A Bayesian approach to probabilistic streamflow forecasts

Hui Wang, Brian Reich and Yeo Howe Lim

ABSTRACT

One-month-ahead streamflow forecasting is important for water utilities to manage water resources such as irrigation water usage and hydropower generation. While deterministic streamflow forecasts have been utilized extensively in research and practice, ensemble streamflow forecasts and probabilistic information are gaining more attention. This study aims to examine a multivariate linear Bayesian regression approach to provide probabilistic streamflow forecasts by incorporating gridded precipitation forecasts from climate models and lagged monthly streamflow data. Principal component analysis is applied to reduce the size of the regression model. A Markov Chain Monte Carlo (MCMC) algorithm is used to sample from the posterior distribution of model parameters. The proposed approach is tested on gauge data acquired during 1961–2000 in North Carolina. Results reveal that the proposed method is a promising alternative forecasting technique and that it performs well for probabilistic streamflow forecasts.

Key words | climate model forecasted precipitation, Gibbs Sampling, Markov Chain Monte Carlo, principal component analysis, water management

Hui Wang (corresponding author)
Bureau of Economic Geology,
Jackson School of Geosciences,
The University of Texas at Austin,
Texas 78758,
USA
E-mail: *hui.wang@beg.utexas.edu*

Brian Reich
Department of Statistics,
North Carolina State University,
Raleigh,
North Carolina 27695,
USA

Yeo Howe Lim
Department of Civil Engineering,
University of North Dakota,
Grand Forks,
North Dakota 58201,
USA

INTRODUCTION

Streamflow forecasting has been researched extensively in the water and hydrological engineering literature for the past several decades (Salas *et al.* 1980; Bartolini & Salas 1993; Lettenmaier & Wood 1993; Lima & Lall 2010) due to its importance in managing water resources systems, e.g. dam operation, water supply and irrigation planning. Generally, forecasting techniques can be characterized as one of the following two types depending on its output format: (1) deterministic (e.g. Hsu *et al.* 1995; Zealand *et al.* 1999); and (2) probabilistic and ensemble streamflow prediction (EPS) (e.g. Day 1985; Werner *et al.* 2004; Grantz *et al.* 2005; Vrugt *et al.* 2006; Wood & Lettenmaier 2006; Johnell *et al.* 2007; Wood & Schaake 2008; Wei & Watkins 2011; Najafi *et al.* 2012). Although deterministic streamflow forecasts have been utilized in water resource planning and management, the utility of ensemble and probabilistic streamflow forecasts are gaining attention (Maurer & Lettenmaier 2004; Golembesky *et al.* 2009; Eum & Kim 2010; Alemu *et al.* 2011). Wang (2012) used streamflow ensembles to investigate different water usage scenarios in deciding water allocation for different users. This study aims to examine multivariate linear Bayesian regression approach to provide probabilistic streamflow forecasts by incorporating gridded precipitation forecasts from climate models, as well as lagged monthly streamflow data.

There are many factors influencing monthly streamflow time series, including all the processes which contribute to the overall water movement cycle, e.g. precipitation, evaporation and infiltration. In building a statistical forecasting model, the components that are found significantly correlated with streamflow are often chosen as predictors. Autocorrelation in monthly streamflow data can often be detected due to storage effect of the basin. Hence, lagged streamflow time series are usually potential predictors for monthly streamflow forecasts (Piechota & Dracup 1999). Recently, studies have found that monthly forecasted precipitation is also highly correlated with streamflow data (e.g. Sankarasubramanian *et al.* 2008; Block *et al.* 2009) and the correlation value varies temporally and spatially. Given the large number of potentially correlated predictors that are available, principal component

doi: 10.2166/hydro.2012.080

analysis (PCA) is an appealing approach for reducing the number of parameters in the regression model (Sankarasubramanian *et al.* 2008). Many similar studies (e.g. Grantz *et al.* 2005; Moradkhani & Meier 2010) addressed predictor selection and reduction dimension.

In this study, PCA is first applied to reduce problem dimension. Further, the Bayesian approach is applied to estimate the parameters of the multivariate linear regression model. From a classical frequentist point of view, the model that is regarded the best fit is the one for which the set of model parameters produces the minimum difference between observation and model output. In contrast, the Bayesian approach regards the model parameter as 'random variables' with uncertainty. Prior information about such random variables can be incorporated; meanwhile, observed streamflow data are used as 'evidence' to update prior information. The Bayesian approach therefore enjoys the advantage of incorporating prior information of the model parameters and streamflow data to obtain the posterior distribution for model parameters of interest. The Bayesian approach has been applied in water resources and hydrological engineering (e.g. Vrugt *et al.* 2008; Jin *et al.* 2010; Wang & Harrison 2013) and has gained popularity due to advances in computing techniques and its advantage of naturally propagating uncertainty of model parameters to prediction distribution.

In this study, predictors for streamflow forecasting are first identified and then PCA is used to reduce the dimension of the regression problem. Bayesian multivariate linear regression is applied to estimate the regression parameters and, finally, the constructed linear regression is utilized to provide streamflow forecasts. As a general procedure proposed in this study, it is tested for monthly streamflow forecasting.

The objectives of this study are to: (1) propose a Bayesian linear regression approach for monthly streamflow forecasts; and (2) examine the performance of the proposed approach in a case study of streamflow prediction in North Carolina (NC), USA. Analysis of streamflow data presents several challenges, including non-normality and collinearity. In the following methodology section, we describe the Box–Cox transformation to deal with non-normality, PCA to deal with colinearity and finally the Bayesian linear regression used to make predictions. There follows a description of the data sources used in this study. The proposed method is applied to a streamflow gauge in North Carolina and the results are described, followed by a concise discussion and final conclusions.

