Journal of Hydrology 497 (2013) 80–96

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts – A Hydrologic Model Output Statistics (HMOS) approach

Satish Kumar Regonda [a,b,*], Dong-Jun Seo [c], Bill Lawrence [d], James D. Brown [e], Julie Demargne [f]

[a] NOAA, National Weather Service, Office of Hydrologic Development, Silver Spring, MD 20910, USA
[b] Riverside Technology, Inc., Fort Collins, CO 80525, USA
[c] Department of Civil Engineering, The University of Texas at Arlington, Box 19308, 416 Yates St. Suite 425, Arlington, TX 76019, USA
[d] NOAA, National Weather Service, Arkansas-Red Basin River Forecast Center, Tulsa, OK 74128, USA
[e] Hydrologic Solutions Limited, Southampton, UK
[f] Consultant, Montpellier, France

ARTICLE INFO

Article history:
Received 20 August 2012
Received in revised form 17 January 2013
Accepted 14 May 2013
Available online 28 May 2013
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Ezio Todini, Associate Editor

Keywords:
Hydrologic ensemble forecasting
Model output statistics
Total predictive uncertainty
Single-valued operational streamflow forecast
Normal quantile transform
Ensemble verification

SUMMARY

We present a statistical procedure for generating short-term ensemble streamflow forecasts from single-valued, or deterministic, streamflow forecasts produced operationally by the U.S. National Weather Service (NWS) River Forecast Centers (RFCs). The resulting ensemble streamflow forecast provides an estimate of the predictive uncertainty associated with the single-valued forecast to support risk-based decision making by the forecasters and by the users of the forecast products, such as emergency managers. Forced by single-valued quantitative precipitation and temperature forecasts (QPF, QTF), the single-valued streamflow forecasts are produced at a 6-h time step nominally out to 5 days into the future. The single-valued streamflow forecasts reflect various run-time modifications, or "manual data assimilation", applied by the human forecasters in an attempt to reduce error from various sources in the end-to-end forecast process. The proposed procedure generates ensemble traces of streamflow from a parsimonious approximation of the conditional multivariate probability distribution of future streamflow given the single-valued streamflow forecast, QPF, and the most recent streamflow observation. For parameter estimation and evaluation, we used a multiyear archive of the single-valued river stage forecast produced operationally by the NWS Arkansas-Red River Basin River Forecast Center (ABRFC) in Tulsa, Oklahoma. As a by-product of parameter estimation, the procedure provides a categorical assessment of the effective lead time of the operational hydrologic forecasts for different QPF and forecast flow conditions. To evaluate the procedure, we carried out hindcasting experiments in dependent and cross-validation modes. The results indicate that the short-term streamflow ensemble hindcasts generated from the procedure are generally reliable within the effective lead time of the single-valued forecasts and well capture the skill of the single-valued forecasts. For smaller basins, however, the effective lead time is significantly reduced by short basin memory and reduced skill in the single-valued QPF.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The importance and value of uncertainty information in hydrologic forecasts is well known (Krzysztofowicz, 2001; NRC, 2004; Roulin, 2007; McCollor and Stull, 2008; Muluye, 2011; Verkade and Werner, 2011; Boucher et al., 2012; Ramos et al., 2012; Dale et al., 2012). Today, quantifying uncertainty in hydrologic and water resources forecasts is recognized as one of the most pressing needs in operational hydrology (NRC, 2006a, 2006b). Because of numerous sources of uncertainty involved in hydrologic prediction, however, accurate modeling of the total predictive uncertainty associated with future streamflow is a major challenge (e.g., Krzysztofowicz, 1999; Seo et al., 2006; Pappenberger and Beven, 2006). Broadly, the various uncertainties in hydrologic prediction may be grouped into two categories, input uncertainty and hydrologic uncertainty. The former comprises uncertainties in the input for hydrologic models, such as precipitation and temperature. The latter comprises uncertainties in the initial conditions, parameters and structures of the hydrologic models, and those due to human control and influences such as flow regulations. Generally speaking, the total predictive uncertainty can be modeled in two ways,

* Corresponding author at: NOAA, National Weather Service, Office of Hydrologic Development, Silver Spring, MD 20910, USA. Tel.: +1 301 713 0640x133; fax: +1 301 713 0963.
E-mail address: satish.regonda@noaa.gov (S.K. Regonda).

0022-1694/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2013.05.028

namely source-specifically or in a lumped fashion, as described below.

The source-specific approach models the various significant uncertainties individually and propagate them in time and space through the hydrologic models, from which the total uncertainty may be quantified with the addition of the residual uncertainty modeled stochastically as necessary (e.g., Krzysztofowicz, 1999; Seo et al., 2006, 2010; Demargne et al., 2013). An example of such an approach for headwater basins is the Bayesian Forecasting System (BFS, Krzysztofowicz, 1999, 2002) which models input and hydrologic uncertainties separately and integrates them via the total probability law following stratified sampling. The Hydrologic Ensemble Forecast Service (HEFS) of the U.S. National Weather Service (NWS) (Demargne et al., 2013) is another source-specific approach in which input and hydrologic uncertainties are modeled separately and then propagated and integrated via random sampling. In general, the source-specific approach may be preferred for modeling the total predictive uncertainty for the following reasons. The input uncertainty depends greatly on the lead time whereas, except for the initial condition uncertainty and the anthropogenic uncertainty, the hydrologic uncertainty does not. Hence, by modeling the latter separately, one may attain parsimony in modeling of hydrologic uncertainty. Also, by modeling the major sources of hydrologic uncertainty individually, the residual uncertainty may be made stochastically as structureless as possible, which reduces the data requirement for and complexity of the stochastic modeling of the hydrologic residual uncertainty. Because it requires modeling of all significant sources of uncertainty, however, the source-specific approach is generally expensive to develop and implement. As an intermediate solution between the source-specific approach for uncertainty modeling and the existing single-valued approach which lacks uncertainty modeling, one may consider generating probabilistic forecasts from the operationally-produced single-valued streamflow forecasts. The aim of this paper is to develop and evaluate such a technique.

