, and Palmer drought severity index (an index of soil moisture) were obtained from the Climate Diagnostic Center Web site (http://www.cdc.noaa.gov).

2.2. Forecast Method

2.2.1. Step 1: Principal Component Analysis

[13] PCA decomposes the space-time multivariate data set (time series of streamflows at six locations of the GRB) into orthogonal space (eigenvectors) and time (principal components) patterns using eigendecomposition [see e.g., *Von Storch and Zwiers*, 1999]. The space-time patterns are ordered according to the percentage of data variance explained (i.e., the first space-time pattern explains most of the data variance), and the leading PCs, corresponding to the space-time patterns that explain most of the data variance, are selected.

[14] PCA on the spring streamflows at the six locations of the GRB resulted in the first PC explaining 87% of the variance, and the remaining five explained 13% of the variance. Clearly, the first PC is the leading mode and furthermore, it had uniform eigenloadings (not shown) and high correlation (>= 0.8) with all the six streamflows. This indicates that the leading PC is a robust indicator of basin-wide streamflow variability.

2.2.2. Step 2: Predictor Selection

[15] Since the leading PC captures most of the data variance of the spring streamflows, we searched for predictors by finding the correlation of the leading PC with large-scale ocean and atmospheric variables across several regions of the world, from preceding seasons. The regions with strong correlations are identified, and predictors are created by averaging the value of the variables from these regions. This is done for each lead time. For example, 1 January forecast considers the large scale ocean-atmospheric variables of November–December season, and similarly 1 April forecast considers the November–March season, etc. In addition, the first PC of the monthly SWE and the average Palmer drought severity index (PDSI) from the GRB region of the antecedent fall were also added to the suite of predictors. The variability of SWE in the basin is quite homogeneous, as can be seen by the fact that the first PC of SWE captures over 70% of the variance in all the months (i.e., 73%, 70%, and 70% variances are explained by 1 February, 1 March, and 1 April SWEs, respectively) furthermore, the first PC of SWE is highly correlated (correlation coefficient >0.70) with the first PC of the spring streamflows (i.e., 0.72, 0.76, and 0.84 respectively for 1 February, 1 March, and 1 April SWEs); hence the first PC of SWE is a good predictor. The PDSI was included because drier conditions in the antecedent fall lead to an increased infiltration in the following spring when the snow starts to melt, and thus reducing the spring streamflows. The details on the diagnostics and selection of the predictors are given by *Regonda et al.* [2006, Table 2].

[16] From the suite of predictors identified from the climate diagnostics, *Regonda et al.* [2006] developed several regression models for the first streamflow PC using a nonparametric local polynomial framework, for each lead time. Each of the regression models consists of a different subset of predictors, and is used to issue ensemble forecasts of the first PC. Since many models have similar skill, it is difficult to identify a single ''best'' model. Consequently, *Regonda et al.* [2006] combine predictions from multiple regression models to obtain multimodel ensemble forecast of the first streamflow PC, and subsequently back transforms the multimodel ensemble forecast to all the six locations in the basin, simultaneously. Categorical forecasts are then issued from the multimodel ensemble forecast.

[17] In this study our goal is to develop and demonstrate the utility of a logistic regression based framework for categorical forecasting. For simplicity at each lead time we used the best predictor set identified by *Regonda et al.* [2006] (i.e., predictors of the model with the lowest error) to produce ensemble streamflow forecasts and the

Table 1. Best Predictor Set for the Logistic Regression at Different Lead Times[a]

Forecast Date	Number of Predictors	Predictor 1	Predictor 2
1 Dec	1	SST^{Dec1}	NA
1 Jan	2	ZW^{Jan1}	PDSI
1 Feb	1	SWE^{Feb1}	NA
1 Mar	2	SWE^{Mar1}	SST^{Mar1}
1 Apr	1	SWE^{Apr1}	NA