METHODOLOGY

Box–Cox transformation

The Box–Cox transformation (Box & Cox 1964) is used to transform a time series which is non-normally distributed to approximate the normal distribution. The Box–Cox transformation is applied to the original time series y_i according to:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y_i), & \text{if } \lambda = 0 \end{cases} \tag{1}$$

where $y_i\ \{i = 1, 2, \ldots, n\}$ is the original time series and $y_i^{(\lambda)}\ \{i = 1, 2, \ldots, n\}$ is the transformed dataset. The parameter is estimated by maximizing the log-likelihood function. It can be easily implemented in Matlab function *boxcox*. After the Box–Cox transformation is applied to obtain the new time series, hypothesis tests, e.g. the Kolmogorov–Smirnov test, can be used to test the normality of transformed time series.

Principal component analysis

PCA (Shaw 2003; Abdi & Williams 2010) is a dimension reduction technique for building predictive models. It aims to replace a large number of correlated predictors with a small number of representative uncorrelated predictors, known as the principal components (PCs). Mathematic derivation and illustrative examples of the feature of PCA can be easily found in the literature (e.g. Joliffe 1996). The main advantage of such an approach is dimension reduction.

Bayesian linear regression

A multivariate linear regression model can be defined:

$$Y = X\beta + \Phi \tag{2}$$

where $Y = \{y_1, y_2, \ldots, y_n\}$ is the dependent variable; X is the $n \times p$ design matrix and the first column are all 1s; $\beta = \{\beta_1, \beta_2, \ldots, \beta_p\}$

is the regression coefficient vector and the first element corresponds to the intercept; and Φ is error with elements following the normal distribution with mean 0 and variance σ^2.

The unknown parameters in Equation (2) are the regression coefficients and error, denoted $\theta = (\beta, \sigma^2)$. In Bayesian context, these unknowns are treated as random variables rather than fixed values. The basic formula for Bayesian theorem is (e.g. Gelman *et al.* 1995; Carlin & Louis 2009):

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{\int P(Y|\theta)P(\theta)\mathrm{d}\theta} \tag{3}$$

The prior distribution reflects knowledge about θ before observing data. In a Bayesian analysis, this uncertainty is quantified with a probability distribution $P(\theta)$. The likelihood term $P(Y|\theta)$ is the distribution of the data given model parameters. In the case of Bayesian linear regression, the likelihood is derived by Equation (2). The denominator is an integration of the product over sample space. At the heart of a Bayesian analysis is the posterior distribution, $P(\theta|Y)$ which represents the current state of knowledge after observing the data. Bayes' Theorem provides a mathematically coherent way to update the prior based on the data to obtain the posterior. Evaluating the posterior is a challenging aspect of a Bayesian analysis, especially when the dimension of sample space is larger than 20 (Gilks *et al.* 1996). Markov Chain Monte Carlo (MCMC) was developed to avoid the calculation of the integration in the denominator of the right-hand side of Equation (3). It is used to sample directly from the posterior distribution $P(\theta|Y)$ and extract information about θ based on these samples.

The MCMC algorithm is used to draw dependent samples, $\theta^{(1)}$, $\theta^{(1)}$, $\theta^{(2)}$, … such that, after some point, the subsequent samples follow the posterior distribution, $P(\theta|Y)$. The initial portion of the chain, referred to as the burn-in, is discarded and the inference is based on the rest of the chain. There are numerous MCMC algorithms in the literature, but most of them are developed from Metropolis–Hasting sampling (Hastings 1970) and Gibbs sampling (Gelfand & Smith 1990). The Gibbs sampling approach is applied here due to its simplicity and the characteristic of the problem defined in this paper.

Gibbs sampling is a convenient algorithm when the priors are conjugate. A prior is called a conjugate prior if the posterior belongs to the same family of the prior probability distributions. One major advantage of using conjugate priors is that an analytical expression can be derived for the posterior and samples can be easily drawn from it (Carlin & Louis 2009). To facilitate the following discussion, a reparameterization of σ^2 is used and its reciprocal is defined as τ. A typical conjugate prior for the parameters is:

$$P(\theta) = P(\beta, \tau) = P(\tau)P(\beta|\tau) \tag{4}$$

where $P(\tau)$ is a gamma distribution with parameter a and b:

$$P(\tau) \propto \tau^{(a-1)}e^{-bt} \tag{5}$$

and $P(\beta|\tau)$ is a normal distribution:

$$P(\beta) \propto \exp\left(-\frac{1}{2\sigma_\beta^2}\sum_{j=1}^{P}\beta_j^2\right) \tag{6}$$

Using such priors ensures that the posterior distribution is analytically derived and samples can be drawn using a Gibbs sampler. Gibbs sampling proceeds by selecting initial values for all parameters, and then updating the parameters one at a time from their full conditional distributions which assume that all other parameters are temporarily fixed. After choosing parameters a, b and σ_β^2 for prior distribution of τ and β, Gibbs sampling is implemented in the following steps.

1. Assign initial values to regression parameters $\beta^{(0)}$ and $\tau^{(0)}$.
2. Sample $\beta^{(j+1)} \mid \tau^{(j)}$ from its conditional density for $\beta^{(j+1)} \sim N(M,V)$ where

$$V = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\tau(j) + \mathbf{I}_p\frac{1}{\sigma_\beta^2}\right)^{-1}$$

$M = \tau^{(j)}V\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix with 1s on the main diagonal and zero everywhere else.

3. Sample $\tau^{(j+1)}|\beta^{(j+1)}$ from its conditional density for

$$\tau^{(j+1)} = \Gamma\left[\frac{n}{2} + a, \left(Y - X\beta^{(j+1)}\right)^{\mathrm{T}}\left(Y - X\beta^{(j+1)}\right) + b\right]$$

4. If j is less than pre-designated iterations J, go to step 2. Otherwise, stop sampling.

Note that in step 2, the updated value of β is used to obtain the sample of τ and the updated τ value is used in step 3 to obtain the sample for β. Convergence diagnostics are applied to determine whether the algorithm has converged and, if it is valid, to make any inference about the posterior distribution of model parameters based on MCMC samples. The Gibbs sampling algorithm is fundamentally different from an optimization algorithm. Rather than stopping once the optimal value has been reached, after the algorithm has converged it is run for many more iterations to produce representative samples from the posterior. It is therefore common practice to run the algorithm for a fixed number of iterations and then retain all samples after a burn-in period.