Applying statistical techniques to single-valued numerical model output to generate probabilistic forecasts has been widely used in weather forecasting, the prime example being Model Output Statistics (MOS, Glahn and Lowry, 1972). In hydrology, however, there are only a few published papers on generating probabilistic forecasts from single-valued, or deterministic, forecasts by modeling the predictive uncertainty in the single-valued hydrologic forecasts (Montanari and Grossi, 2008; Reggiani and Weerts, 2008; Coccia and Todini, 2011; Bogner and Pappenberger, 2011; Weerts et al., 2011; Smith et al., 2012). Much research has been focused on producing probabilistic forecasts from hydrologic simulations, which account only for hydrologic uncertainty (Montanari and Brath, 2004; Seo et al., 2006; Chen and Yu, 2007; Hantush and Kalin, 2008; Montanari and Grossi, 2008; Todini, 2008; Bogner and Pappenberger, 2011; Zhao et al., 2011; Brown and Seo, 2012). Although one could apply similar techniques to model each of the hydrologic and total uncertainties, the same performance may not be expected because the hydrologic and total uncertainties differ in structure. For example, the total uncertainty is lead-time dependent whereas the hydrologic uncertainty, except for the initial condition uncertainty and certain anthropogenic uncertainties, is not. Many techniques that model either hydrologic or total uncertainty convert streamflow variables into Gaussian space via Normal Quantile Transform (NQT). The transformation is done mainly because marginally-normal variables are more likely to satisfy multivariate normality, under which many statistical estimation techniques are optimal in the mean square error sense. The modeling techniques reported in the literature mainly belong to the family of linear regression or time series modeling techniques. Montanari and Grossi (2008) divide the forecast error into two types, positive and negative errors, and develop regression models in the normal space for each type. The regression models express the forecast error as a function of the forecast flow, forecast errors at preceding time steps and observed antecedent precipitation. For a given flow forecast, an ensemble of errors is produced from the regression models. The errors are then added to the forecast in proportion to the climatological frequency of occurrence of the two types of error to obtain the final ensemble. The authors develop the regression models lead time-specifically without modeling the time-dependent structure of the errors. As such, the resulting ensemble members at successive lead times do not preserve the temporal correlation of the observed flow. Note that, for applications that support decisions on how water is to be managed over a range of time scales, it is important that the ensemble traces possess realistic temporal correlation structure. Reggiani and Weerts (2008) apply a Bayesian processor and estimate the conditional probability distribution function of water level forecasts given all information (upstream observations and deterministic forecasts) at the time of the forecast. The Bayesian processor combines prior and likelihood distributions in the normal space, and estimates season- and forecast lead time-specific posterior distributions of water levels. Coccia and Todini, 2011 model the joint distribution of forecast and observed flows in the normal space by considering the entire data set as one region or dividing the entire normal region into one or more sub-regions, defined as truncated normal distributions, and estimate the predictive uncertainty accordingly for each lead time as a confidence interval. Weerts et al. (2011) estimate the full probability distribution of forecast error by estimating a set of quantiles of the distribution via quantile-specific regression equations. The forecast error distributions specific to lead time are estimated and combined with forecasts to yield the full probability distribution of forecasts. Smith et al. (2012) modeled observed flows as a function of deterministic model predictions plus noise with a simple linear regression model, and model the regression coefficient or adaptive gain, i.e., time varying correction for the bias in the model forecast, using time series modeling techniques. Unlike the above studies, Bogner and Pappenberger (2011) decomposed the forecast output into multiple temporal scales using wavelet transformation and then modeled across scales with a Vector-AutoRegressive model with eXogenous input (VARX) and reconstructed back into the original time domain.

The objective of this work is to develop a technique that, given the operationally-produced single-valued forecast, produces an ensemble streamflow forecast that is reliable, i.e., probabilities of predicted events match observed frequencies, and captures the skill in the single-valued forecast. We refer to this technique as Hydrologic Model Output Statistics (HMOS). The HMOS technique is similar to the above-mentioned techniques in that it is based on bivariate modeling of the observed and forecast flows in the Gaussian space via linear regression, but differs in that it produces ensemble traces that preserve the serial correlation over successive lead times. The novel aspects and new contributions of this work include: (i) the use of the operationally-produced single-valued forecast which reflects various run-time modifications (see Section 2 for details); (ii) parsimonious modeling of the serial correlation in ensemble traces of streamflow at successive lead times; (iii) parameter optimization that explicitly uses an ensemble forecast verification metric, i.e., the mean Continuous Ranked Probability Score (CRPS), and (iv) quantitative assessment of the effective lead time in the single-valued forecast as a function of its magnitude and the magnitude of the quantitative precipitation forecast (QPF).

This paper is organized as follows. The problem and its formulation, proposed solution, and technical details, such as the estimation of model parameters, are detailed in Section 2. The study area and data used are described in Section 3. Categorization of flow and QPF, and results and analysis of parameters are detailed in

Section 4. Evaluation of model-generated probabilistic forecasts and visual examination of example probabilistic forecasts are discussed in Section 5 followed by concluding remarks in Section 6.

2. Methodology

In this section, we describe and formulate the problem and detail the proposed solution.

2.1. Problem description

The hydrologic forecasters at the National Weather Service (NWS) River Forecast Centers (RFC) produce streamflow forecasts by running hydrologic models with quantitative precipitation and temperature estimates and forecasts (QPE, QTE, QPF, and QTF). In this forecast process, the forecasters may make various run-time modifications, or MODs, to the input data, model states, model output and, if necessary, model parameters based on real-time hydrologic and hydrometeorological observations and forecasters' knowledge of the basin, the river, the models being used and their tendencies and biases, and the quality of the data used. Currently, the short-term river stage forecast produced by the RFCs is a single-valued prediction of the instantaneous river stage at a 6-h interval nominally for 5 days into the future. It is considered the "best" single-valued prediction in that it reflects what the forecasters consider to be the best available single-valued forcing estimates and forecasts, and the various MODs they may make. Given the above single-valued streamflow forecast, our problem is to produce an ensemble streamflow forecast that is reliable, i.e., probabilistically unbiased, and that captures the skill in the single-valued forecast.

Because the operationally-produced single-valued forecasts already reflect all sources of error in the end-to-end forecast process, it is not possible to separate input and hydrologic errors. One of the most often used operations at the RFCs is "Adjust-Q", which interpolates streamflow or stage between the most recent observation and the single-valued forecast valid at some near-future time. For headwater basins, the effect sought from Adjust-Q on streamflow prediction is similar to that of updating the initial conditions of the rainfall-runoff model by assimilating recent streamflow observations. With Adjust-Q, however, no adjustments can be made to the model initial conditions. The proposed procedure utilizes the most recent real-time streamflow observation as an additional conditioning variable, or predictor, for the Adjust Q-like effect.