[a]SST^{Dec1} is the averaged sea surface temperature anomalies difference between (14.3°–41.0°N, 71.2°–39.4°W) and (42.9°–52.4°N, 176.2°E–150.0°W) for 1 December; SST^{Mar1} is between (21.9°–42.9°N, 60.0°–30.0°W) and (23.8°–10.5°S, 110.6°–78.8°W) for 1 March. ZW^{Jan1} is the averaged zonal wind difference between (50.0°–60.0°N, 125.0°–100.0°W) and (60.0°–65.0°N, 155.0°–142.5° W) for 1 January. PDSI is August–October Palmer drought severity index of preceding year averaged over climate division 2 (which includes the GRB) of Colorado; SWE^{Feb1} corresponds to the first PC of snow water equivalent in the basin on 1 February (and similarly for SWE^{Mar1} and SWE^{Apr1}).

corresponding categorical probabilities. That is, the multimodel forecasts are not used. The same combination of predictors from the single best local polynomial model is then used as explanatory variables in the logistic regression model to estimate categorical forecast probabilities directly. These predictor sets are described in Table 1 for different lead times. A more formal approach would be to identify the best subset of predictors using the logistic regression.

2.2.3. Step 3: Categorical Forecast Framework—Logistic Regression

[18] As described earlier, the goal of this study is to develop a framework that issues categorical forecasts of streamflows. To this end, probabilities corresponding to events of different threshold values need to be estimated. This consists of the following steps.

[19] 1. A threshold value, say the 20th percentile, is chosen for the first PC.

[20] 2. The first PC time series is reduced to a binary (1 if the PC value exceeds the threshold, 0 otherwise) series.

[21] 3. Logistic regression is employed [*Hosmer and Lemeshow*, 1989] to obtain the threshold probabilities. This method was outlined in the Introduction but is repeated here for completeness:

$$P_{logit} = \Pr(Q_i|X_1,\ldots,X_n) = 1 - \frac{1}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\ldots\ldots+\beta_nX_n)} \quad (4)$$

where P_{logit} is the probability of the event of interest (e.g., PC value exceeding the 20th percentile) and (β_1, β_2, … β_n) are regression coefficients for each of the predictor variables (X_1, X_2, …, X_n). Correspondingly, the probability of nonexceedance of the threshold value is ($1 - P_{logit}$). The regression coefficients are obtained via maximum likelihood estimation procedure (for details see, *Hosmer and Lemeshow* [1989]). We used the library "glm" in the free software R (http://www.r-project.org).

[22] 4. Typically, PCA tends to change the distribution of the original variables to normal distribution. Thus, if the distribution of the original variables is nonnormal then this transformation can lead to incorrect categorical probabilities. We performed a Kolmogorov-Smirnov (KS) test on spring season streamflows at each location to test the normality of the distribution. In addition, we also tested the distributional difference between the first PC and each of the streamflows. All the spring flows, except Taylor River, were found to be normally distributed also, the distribution of all the seasonal streamflows were found to be as that of the first PC, at 95% confidence level. These tests suggest that the PC analysis retained the distributional properties of the original data consequently, the rank probability as well. Furthermore, the first PC was found to be highly positively correlated with the streamflows in the basin and explains 87% of the variance. These diagnostics indicate that we can translate the probabilistic forecast of the PC directly to the streamflows at all the six locations. Thus the P_{logit} value obtained from equation (4) (step 3) for a given threshold (say the 20th percentile of the PC value) is interpreted as the probabilistic forecast of same threshold (i.e., 20th percentile) exceedance of the streamflows.

[23] 5. Steps 1–4 are repeated for several thresholds.

[24] This framework is applied in a cross-validated mode to obtain the probabilistic streamflow forecasts for all the years at the six locations for different lead times.