The posterior mean of MCMC chain of each regression parameter can be used to provide single-value streamflow forecasts if only interested in deterministic streamflow forecasts. Ensemble streamflow forecasts can be obtained if each set of MCMC samples in the posterior distribution is utilized to run the linear regression model. This can also be used to develop probabilistic streamflow forecasts, e.g. probabilistic forecasting of streamflow falling into 'below normal' (BN), 'normal' (N) or 'above normal' (AN) categories.

DATA

Streamflow data

Monthly streamflow data for USGS gauge number 02102000 are downloaded from surface water USGS portal (http://waterdata.usgs.gov/nwis). The drainage area contributing to this gauge is 1,434 square miles (3,714 km^2) and it is located 1 mile (1.61 km) upstream of Lockville dam, which is usually operated for hydroelectricity generation. The location of the gauge is represented as a triangle in Figure 1. Streamflow forecasting is of great importance for planning hydroelectricity generation. In this study, we focus on the dataset during the period 1961–2007. As shown in Figure 2, data are heavily right skewed. The whole dataset is split into two: data from the years 1961–1996 are used for calibration and from years 1997–2007 for model validation.

Climate forecasts

For climate forecasts, we consider one-month-ahead retrospective precipitation forecasts from ECHAM4.5 (the European Centre for Medium-Range Weather Forecasts, Hamburg; Roeckner *et al.* 1996) forced with constructed

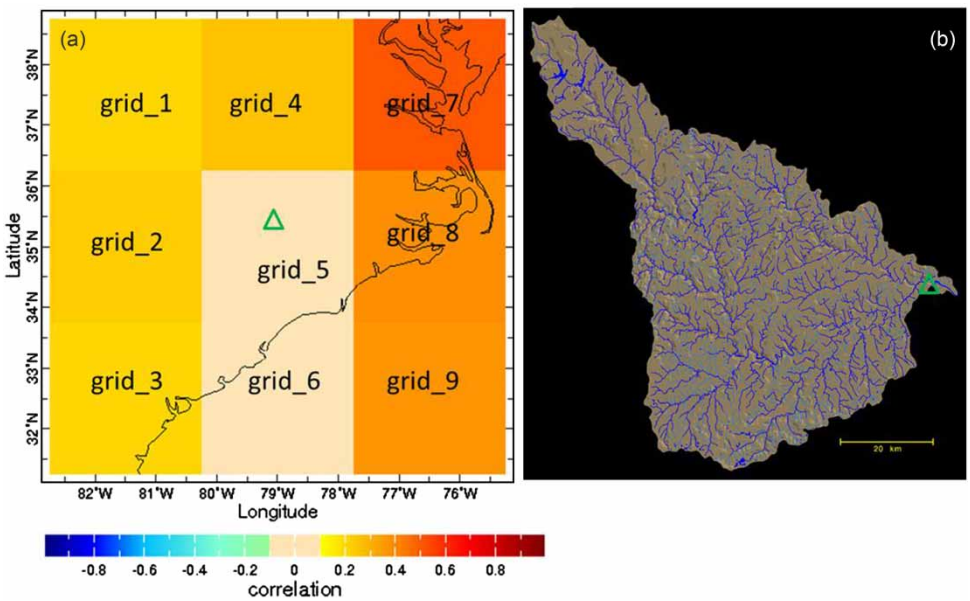

Figure 1 | (a) Correlation between streamflow data at the gage represented by a triangle and retrospective monthly precipitation data at neighboring grids or resolution 2.5° × 2.5°. (b) The watershed contributing to the streamflow gauge.

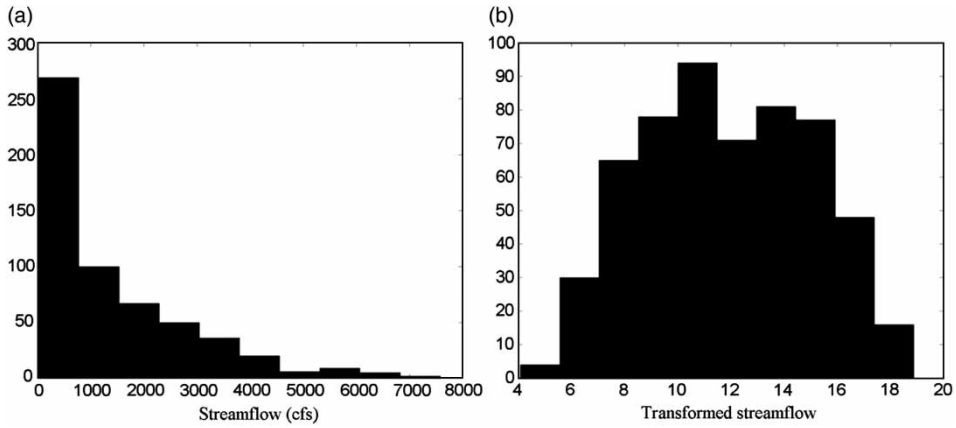

Figure 2 | (a) Histogram of monthly streamflow data and (b) Box–Cox transformed dataset at gauge 02102000.

analogue sea-surface temperatures (SSTs) (Li & Goddard 2005). These retrospective forecasts are available for 7 months ahead and are updated every month from January 1957. Gridded monthly precipitation forecasts for nine grids of 2.5° latitude $\times$ 2.5° longitude were downloaded from http://iridl.ldeo.columbia.edu.

RESULTS AND ANALYSIS

Box–Cox transformation

Figure 3(b) shows that the transformed data after a Box–Cox transformation with $\lambda = 0.159$ is approximately normal and it passes the Kolmogorov–Smirnov test of normality. There are two advantages of using the Box–Cox transform: one is that it avoids negative streamflow values once it is being transformed back to the original space; the other is that normally distributed time series of streamflow satisfy the assumption of Equation (2).