Because QPF and QTF are very large sources of error in streamflow prediction, their skill, as well as how the RFC may use them, greatly impacts the skill in the single-valued streamflow forecast and its dependence on lead time. At the NWS Arkansas-Red River Basin River Forecast Center (ABRFC), the area of this study, QPF is by far the most important forcing. As such, we only consider QPF for future forcing input in this work. At ABRFC, QPF is usually input only up to 12 or 24 h into the future and assumed zero thereafter for single-valued streamflow forecasting. The maximum lead time of QPF and QTF that are input to the hydrologic model varies from one RFC to another. Longer-lead QPFs may also be used as necessary, e.g., to produce contingency forecasts. Given the above, we formulate the problem as modeling the conditional probability distribution of observed flows at future time steps given the single-valued streamflow forecasts and QPF valid at those time steps, streamflow observations valid at the recent time steps, and generate ensemble traces from the conditional distribution.

2.2. Problem formulation

As posed above, our problem may be stated in generality as estimating the conditional probability, $\mathrm{Prob}[Q_{o,k+1}, Q_{o,k+2}, \ldots, Q_{o,k+n} | \mathbf{Q_f} = \mathbf{q_f}, \mathbf{Q_o} = \mathbf{q_o}, \mathbf{P} = \mathbf{p}]$, where $\mathbf{Q_o} = \{Q_{o,k}, Q_{o,k-1}, \ldots\}$, $\mathbf{Q_f} = \{Q_{f,k+1}, Q_{f,k+2}, \ldots, Q_{f,k+n}\}$, $\mathbf{P} = \{P_{k+1}, P_{k+2}, \ldots, P_{k+n}\}$. In the above, k denotes the current time step, n denotes the number of 6-h time steps in the forecast horizon, and $Q_{o,k}$, $Q_{f,k}$, and P_k denote the observed streamflow, forecast streamflow and QPF valid at time step k, respectively. For notational brevity, we shorthand the conditioning event $\{\mathbf{Q_f} = \mathbf{q_f}, \mathbf{Q_o} = \mathbf{q_o}, \mathbf{P} = \mathbf{p}\}$ as $\{\mathbf{Q_f}, \mathbf{Q_o}, \mathbf{P}\}$ throughout the rest of this paper. The above conditional probability may be written as:

$$\begin{aligned} &Prob[Q_{o,k+1}, Q_{o,k+2}, \ldots, Q_{o,k+n} | \mathbf{Q_f}, \mathbf{Q_o}, \mathbf{P}] \\ &\quad = Prob[Q_{o,k+n} | Q_{o,k+n-1}, \ldots, Q_{o,k+1}, \mathbf{Q_f}, \mathbf{Q_o}, \mathbf{P}] \\ &\quad \bullet Prob[Q_{o,k+n-1} | Q_{o,k+n-2}, \ldots, Q_{o,k+1}, \mathbf{Q_f}, \mathbf{Q_o}, \mathbf{P}] \\ &\qquad \vdots \\ &\quad \bullet Prob[Q_{o,k+2} | Q_{o,k+1}, \mathbf{Q_f}, \mathbf{Q_o}, \mathbf{P}] \\ &\quad \bullet Prob[Q_{o,k+1} | \mathbf{Q_f}, \mathbf{Q_o}, \mathbf{P}] \end{aligned} \tag{1}$$

Under the assumption that the observed streamflow is Markovian, we may approximate the above as:

$$\begin{aligned} &Prob[Q_{o,k+1}, Q_{o,k+2}, \ldots, Q_{o,k+n} | \mathbf{Q_f}, \mathbf{Q_o}, \mathbf{P}] \\ &\quad \approx Prob[Q_{o,k+n} | Q_{o,k+n-1}, \mathbf{Q_f}, \mathbf{P}] \\ &\quad \bullet Prob[Q_{o,k+n-1} | Q_{o,k+n-2}, \mathbf{Q_f}, \mathbf{P}] \\ &\qquad \vdots \\ &\quad \bullet Prob[Q_{o,k+2} | Q_{o,k+1}, \mathbf{Q_f}, \mathbf{P}] \\ &\quad \bullet Prob[Q_{o,k+1} | Q_{o,k}, \mathbf{Q_f}, \mathbf{P}] \end{aligned} \tag{2}$$

Noting that, in $\mathbf{Q_f}$, one may expect $Q_{f,k+i}$ to be generally the most skillful predictor for $Q_{o,k+i}$, we further approximate the above as:

$$\begin{aligned} &Prob[Q_{o,k+1}, Q_{o,k+2}, \ldots, Q_{o,k+n} | \mathbf{Q_f}, \mathbf{Q_o}, \mathbf{P}] \\ &\quad \approx Prob[Q_{o,k+n} | Q_{o,k+n-1}, Q_{f,k+n}, \mathbf{P}] \\ &\quad \bullet Prob[Q_{o,k+n-1} | Q_{o,k+n-2}, Q_{f,k+n-1}, \mathbf{P}] \\ &\qquad \vdots \\ &\quad \bullet Prob[Q_{o,k+2} | Q_{o,k+1}, Q_{f,k+2}, \mathbf{P}] \\ &\quad \bullet Prob[Q_{o,k+1} | Q_{o,k}, Q_{f,k+1}, \mathbf{P}] \end{aligned} \tag{3}$$

Because precipitation has a mixed distribution and the relationship between precipitation and the rainfall-to-runoff processes are highly nonlinear and time-delayed, we do not consider it practical to directly model the conditional probability distributions in Eq. (3). Instead, we stratify the conditional probabilities in Eq. (3) according to the categorical magnitude of QPF in a lead time-dependent way. In addition to the QPF, the above conditional probability is also stratified according to the magnitude of the forecast streamflow similarly to Seo et al. (2006) to reflect the generally disparate temporal correlation structure between high and low flows. The details on categorization of flow and QPF are given in Section 4.1. Below, we describe how $Prob[Q_{o,k+1} | Q_{o,k} = q_{o,k}, Q_{f,k+1} = q_{f,k+1}, p_l^j \leqslant P_{k+1} < p_u^j]$, where p_l^j and p_u^j denote the lower and upper bounds for the jth category of QPF, respectively, is modeled.

2.3. Proposed solution

The model used to estimate the conditional probability distribution of observed flow given forecast flow in Eq. (3) and technical details such as tail modeling and how model parameters estimated are described in this section.