[25] An alternative approach would be to fit logistic regression models to each individual streamflow gauge in the GRB. The PCA step of *Regonda et al.* [2006] has important advantages, as it preserves the spatial correlation structure over the basin when forecasts of the PC are disaggregated to individual stream gauges. In the logistic regression model developed in this paper, the probability for the PC is distributed uniformly over the basin, and we hence lose the capabilities for spatial disaggregation in the original *Regonda et al.* [2006] method. Nevertheless, use of the PC has other advantages, as it filters the data and reduces noise in the predictand variable. Moreover, use of the PC for logistic regression provides scope to compare the logistic regression approach with the original *Regonda et al.* [2006] method.

2.3. Forecast Evaluation

[26] The performance of the categorical probabilistic forecasts issued from the logistic regression framework is evaluated using the Brier skill score (BSS). This is a widely used measure to verify categorical probabilistic forecasts [*Wilks*, 1995]. The BSS is computed for each threshold and is defined as [*Wilks*, 1995]:

$$BSS = 1 - \frac{BS_{forecast}}{BS_{clim}} \quad (5)$$

where $BS_{forecast}$ and BS_{clim} are the Brier score (BS) corresponding to the logistic regression (or best model ensemble) forecast and climatology, respectively. The BS is the mean squared difference of forecasted probabilities and observations (equation (6)) and is defined as

$$BS = \frac{\sum_{i=1}^{N}(p_i - o_i)^2}{N} \quad (6)$$

where p_i refers to the forecast probabilities for a given threshold value (i.e., either estimated from equation (4) or