Principal component analysis

Lagged streamflow data are usually significantly correlated with current-month streamflow due to system memory effect for watersheds where groundwater or surface water storage plays an important role in the hydrological cycle. Figure 3 shows the correlation between monthly streamflow and lag-1

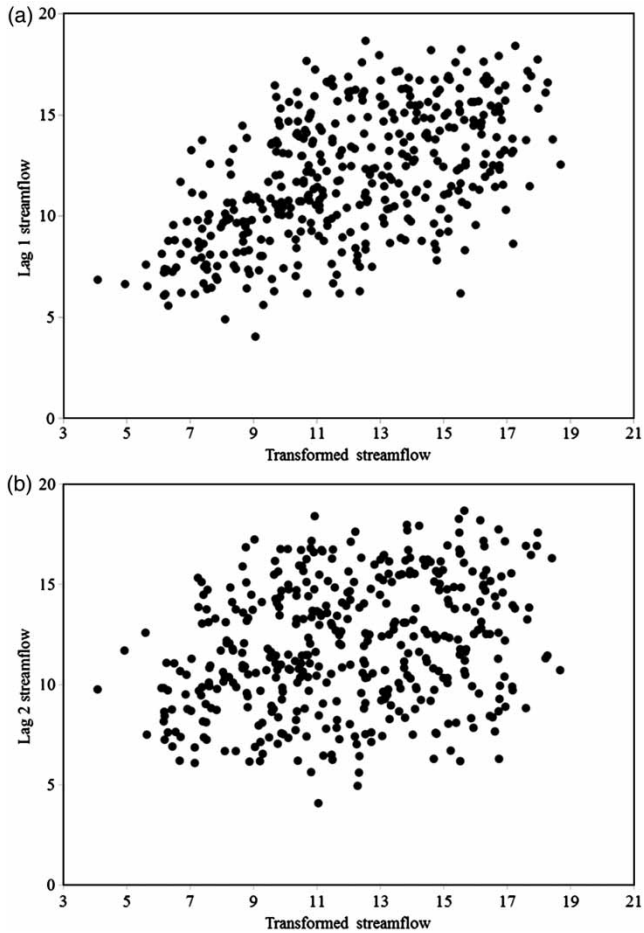

Figure 3 | (a) Scatter plot between transformed streamflow data and lag-1 streamflow time series (correlation 0.56) and (b) scatter plot between transformed streamflow data and lag-2 streamflow time series (correlation 0.27).

streamflow, as well as monthly streamflow and lag-2 streamflow time series. Correlation values are 0.56 and 0.27 respectively, both of which are significant at 5% significance level.

It is obvious that precipitation is usually well correlated with streamflow. However, if precipitation observation is used to build the regression model for streamflow forecasting, precipitation observation data cannot be obtained until the end of the month of interest. This poses a challenge in using observed precipitation for streamflow forecasting purposes. To alleviate this problem, precipitation forecasts may be used as an approximation to precipitation observation (e.g. Piechota & Dracup 1999). Recent studies have shown that climate forecasted precipitation is correlated with streamflow data for places where climate models have good performance (e.g. Sankarasubramanian *et al.* 2008). Figure 1(a) shows the correlation between one-month lead time precipitation forecasts from ECHAM4.5 at neighboring grids and the concurrent monthly streamflow data. For seven out of nine grids, there is significant correlation between the streamflow data and the retrospective precipitation forecasts with a minimum correlation value of 0.214 and a maximum of 0.578.

In total, nine variables are identified as the predictors for the monthly streamflow series. These are lag-1 streamflow, lag-2 streamflow and ECHAM4.5 forecasted monthly precipitation over seven neighboring grids of the streamflow guage. To reduce the model dimension, PCA is applied to those nine variables.

As shown in Figure 4(a), the first three PCs explain 64.2, 15.3 and 12.3% of the totally variance exhibiting in the nine variables. Over 90% of the variance can be explained by those PCs; they are therefore identified as the three predictors for monthly streamflow forecasting. The dimension is reduced significantly from nine to three by choosing three PCs and abandoning the rest. Apart from the dimension reduction, correlation among PCs is zero since they are orthogonal to each other. Figure 4(b) shows the coefficients of the variables contributing to the PCs. The absolute values of the coefficients represent the contribution to the PCs. For example, lag-1 streamflow data, lag-2 streamflow data, precipitation forecasts over grid 1, grid 2 and grid 3 are the major components of the first PC (PC_1), while the other variables have negligible contribution to PC_1.

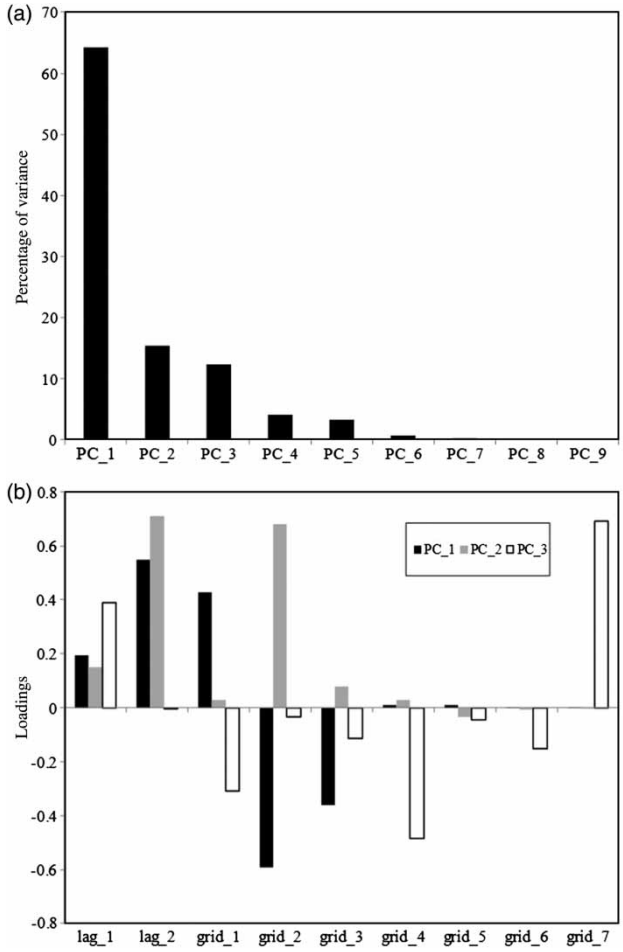

Figure 4 | (a) Variance explained by the principal components and (b) coefficients associated with each of the nine variables in constructing the first three principal components.