2.3.1. Model description

We model $Prob[Q_{o,k+1} | Q_{o,k} = q_{o,k}, Q_{f,k+1} = q_{f,k+1}, p_l^j \leqslant P_{k+1} < p_u^j]$ via a combination of probability matching and linear regression in the bivariate normal space. The model used is the first-order autoregressive model with a single exogenous variable, or

ARX(1,1), in the normal space (Box and Jenkins, 1976). It uses the prior observation and model forecasts as predictors and constrains the regression coefficients to sum to unity to achieve unbiasedness (see the kriging literature, e.g., Journel and Huijbregts, 1978). The AR(1) model is based on the assumption that the observed streamflow is Markovian (Krzysztofowicz, 1999, 2002). It implies that the streamflow forecast at one-step ahead in the normal space is dependent on the streamflow at the current time step and independent of streamflow at the preceding time steps. In reality, a higher-order model may be better. Our experience, however, is that additional complexity of the higher-order model is seldom justified given the increased data requirements. As noted, usually only a limited amount of historical operational forecasts is available in practice. As such, the use of the most parsimonious model, such as the one used in this work, is often the only practical choice. There are, however, cases when AR(1), and hence the Markovian assumption, may not be reasonable such as when the streamflow is regulated.

The model has the following form:

$$Z_{o,k+1} = (1 - b_{k+1})Z_{o,k} + b_{k+1}Z_{f,k+1} + E_{k+1} \tag{4}$$

where $Z_{o,k+1}$ and $Z_{o,k}$ denote the NQT-transformed observed flow at time steps $k+1$ and k, respectively, b_{k+1} denotes the regression coefficient at time step $k+1$, $Z_{f,k+1}$ denotes the NQT-transformed forecast flow valid at time step $k+1$, and E_{k+1} denotes the residual error at time step $k+1$. At time step $k+1$, $Z_{o,k}$ corresponds to the actual observed streamflow at time step k, whereas time step $k+2$ onwards $Z_{o,k}$ is the estimated observed value at the preceding time step, i.e., $Z_{o,k+1}$, using Eq. (4).

The above model is similar to that used by Seo et al. (2006). Unlike Seo et al. (2006), however, in which the exogenous variable is the model-simulated streamflow, here it is the operationally-produced single-valued streamflow forecast whose skill depends on a multitude of factors such as the availability and quality of QPF, the RFC's choice of the maximum lead time for QPF, the choice and extent of MODs, etc. As such, the parameter b_{k+1} in Eq. (4) is necessarily lead time-dependent in a rather complex way. Therefore, the one-step transition equation is developed for each lead time and marched forward until the end of the forecast horizon.

Empirical analysis shows that often significant correlation exists between $Z_{o,k}$ and E_{k+1} in Eq. (4), and that the residual error, E_{k+1}, is correlated in time. As expected, the above correlation is quite strong at large lead times where the basin memory and the effects of QPF wear off. This temporal dependence is modeled in this work as first-order autoregressive, or AR(1):

$$E_{k+1} = \frac{\sigma_{E_{k+1}}}{\sigma_{E_k}} \rho(E_{k+1}, E_k)E_k + W_{k+1} \tag{5}$$

where $\sigma_{E_{k+1}}$ and σ_{E_k} denote the standard deviations of E_{k+1} and E_k, respectively, $\rho(E_{k+1}, E_k)$ denotes the serial correlation between E_{k+1} and E_k, and W_{k+1} represents the white noise. It is assumed that E_k and W_{k+1} are statistically independent. Empirical analysis shows that AR(1) is not always satisfactory. A more complex model, however, may not be desirable due to the operational requirement that the procedure must work reasonably well, even with limited data availability for parameter estimation. From Eq. (5), we have:

$$\sigma^2_{W_{K+1}} = (1 - \rho^2(E_{k+1}, E_k))\sigma^2_{E_{k+1}} \tag{6}$$

In reality, the residual error, E_{k+1}, and the most recently observed flow in the normal space, $Z_{o,k}$, are correlated. As such, estimated observed flows will not in general preserve the error variance in the predictand, $Z_{o,k+1}$. In this work, rather than explicitly modeling the dependence between E_{k+1} and $Z_{o,k}$, we scale E_{k+1} in such a way that the error variance in $Z_{o,k+1}$ is preserved. For details, the reader is referred to Appendix A. This approximation is motivated by parsimony and algorithmic simplicity.

The conditional mean (and, similarly, conditional variance) of observed streamflow at future time steps may also be evaluated directly via numerical integration:

$$\begin{aligned} E[Q_{o,k+1}|Q_{o,k}, Q_{f,k+1}, P_{k+1}] &= \int_0^\infty Q_{o,k+1} f(Q_{o,k+1}|Q_{o,k}, Q_{f,k+1}, P_{k+1}) dQ_{o,k+1} \\ &= \int_0^\infty NQT^{-1}(Z_{o,k+1}) f(Z_{o,k+1}|Z_{o,k}, Z_{f,k+1}, P_{k+1}) dZ_{o,k+1} \end{aligned}$$

where $NQT^{-1}()$ denotes the normal quantile inverse-transformation of the variable parenthesized, and the conditional probability density function (PDF) of $Z_{o,k+1}$, $f(Z_{o,k+1}|Z_{o,k}, Z_{f,k+1}, P_{k+1})$, is given by $\mathbf{N}((1 - b_{k+1})\ Z_{o,k} + b_{k+1}\ Z_{f,k+1},\ \sigma^2_{Z_{o,k+1}})$. Eq. (7) is useful for verifying the ensemble results and for producing ensemble mean forecasts that are not subject to sampling errors due possibly to limited ensemble size.

2.3.2. Tail modeling

For Eq. (4), the NQT maps streamflow to the standard normal space via probability matching. We calculate the empirical cumulative distribution functions (CDF) of streamflow using the Weibull plotting position (Wilks, 2011). Because the CDFs are empirical, the deviates of $Z_{o,k+1}$ that exceed the historical minima and maxima will not have corresponding historically observed flow values in the original space. For this reason, some form of extrapolation of empirical CDFs is necessary. Bogner et al. (2012) discuss the difficulties involved in extrapolating CDFs and propose a novel approach. In this study, we used the hyperbolic approximation (Deutsch and Journel, 1998) for the uppermost-tail end of the distribution and linear interpolation for the lowermost-tail end as in Seo et al. (2006). The hyperbolic approximation is very flexible in that one can easily control the fatness of the tail with a single parameter.