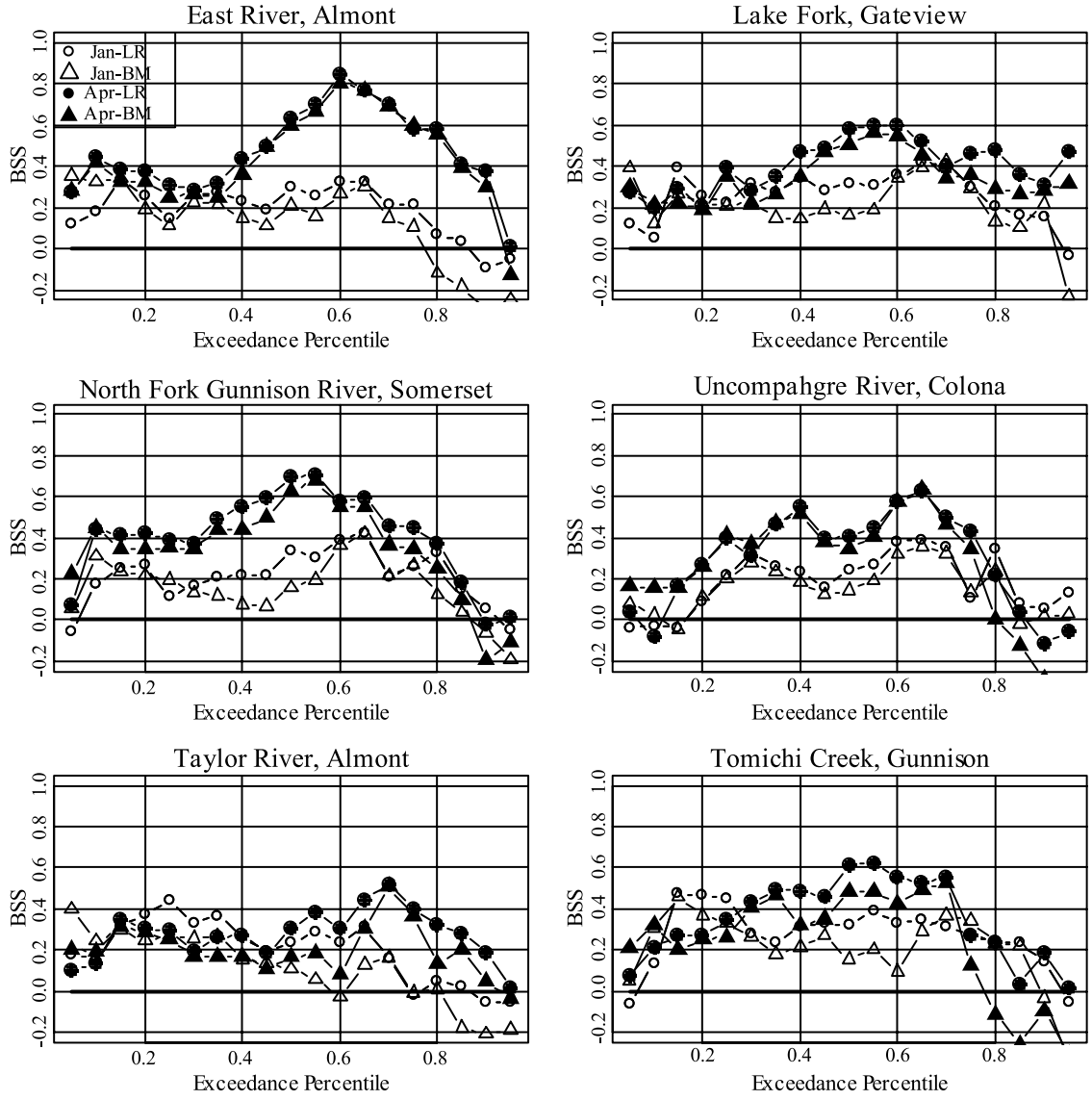

Figure 2. BSS of forecasts from the logistic regression (LR) framework (circles) and the "best model" (BM) ensemble of *Regonda et al.* [2006] (triangles), issued on 1 January (open circles and open triangles) and 1 April (solid circles and solid triangles) at the six locations in the GRB. Exceedance percentiles and BSS are plotted on x and y axes, respectively.

from the best model ensemble); o_i refers to the outcomes; o_i is 1 if the observed flow exceeds the threshold volumes and 0 otherwise; N is number of forecasts, in this case the number of years. Replacing p_i with climatological probability in equation (6) results in the BS_{clim}; for example, the climatological probability of 20th percentile exceedance will be 0.8. The BSS values range from negative infinity to 1; negative values of BSS indicate forecast performance worse than climatology, and positive values indicate forecast performance better than climatology. A BSS of "0" implies that the forecast accuracy is the same as that for climatological forecasts. In contrast, a BSS of 1 can only occur for perfect forecasts.

3. Results

[27] We estimated the cross-validated BSS for the categorical streamflow forecasts issued from the logistic regression framework developed in this paper and also from the best model ensemble of *Regonda et al.* [2006], at all the six locations and at different lead times. Figure 2 displays the BSS values for the forecasts issued on 1 January and 1 April. It can be seen that the BSS values from both the approaches are greater than climatology, suggesting skillful long-lead forecasts. Forecasts issued on 1 April show better performance relative to those issued on 1 January. This increase in performance with a decrease in lead time is to be expected; on 1 April, almost all of the seasonal snow is present on the ground in the basin, thus providing accurate information of the resulting streamflow from the melt. Lower forecast performance is observed at higher threshold values at both lead times. This is because of fewer (rare) events, which results in fewer data points making the logistic regression unstable [*Bradley et al.*, 2003; *Clark and Slater*, 2006]. The skill from both methods is comparable. However, the logistic regression framework can

directly provide the categorical forecast at less computation cost which makes it quite attractive.

4. Summary

[28] We developed a simple and direct method to produce probabilistic categorical streamflow forecasts. In this, the categorical forecasts of the leading mode (or principal component) of the spatial streamflow at different thresholds is estimated via logistic regression. Large-scale climate features are used to obtain the predictors, and used in the logistic regression. These categorical forecasts are then uniformly transferred to the corresponding categorical forecast of streamflows at all the locations. The framework was applied to categorical forecasts of the April–June (spring) streamflows in the GRB at several month lead times starting from 1 December. The forecasts exhibited significant skill even at long lead times. Furthermore, the skills were comparable or better in some cases to those obtained from a best model nonparametric regression based forecasts of *Regonda et al.* [2006]. We also estimated skill scores from the multimodel ensemble forecasts of *Regonda et al.* [2006] and comparable results were observed. Both these approaches, we feel are complimentary and serve different purposes: The logistic regression method will be useful if a quick categorical forecast is required, while the ensemble approach can provide the entire probability density function of the streamflows that can be used to drive decision support models.