Markov Chain Monte Carlo diagnostic

A MCMC chain is composed of samples from all iterations. As an example, Figure 5 shows the samples from iteration 1,000 to iteration 1,500. Before drawing conclusions about the posterior distribution of interested parameters from a MCMC chain, it is necessary to examine its convergence. A converged MCMC chain indicates that the samples are from the posterior distribution of the estimated parameters. There are many different ways to analyze the convergence of the chain (Cowles & Carlin 1996), some of which calculate the statistics from the samples of the chain. MCMC samples for parameters which pass the convergence diagnostics; they can be used to infer posterior distributions. Figure 6 shows

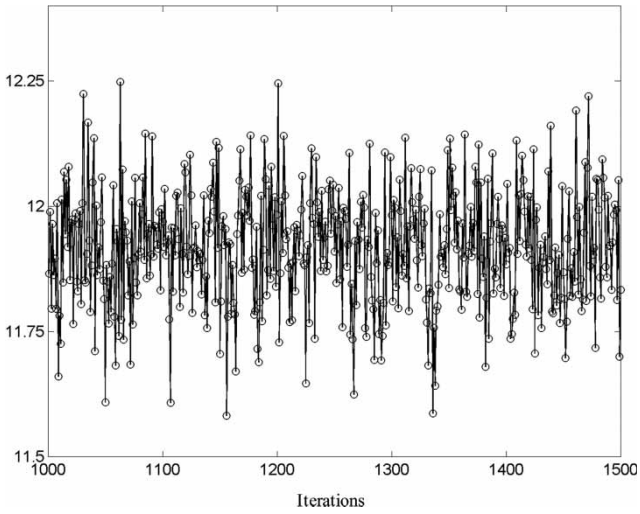

Figure 5 | Excerpt of MCMC sampling for β_1.

the running mean of the five parameters. The total length of the MCMC chain is 20,000 and the first 1,000 iterations are treated as 'burn-in'. The last 19,000 samples are used to draw conclusions about the regression parameters.

Figure 7 shows the histogram of the posterior samples for each parameter. The posterior mean of the MCMC samples is often suggested as the best estimate of a parameter (Carlin & Louis 2009), and is shown as the vertical line in each panel of Figure 7. Similar to the confidence interval in classical statistics approach, Bayesian sampling

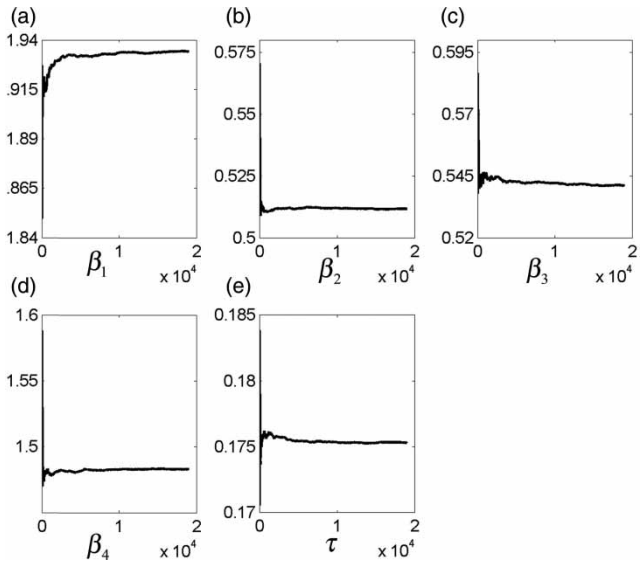

Figure 6 | Running mean of the MCMC chain for all parameters.

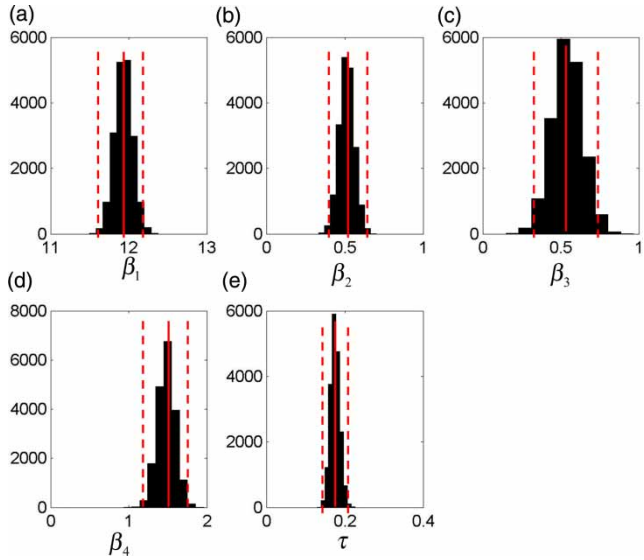

Figure 7 | Posterior distribution and 95% credible interval of the model parameters.

can offer a credible interval. The difference in the calculations of confidence interval and credible interval is that the latter is obtained with both prior information, e.g. prior probabilities, and data, e.g. streamflow observations. The 95% credible intervals are represented as the two dashed lines in each panel. For example, a 95% credible interval for β_1 indicates that 95% of the mass of the probability density resides between 11.71 and 12.16.