2.3.3. Parameter estimation

For parameter estimation, the forecasts and the verifying observations are grouped into predefined categories (see Table 1), for which separate regression models, i.e., Eq. (4), are developed for each lead time. Parameter estimation comprises the following steps for each category and for each lead time:

(1) Calculate the empirical CDF of the observed and forecast streamflows and transform them into standard normal deviates.
(2) Select a limited number of different values of b_{k+1}, and for each b_{k+1}.
(3) Calculate sample mean and variance of E_{k+1}.
(4) Estimate $\sigma^2_{W_{k+1}}$ via Eq. (6) from the sample statistics of the terms in the right hand side of Eq. (6).
(5) Generate a random deviate of W_{k+1} from $\mathbf{N}(0, \sigma^2_{W_{k+1}})$.
(6) Generate a trace of E_{k+1} via Eq. (5), using the random deviate of W_{k+1}.
(7) Generate a trace of $Z_{o,k+1}$ using Eq. (4) and back-transform (i.e., inverse NQT) it into the flow space.
(8) Repeat steps 5 through 7 and generate desired number of ensemble streamflow forecasts.
(9) Generate ensemble streamflow forecasts for all selected limited number of b_{k+1} values, calculate the mean CRPS, and identify the value of b_{k+1} that minimizes the mean CRPS; using this value as the best initial guess, optimize b_{k+1} using Brent's method (Press et al., 1988).

The above steps are repeated marching forward in time until the end of the forecast horizon is reached. Thus, optimal b_{k+1} values

Table 1
Details of categorization of flow and QPF; Q_t^{pm} is the probability-matched single-valued forecast valid at time t; Q_{50} is the climatological median of observed flow; $QPF_{\leqslant 12}$ is the QPF valid for the first 12 h; $QPF_{\leqslant 24}$ is the QPF valid for the first 24 h; $QPF_{>12}$ is the QPF valid for the first 12–48 h of lead time; all QPFs are in mm.

Category	Lead time ⩽ 12 h	Lead time > 12 h
High flow, zero QPF	$Q_t^{pm} \geqslant Q_{50},\ QPF_{\leqslant 24} = 0$	$Q_t^{pm} \geqslant Q_{50},\ QPF_{\leqslant 24} = 0$
High flow, moderate QPF	$Q_t^{pm} \geqslant Q_{50},\ 0 < QPF_{\leqslant 24} < 12.5$	$Q_t^{pm} \geqslant Q_{50},\ 0 < QPF_{\leqslant 24} < 12.5$
High flow, large QPF	$Q_t^{pm} \geqslant Q_{50},\ QPF_{\leqslant 24} \geqslant 12.5$	$Q_t^{pm} \geqslant Q_{50},\ QPF_{\leqslant 24} \geqslant 12.5$
Low flow, zero QPF	$Q_t^{pm} < Q_{50},\ QPF_{\leqslant 12} = 0$	$Q_t^{pm} < Q_{50},\ QPF_{>12} = 0$
Low flow, moderate QPF	$Q_t^{pm} < Q_{50},\ 0 < QPF_{\leqslant 12} < 12.5$	$Q_t^{pm} < Q_{50},\ 0 < QPF_{>12} < 12.5$
Low flow, large QPF	$Q_t^{pm} < Q_{50},\ QPF_{\leqslant 12} \geqslant 12.5$	$Q_t^{pm} < Q_{50},\ QPF_{>12} \geqslant 12.5$

are estimated sequentially for each category for all lead times. Once optimal b_{k+1} and corresponding error statistics are generated, above steps 1 through 8 are used to generate ensemble streamflows.

3. Study area and data used

To evaluate the proposed model, we chose 6 basins in the ABRFC's service area. Fig. 1 shows the location of the study basins. The basins vary considerably in size but experience similar mean annual precipitation. In this paper, we present the results only for three basins as they are representative of the other basins (see Table 2). The single-valued forecasts used are the operational forecasts produced by the ABRFC for public release by the Weather Forecast Offices (WFOs). They are at a 6-h interval for 5 days into the future. The period of record is from February, 1997, to October, 2008. The first 2 years of the archive have forecasts only for 4 days into the future. River forecasts are typically issued once a day between 12Z and 18Z. If there is flooding, the forecasts may be updated at every 6 h. In this study, we used only those historical forecasts that have both forecasts and verifying observations for at least 4.5 days within the 5-day forecast horizon. This criterion yielded an approximately equivalent sample size of 6–8 years in all basins. At ABRFC, the QPF is currently input only for the first 12 h into the future beyond which zero precipitation is assumed. The RFC began this practice on July 16, 2001. Before then, QPF was input for the first 24 h of lead time. The river stage forecasts were converted to discharge forecasts using the current rating curves, i.e., the relation between discharge and stage, used operationally at the RFC. In reality, rating curves may change over time due to, e.g., scouring and deposition of silt. As such, we are neglecting the uncertainties associated with possible changes in the rating curves in the period of record. Due to lack of data, however, this and other observational uncertainties (e.g., instrument, reporting and precision errors) are not considered in this work.

4. Application

Forecasts and verifying observations are grouped into several categories, and the model is applied for each category and for each lead time. In this section, we provide details on the categories and how they are formed, followed by an interpretation of the model parameters.

4.1. Categorization of Flow and QPF

The ARX(1,1) model of Eq. (4) is conditioned categorically on the magnitude of the probability-matched single-valued streamflow forecast and the magnitude of the appropriate QPF (see Table 1). The probability-matched single-valued forecast is obtained by matching the CDF of the forecast flow with that of the verifying observed flow (e.g., Hashino et al., 2007). The probability-matched forecast is used mainly to reduce the impact of forecast biases. In this study, the climatological median (i.e., 50th percentile) of ob-

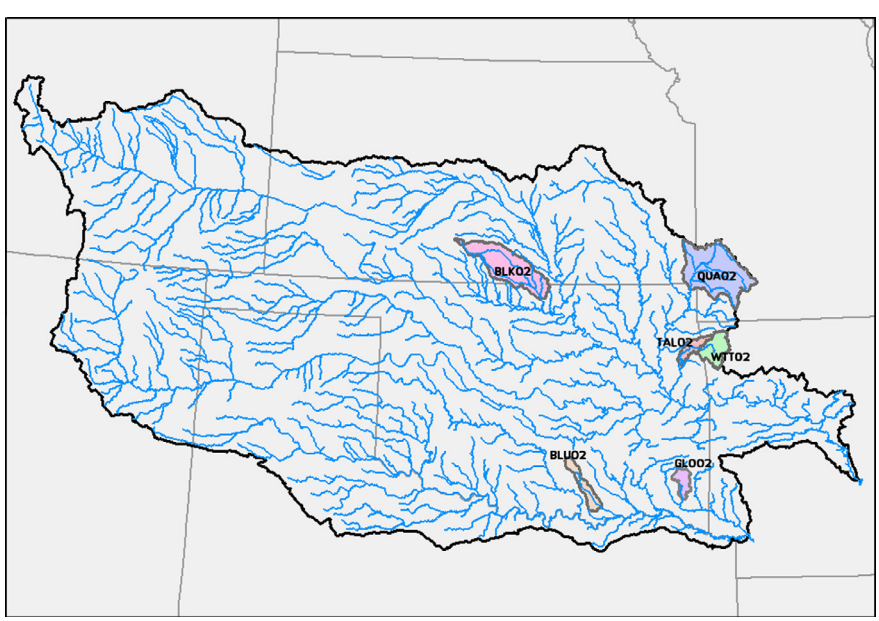

Fig. 1. Map of the study locations and drainage area.