[29] The framework developed in this research is flexible and simple to implement. It works very well if the leading mode captures most of the data variance, and has uniform Eigenloadings and high correlations with all the basin streamflows, such as the case in the application to the GRB that we demonstrated. If there are more than one leading PC that capture a significant part of the spatial variance, then this framework can be applied to all the significant leading PCs and the estimated categorical forecasts from each of the PCs optimally combined following *Rajagopalan et al.* [2002]. Other potential improvements could include optimally combining categorical forecasts from multimodels and also using the nonparametric logistic regression [*Loader*, 1999] method.

[30] **Acknowledgments.** The first author thanks Patricia Weis-Taylor, Veerender Garg, Eric Noble, and Paul Block for their editorial suggestions on earlier versions of the manuscript. Support for this study by the NOAA Regional Integrated Sciences and Assessment Program (NOAA Cooperative Agreement NA17RJ1229) and the Center for Advanced Decision Support in Water and Environmental Systems (CADSWES) at the University of Colorado, Boulder, are thankfully acknowledged. The third author acknowledges support from the New Zealand Foundation for Research Science and Technology (contract C01X0401). Thanks are also owed to three anonymous reviewers and the Associate Editor for their insightful comments and suggestions.

References

Bradley, A. A., T. Hashino, and S. S. Schwartz (2003), Distributions-oriented verification of probability forecasts for small data samples, *Weather Forecast.*, *18*, 903–917.

Clark, M. P., and L. E. Hay (2004), Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, *J. Hydrometeorol.*, *5*, 15–32.

Clark, M. P., and A. G. Slater (2006), Probabilistic quantitative precipitation estimation in complex terrain, *J. Hydrometeorol.*, *7*, 3–22.

Day, G. N. (1985), Extended streamflow forecasting using NWSRFS, *J. Water Resour. Plann. Manage.*, *111*, 157–170.

Garen, D. C. (1992), Improved techniques in regression-based streamflow volume forecasting, *J. Water Resour. Plann. Manage.*, *118*, 654–670.

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.

Hosmer, D. S., and S. Lemeshow (1989), *Applied Logistic Regression*, John Wiley, Hoboken, N. J.

Lall, U. (1995), Recent advances in nonparametric function estimation, *U.S. Natl. Rep. Int. Union Geod. Geophys. 1991–1994, Rev. Geophys.*, *33*, 1093–1102.

Loader, C. (1999), *Local Regression and Likelihood*, Springer, New York.

Piechota, T. C., F. H. S. Chiew, J. A. Dracup, and T. A. McMahon (1998), Seasonal streamflow forecasting in eastern Australia and the El Nino–Southern Oscillation, *Water. Resour. Res.*, *34*, 3035–3044.

Rajagopalan, B., U. Lall, and S. Zebiak (2002), Optimal categorical climate forecasts through multiple GCM ensemble combination and regularization, *Mon. Weather Rev.*, *130*, 1792–1811.

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.

Von Storch, H., and F. W. Zwiers (1999), *Statistical Analysis in Climate Research*, Cambridge Univ. Press, New York.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, Elsevier, New York.

M. Clark, NIWA, PO Box 8602, Christchurch, New Zealand.

B. Rajagopalan and S. K. Regonda, Department of Civil, Environmental and Architectural Engineering, University of Colorado, Campus Box 421, Boulder, CO 80309-0421, USA. (satish-kumar.regonda@colorado.edu)