To provide monthly streamflow forecasts, the posterior mean of the regression parameters can be used if we are only interested in deterministic streamflow forecasts. This study aims to investigate the capacity of the regression model to offer probabilistic forecasting. All posterior samples of regression parameters are therefore used to provide one-month-ahead streamflow forecasts and ensemble streamflow forecasts, comprising 19,000 individual forecasting members. By doing so, we account for uncertainty in the regression parameters.

Streamflow forecast performance

Streamflow data for 1996–2007 are used for validation purposes. Starting from January 1996, monthly streamflow forecasts are based on the calibrated regression model built using data from 1961 to 1995 via the Bayesian MCMC approach. For example, streamflow forecasting for January

1996 is based on observations from the last two months and retrospective monthly precipitation forecasts for January 1996, all of which are used to construct the three PCs. When forecasting streamflow for February 1996, streamflow data from January 1996 is incorporated in constructing the PCs.

As described in the previous section, all posterior samples of regression parameters are used to provide one-month-ahead streamflow forecasts and ensemble streamflow forecasts comprising 19,000 individual forecasting members. Each forecast is transformed back to the original space, and this can be easily done via Equation (1). The mean value of the probabilistic streamflow forecasts can be used as deterministic forecasts (Figure 8), whereas there are two additional outputs derived from ensemble forecasts. One is the credible interval as shown in Figure 8; the 95% credible interval is the range between 2.5% percentile and 97.5% of the ensemble forecasts for a specific month. The other is probabilistic forecasts of streamflow residing in different categories, such as below normal (BN), normal (N) and above normal (AN). These three categories are divided by 33% percentile and 67% percentile of the long-term mean of the month of interest. Figure 9 shows the probabilistic forecasts for each category over the validation period.

There are many different ways to verify probabilistic forecasts (Wilks 2006); ranked probabilistic score (RPS) is chosen in this study due to its simplicity. RPS summarizes the sum of square of errors in the cumulative probabilities of the given

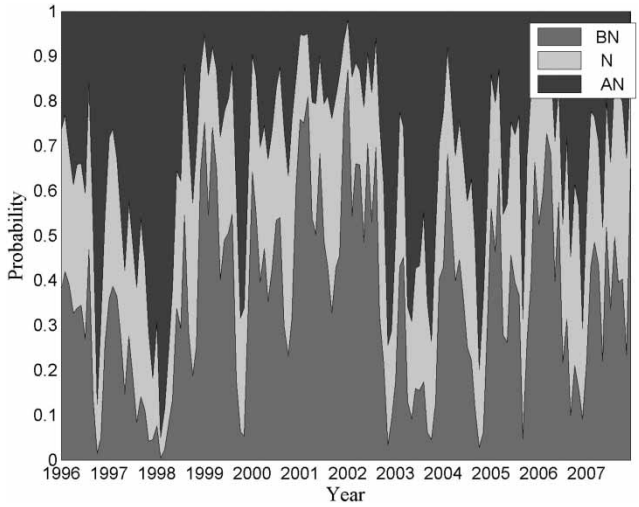

Figure 9 | Monthly probabilistic forecasts over the three categories for every month during the validation period.

categorical forecast and the observed category, which has an assigned value of 1. The lower the value of RPS, the smaller is the error between accumulative probabilities and observed category. When RPS is 0, this indicates perfect forecasts. For example, 100% of the probability mass resides in the BN category and the observation is BN. For three-category forecasts used in this study, the upper boundary of RPS is 2 when none of the probability mass resides in the observation category. A base model is climatology, where the same amount of probability mass is assigned for each of the three categories. Figure 10 shows the average RPS for different months over all years for the proposed method and climatology.

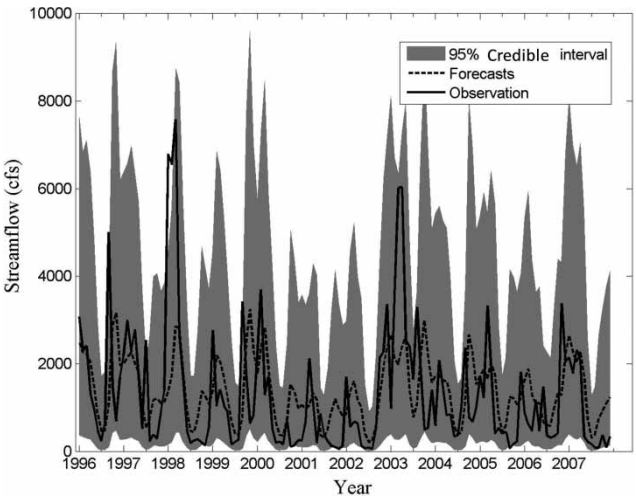

Figure 8 | Comparison between streamflow observation and forecasts; the shaded area denotes 95% credible interval deriving from ensemble streamflow forecasts.

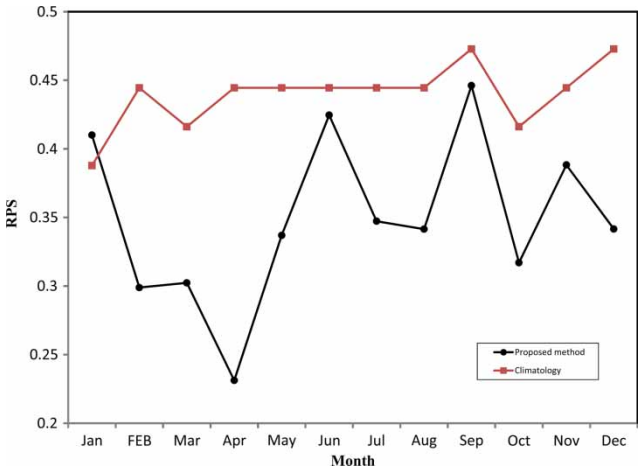

Figure 10 | Comparison of average rank probability score for every month during 1996–2007.