Table 2
Selected three locations in the Arkansas-Red Basin River Forecast (ABRFC)'s service area. Basins are sorted by increased drainage area.

Basin	Drainage area (km^2)	Annual rainfall (mm)	# Of days with held	# Of cross validation periods	Climatological observed flows exceeding below mentioned percentiles (cms)		
					50th	90th	95th
Illinois River near Tahlequah, OK [TALO2]	2484	838	389	7	13	53	88
Illinois River near Watts, OK [WTTO2]	1615	1171	349	8	10	33	54
Blue River near Blue, OK [BLUO2]	1233	1092	354	7	3	11	21

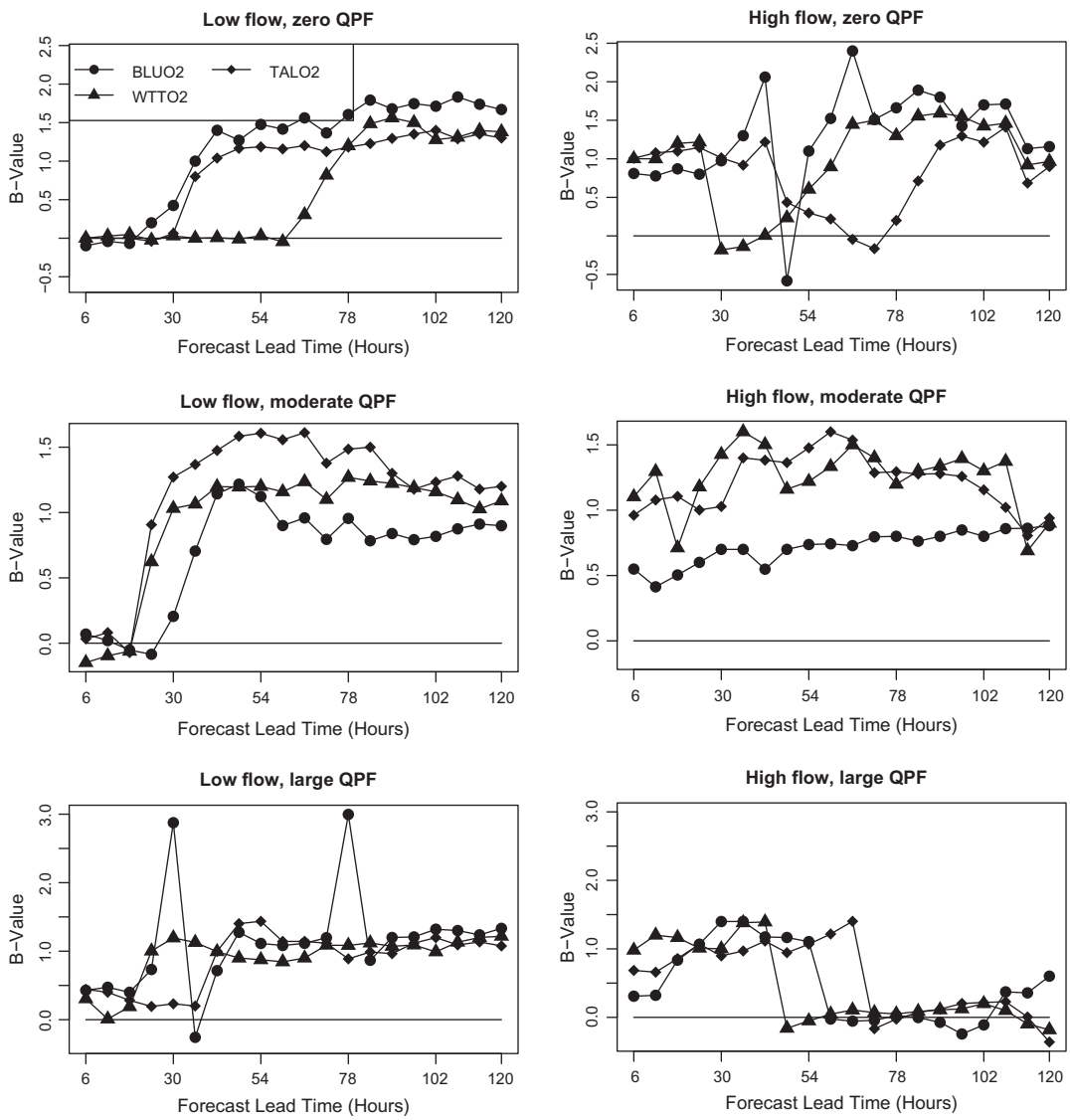

Fig. 2. Optimal regression constants (i.e., *b*-values) for all five different flow regimes and for the entire forecast horizon. Each panel corresponds to a flow regime and each curve corresponds to a specific basin.

served flow was used to categorize the bias-adjusted (i.e. via probability matching) forecast flow as low (i.e., below-median) or high (i.e., above-median). Note that the "low" and "high" designation is only for categorization only and should not be interpreted literally.

For categorization of QPF, we used three different QPFs in this work, i.e., the QPF valid for the first 12 h of lead time, labeled $QPF_{\leqslant 12}$, the QPF valid for the first 24 h of lead time, labeled $QPF_{\leqslant 24}$, and the QPF valid for the first 12–48 h of lead time, labeled $QPF_{>12}$. As the current operational practice of using the first 12 h QPF indicates, $QPF_{\leqslant 12}$ is a skillful predictor of future flow conditions, but not for large flow conditions that may result from significant precipitation beyond the first 12 h. These large flow conditions may be identified by conditioning on QPF beyond the first 12 h. As such, we include $QPF_{\leqslant 12}$ and $QPF_{>12}$ (i.e. QPF valid for 12–48 h ahead) in the categorical predictor set for low flows for the first 12 h of lead time and thereafter, respectively. Thus, for low flows, QPF categorization is lead time-dependent and allows us to use different sets of predictors depending on the lead time. For large flows, we use $QPF_{\leqslant 24}$. The $QPF_{\leqslant 24}$, i.e., the QPF valid for the first 24-h, is a skillful predictor for large flows that may occur due to the QPF outside the first 12-h window. For large flows, QPF valid over the second 24 h of lead time, however, loses skill very quickly. Therefore, only the

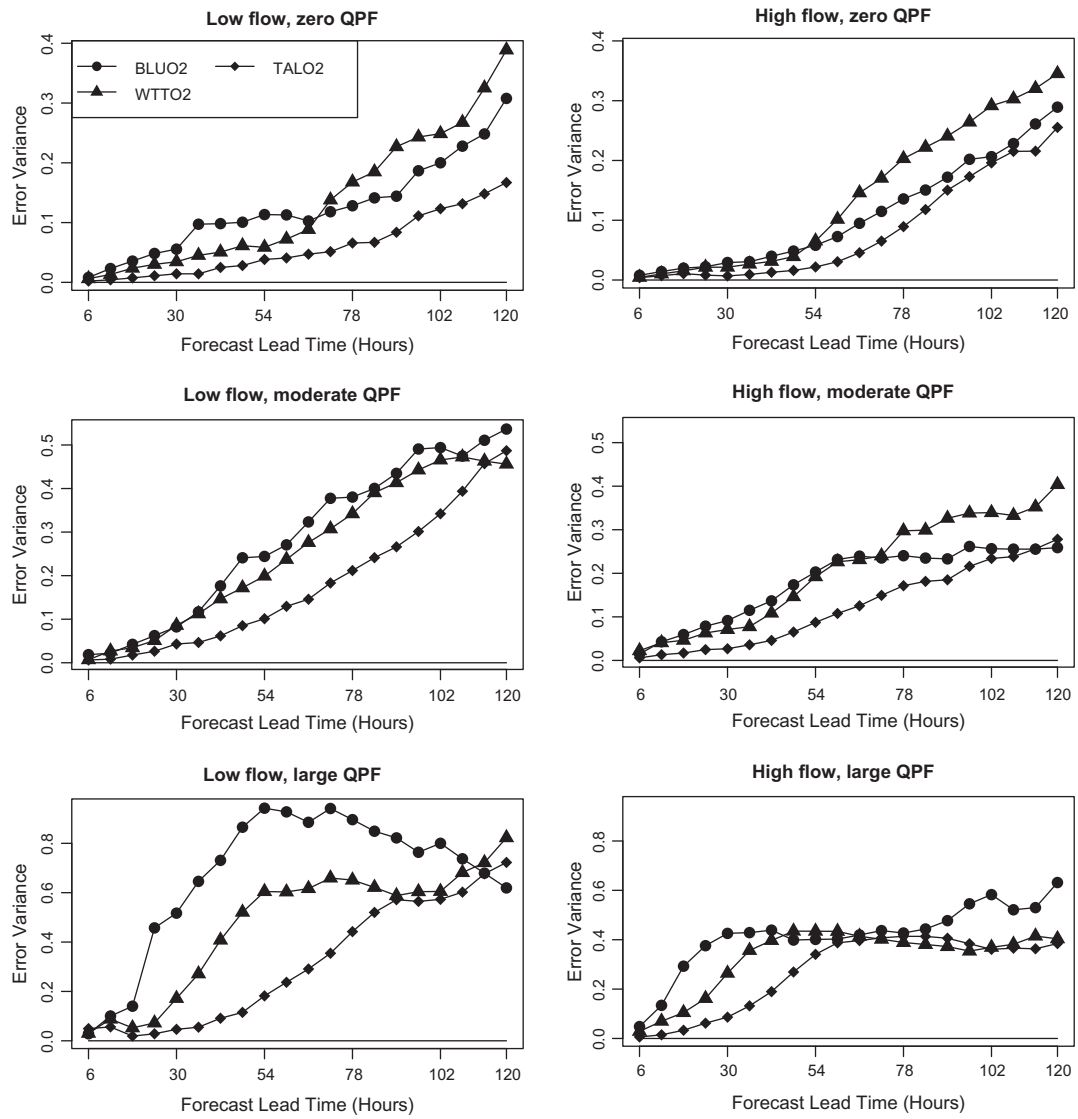

Fig. 3. Error variance for all six different flow regimes and for the entire forecast horizon. Each panel corresponds to a flow regime and each curve corresponds to a specific basin.

$QPF_{\leqslant 24}$ is considered. In this study, QPF amounts of 0 mm, between 0 and 12.7 mm, and above 12.7 mm are used as the thresholds for categorization. Table 1 lists all categories of forecast flow and QPF used in this work for categorical conditioning of $\mathbf{Q_f}$ and $\mathbf{P}$ in Eq. (3).

4.2. Parameter estimation results

Parameter estimation solves for the regression coefficients and the associated sample statistics, such as error variance, and provides information to assist in interpreting the streamflow forecasts. Below, we interpret the regression coefficients and the error variance.

4.2.1. Regression coefficients, b_{k+1}

Figs. 2 and 3 show the optimal b_{k+1} values and the corresponding error variance estimates (i.e., sample variances of E_{k+1} in Eq. (4)), respectively, for the entire 5-day forecast horizon for all six categories for all three basins. Each panel in the figures corresponds to a specific category of the flow and QPF combination, and each curve corresponds to a basin. To aid interpretation of Fig. 2, we examine the qualitative behaviors of b_{k+1} through its least-squares solution in the normal space using Eq. (4) (see also Seo et al., 2006):

$$b_{k+1} = \frac{1 - \rho(Z_{o,k+1}, Z_{o,k}) - \rho(Z_{f,k+1}, Z_{o,k}) + \rho(Z_{f,k+1}, Z_{o,k+1})}{2\{1 - \rho(Z_{f,k+1}, Z_{o,k})\}} \tag{8}$$

where $\rho(,)$ denotes the correlation between the two variables in the argument. If the single-valued forecast has no skill, we have $\rho(Z_{f,k+1}, Z_{o,k+1}) = 0$ and may assume $\rho(Z_{f,k+1}, Z_{o,k}) \approx 0$, with which Eq. (8) is reduced to $b_{k+1} \approx \{1 - \rho(Z_{o,k+1}, Z_{o,k})\}/2$. In very high flow situations, we have $\rho(Z_{o,k+1}, Z_{o,k}) \approx 0$ and hence $b_{k+1} \approx 1/2$, i.e., equal weights are given to the observed flow at time step k and the model forecast for time step $k+1$. In very low flow situations, we have $\rho(Z_{o,k+1}, Z_{o,k}) \approx 1$ and hence $b_{k+1} \approx 0$, i.e., no weight is given to the model forecast. The $b_{k+1} \approx 1/2$ result above is a reflection of the fact that the constrained optimal solution for combining two independent predictions of the same variability with no a priori information is the arithmetic average of the two (Schweppe, 1973). In the other extreme, if the single-valued forecast is perfect, i.e., $\rho(Z_{f,k+1}, Z_{o,k+1}) \approx 1$, we have $b_{k+1} = 1$. In reality, b_{k+1} is optimized under the mean CRPS criterion in the original space and the arguments in the correlation coefficients are the regression-predicted observed flow rather