Comparison with other models

This study aims to examine the proposed Bayesian regression method which explicitly incorporates precipitation forecasts from climate models to provide streamflow forecasts.

In the comparison below, the proposed approach is referred to as model A. It is compared with another three models, namely models B, C and D. Model B is the same except that the ordinary least squares (OLS) approach (and not Bayesian) is applied to obtain regression coefficients. Model C is the same as model B except that PCA is not used and all the nine predictors are represented in the design matrix. To examine the utility of precipitation forecasts from ECHAM4.5, only lagged streamflow data are used as predictors in model D to build the regression. Model settings and solving techniques are listed in Table 1.

Long-term bias, which is the difference between the average of streamflow observation during the validation period and the average of streamflow forecasts, is of importance to evaluate a regression model. Long-term bias is 149.5, 207.7, 226.6 and 292.6 for models A, B, C and D, respectively.

The Nash–Sutcliffe efficiency coefficient is often used to assess the predictive performance of a hydrological model and it is calculated:

$$E = 1 - \frac{\sum_{t=1}^{T} (Q_0^t - Q_m^t)^2}{\sum_{t=1}^{T} (Q_0^t - \overline{Q_0})^2} \quad (7)$$

Table 1 | Different models considered in this study

Model	Predictors	PCA	Solving technique	Correlation
A	Lag-1 and lag-2 streamflow, precipitation forecasts from 7 grids	Yes	Bayesian MCMC	0.72
B	Lag-1 and lag-2 streamflow, precipitation forecasts from 7 grids	Yes	OLS	0.70
C	Lag-1 and lag-2 streamflow, precipitation forecasts from 7 grids	No	OLS	0.69
D	Lag-1 and lag-2 streamflow	No	OLS	0.58

where Q_0^t is observed streamflow; Q_m^t is streamflow prediction, $\overline{Q_0}$ is averaged streamflow during validation period and T is the total number of forecasting periods. Nash–Sutcliffe coefficients are 0.40, 0.42, 0.41 and 0.26 for models A, B, C and D, respectively. There is therefore no substantial difference between the first three models, while model D has the poorest performance.

Similar to the Nash–Sutcliffe efficiency coefficient, the coefficient of determination (r^2) is often used to indicate how much variance evident in the streamflow data during the validation period could be explained by the regression model. Correlation value (r) is shown in Table 1, and it can be seen that model A is of the highest correlation and model D is of the lowest correlation.

DISCUSSION

One advantage of the Bayesian approach is that prior information of the parameters of interest, e.g. mean and variance for the multivariate normal distribution of β_0, β_1, β_2 and β_3 can be easily incorporated into consideration. One question raised is whether this approach is robust, such that it is not sensitive to the availability of prior information? There are at least two different decisions to make: one is the form of probability density function for prior, e.g. gamma distribution for τ and normal distribution for β; the other is the parameters of the probability density function once a specific function is chosen. In this study, two parameters (a and b) of the gamma distribution of τ, as well as prior mean and variance of the normal distribution for β, are tested for different values. We find that results are not sensitive to the choice of priors for these data.

The main purpose of this study is to investigate the scheme providing adaptive monthly streamflow forecasts. This scheme enjoys the flexibility of extending to streamflow forecasts over different temporal scales, e.g. seasonal and annual, since probabilistic streamflow has its own application in water resources planning and management. There are many different ways to further fine-tune the Bayesian linear regression model, e.g. selecting different predictors for different forecasting months (Wei & Watkins 2011; Schepen *et al.* 2012). In the forecast verification section,

a general assessment is therefore made other than calibrating the regression model for each month.

CONCLUSION

In this study, an alternative scheme of probabilistic monthly streamflow forecasts is investigated. This scheme incorporates lagged streamflow data, as well as retrospective climate forecasts of gridded precipitation. PCA is applied to reduce problem dimension and illuminate correlation among the predictors. The first three PCs explain more than 90% of the variance exhibiting in predictors, and are used as predictors. At the same time, the Box–Cox transform is applied to the original monthly streamflow data and the transformed dataset follow approximate normal distribution. The Bayesian approach is used to estimate the parameters of the linear regression model. The conjugate prior for the parameters of interest is used and this facilitates the MCMC sampling process, where the Gibbs sampling approach is applied to sample from the posterior distribution of parameters. An illustrative example demonstrates the effectiveness and robustness of the proposed approach. Further studies extending the proposed scheme to seasonal/annual time scales and selecting the best predictors for different seasons are needed.

ACKNOWLEDGEMENTS

The authors thank the three anonymous reviewers for their insightful comments.

REFERENCES

Abdi, H. & Williams, L. J. 2010 Principal component analysis. Wiley Interdisciplinary Reviews. *Computational Statistics* **2**, 433–459.

Alemu, T. E., Palmer, N. P., Polebitski, A. & Meaker, B. 2011 Decision support system for optimizing reservoir operations using ensemble streamflow predictions. *Journal of Water Resource Planning and Management* **137**, 72–82.

Bartolini, P. & Salas, J. D. 1993 Modeling of streamflow processes at different time scales. *Water Resources Research* **29** (8), 2573–2588.

Block, P., Souza Filho, A., Sun, L. & Kwon, H. 2009 A streamflow forecasting framework using multiple climate and hydrological models. *Journal of the American Water Resources Association* **45** (4), 828–843.

Box, G. E. P. & Cox, D. R. 1964 An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26** (2), 211–252.

Carlin, B. P. & Louis, T. A. 2009 *Bayesian Methods for Data Analysis*. CRC Press, Boca Raton, Florida.

Cowles, M. K. & Carlin, B. P. 1996 Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* **91** (434), 883–904.

Day, G. N. 1985 Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management* **111** (2), 157–170.

Eum, H.-I. & Kim, Y.-O. 2010 The value of updating ensemble streamflow prediction in reservoir operations. *Hydrological Processes* **24**, 2888–2899.

Gelfand, A. E. & Smith, A. F. M. 1990 Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association* **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. S. 1995 *Bayesian Data Analysis*. Chapman & Hall, London.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. 1996 Introducing Markov Chain Monte Carlo. In: *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson & D. J. Spiegelhalter, eds). Chapman & Hall, London.

Golembesky, K., Sankarasubramanian, A. & Devineni, N. 2009 Improved drought management of Falls Lake Reservoir: role of multimodel streamflow forecasts in setting up restrictions. *Journal of Water Resources Planning and Management* **135**, 188.

Grantz, K., Rajagopalan, B., Clark, M. & Zagona, E. 2005 A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resources Research* **41**, W1040.

Hastings, W. K. 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hsu, K., Gupta, H. V. & Sorooshian, S. 1995 Artificial neural network modeling of the rainfall–runoff process. *Water Resources Research* **31** (10), 2517–2530.

Jin, X., Wang, H. & Ranjithan, S. R. 2010 Bayesian inference of groundwater contamination source. World Environmental and Water Resources Congress 2010: Challenges of Change, 16–20 May 2010, Providence, RI.

Johnell, A., Lindstrom, G. & Olsson, J. 2007 Deterministic evaluation of ensemble streamflow predictions in Sweden. *Nordic Hydrology* **38** (4–5), 441–450.

Joliffe, I. T. 1996 *Principal Component Analysis*. Springer-Verlag, New York.

Lettenmaier, P. D. & Wood, F. E. 1993 Hydrological forecasting. Chapter 26 In: *Handbook of Hydrology* (D. Maidment, ed.). McGraw-Hill, New York.

Li, S. & Goddard, L. 2005 Retrospective Forecasts with ECHAM4.5 AGCM IRI. Technical Report, 05-02 December,

International Research Institute for Climate and Society, University of Columbia, NY.

Lima, C. & Lall, U. 2010 Climate informed monthly streamflow forecasts for the Brazilian hydropower network using a periodic ridge regression model. *Journal of Hydrology* **380**, 438–449.

Maurer, E. P. & Lettenmaier, D. P. 2004 Potential effects of long-lead hydrologic predictability on Missouri River main-stem reservoirs. *Journal of Climate* **17** (1), 174–186.

Moradkhani, H. & Meier, M. 2010 Long-lead water supply forecast using large-scale climate predictors and independent component analysis. *Journal of Hydrologic Engineering* **15**, 744.

Najafi, M. R., Moradkhani, H. & Piechota, T. 2012 Ensemble streamflow prediction: climate signal weighting methods vs. climate forecast system reanalysis. *Journal of Hydrology* **442–443**, 105–116.

Piechota, C. T. & Dracup, A. J. 1999 Long-range streamflow forecasting using EI Nino-Southern Oscillation indicators. *Journal of Hydrologic Engineering* **4**, 144–151.

Roeckner, E., Arpe, K., Bengtsson, L., Christoph, M., Claussen, M., Dümenil, L., Esch, M., Giorgetta, M., Schlese, U. & Schulzweida, U. 1996 The atmospheric general circulation model ECHAM-4: Model description and simulation of present-day climate. Report 218. Retrieved at: http://www.mpimet.mpg.de/fileadmin/publikationen/Reports/MPI-Report_218.pdf.

Salas, J. D., Delleur, J. W., Yevjevich, V. & Lane, W. L. 1980 *Applied Modeling of Hydrologic Time Series*. Water Resources Publication, Littleton, Colorado.

Sankarasubramanian, A., Lall, U. & Espinueva, S. 2008 Role of retrospective forecasts of GCMs forced with persisted SST anomalies in operational streamflow forecasts development. *Journal of Hydrometeorology* **9**, 212–227.

Schepen, A., Wang, Q. J. & Robertson, D. 2012 Evidence for using lagged climate indices to forecast Australian seasonal rainfall. *Journal of Climate* **25** (4), 1230–1246.

Shaw, P. J. A. 2003 *Multivariate Statistics for the Environmental Sciences*. Hodder Arnold, London.

Vrugt, J. A., Braak, C. J. F., Clark, M. P., Hyman, J. M. & Robinson, B. A. 2008 Treatment of input uncertainty in hydrological modeling: doing hydrology backwards with Markov Chain Monte Carlo simulation. *Water Resources Research* **44**, 10. 1029/2007WR006720.

Vrugt, J. A., Hoshin, V. G., Ó Nualláin, B. & Bouten, W. 2006 Real-time data assimilation for operational ensemble streamflow forecasting. *Journal of Hydrometeorology* **7**, 548–565.

Wang, H. 2012 Improved Streamflow in Adaptive Reservoir Operation. PhD Thesis, North Carolina State University, North Carolina.

Wang, H. & Harrison, K. W. 2013 Bayesian update method for contaminant source characterization in water distribution systems. *Journal of Water Resources Planning and Management* **139** (1), 13–22.

Wei, W. & Watkins, W. D. 2011 Probabilistic streamflow forecasts based on hydrologic persistence and large-scale climate signals in Central Texas. *Journal of Hydroinformatics* **13** (4), 760–774.

Werner, K., Brandon, D., Clark, M. & Gangopadhyay, S. 2004 Climate index weighting schemes for NWS ESP-based seasonal volume forecasts. *Journal of Hydrometeorology* **5**, 1076–1090.

Wilks, D. 2006 *Statistical Methods in the Atmospheric Sciences*, 2nd ed. Elsevier, Amsterdam.

Wood, A. W. & Lettenmaier, D. P. 2006 A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bulletin of the American Meteorological Society* **87** (12), 1699–1712.

Wood, A. W. & Schaake, J. C. 2008 Correcting errors in streamflow forecast ensemble mean and spread. *Journal of Hydrometeorology* **9**, 132–148.

Zealand, C. M., Burn, H. D. & Simonovic, P. S. 1999 Short term streamflow forecasting using artificial neural networks. *Journal of Hydrology* **214**, 32–48.

First received 22 April 2012; accepted in revised form 22 August 2012. Available online 20 November 2012