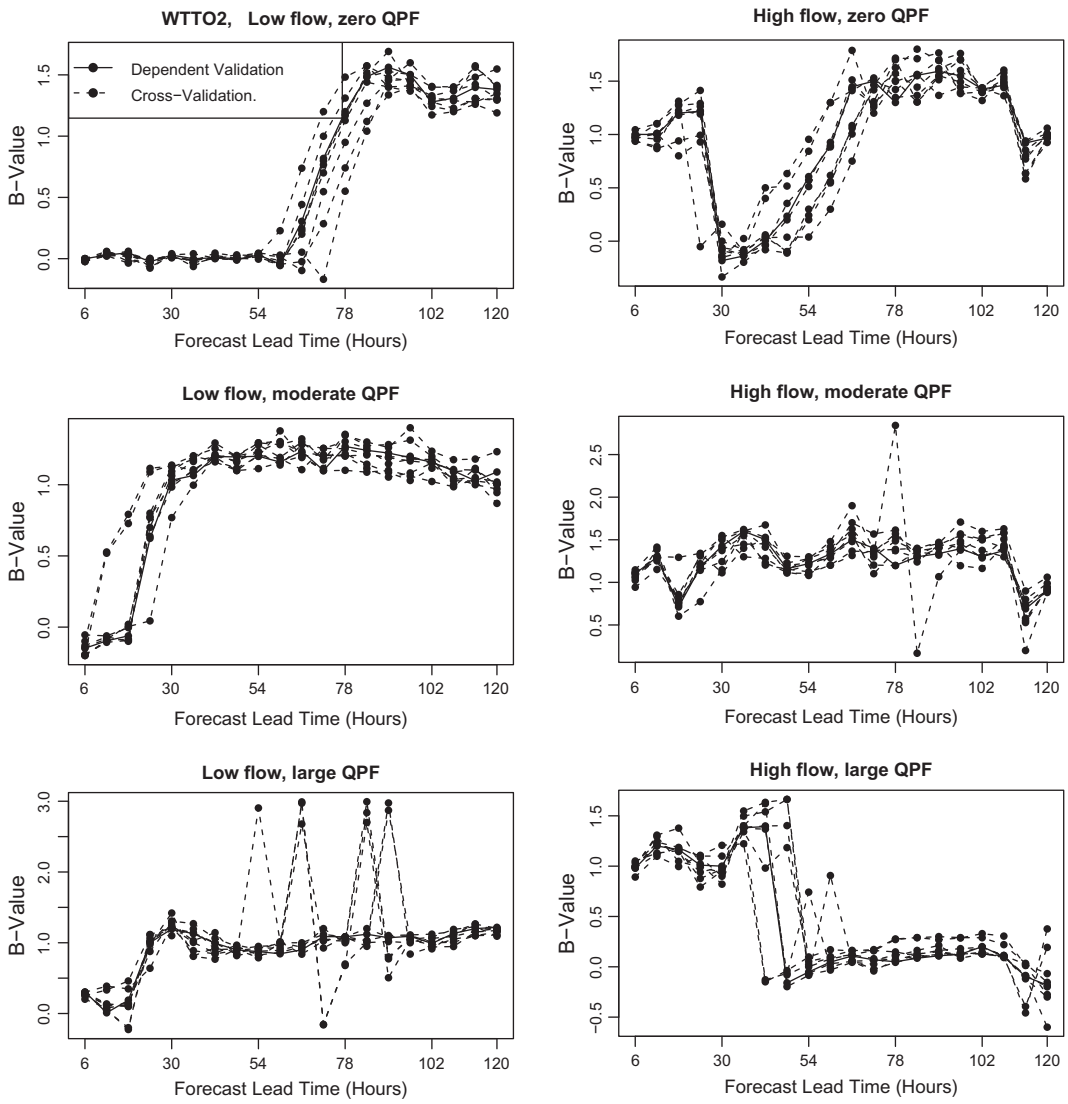

Fig. 4. Optimal regression constants (i.e., *b*-values) estimated in cross validation mode for all six different flow regimes, for the entire forecast horizon for the WTTO2 basin; each dashed line corresponds to a new data set formed via dropping a period of data, whereas the solid line corresponds to entire data set.

than the actual observed flow. As such, one may not expect the above behaviors necessarily to hold. The above observations nevertheless help understand the b_{k+1} patterns shown in Fig. 2 as explained below.

Fig. 2 shows that the optimal values of b_{k+1} range from slightly below zero to greater than unity. Qualitatively, the general patterns for b_{k+1} are similar among the three basins, but with exceptions. For BLUO2, which is the smallest of the three basins shown, there are instances where b_{k+1} spikes up or down rather abruptly. Examination of the mean CRPS as a function of b_{k+1} (not shown) indicates that the objective function is often not very sensitive to b_{k+1} particularly outside of $[0,1]$, and that many different b_{k+1} values over a wide range may yield ensembles of comparable quality. Given the first of the two observations above, we might have constrained the feasible region of b_{k+1} to be $[0,1]$ (see Gneiting et al., 2005, for example).

The following observations may be made in Fig. 2. For forecasts in the below-median flow conditions (left panels), the b_{k+1} values are consistently small for the first few lead times and large for longer lead times. Small values of b_{k+1} indicate a relatively large contribution from the recently observed flow in the predictand due to increased persistence in the low flow (i.e. below median) conditions. As expected, the contribution from persistence is the strongest for forecasts in the below-median flow conditions with zero QPF (upper left). With increasing QPF, however, persistence is either limited only to the first few lead times (middle left) or nearly absent (lower left). Large values of b_{k+1} indicate larger contributions from the operationally-produced single-valued forecast; they are seen at long lead times in the below-median flow conditions (left panels) where the influence of persistence and the effect of basin memory diminish. For forecasts in the above-median flow conditions (right panels), the values of b_{k+1} are consistently close to unity for the first few lead times. The b_{k+1} values for BLUO2, which has the smallest drainage area of the three basins, are generally smaller due presumably to lower skill in the operationally-produced single-valued forecast and shorter basin memory. The large b_{k+1} values for high flow (i.e. above median) conditions indicate skill in the single-valued forecast. It is interesting to note that, for forecasts in the high-flow conditions with zero QPF (upper right), b_{k+1} decreases to rather small values at the intermediate lead times, due presumably to persistence in basin memory in the absence of significant precipitation forcing. The sudden drop

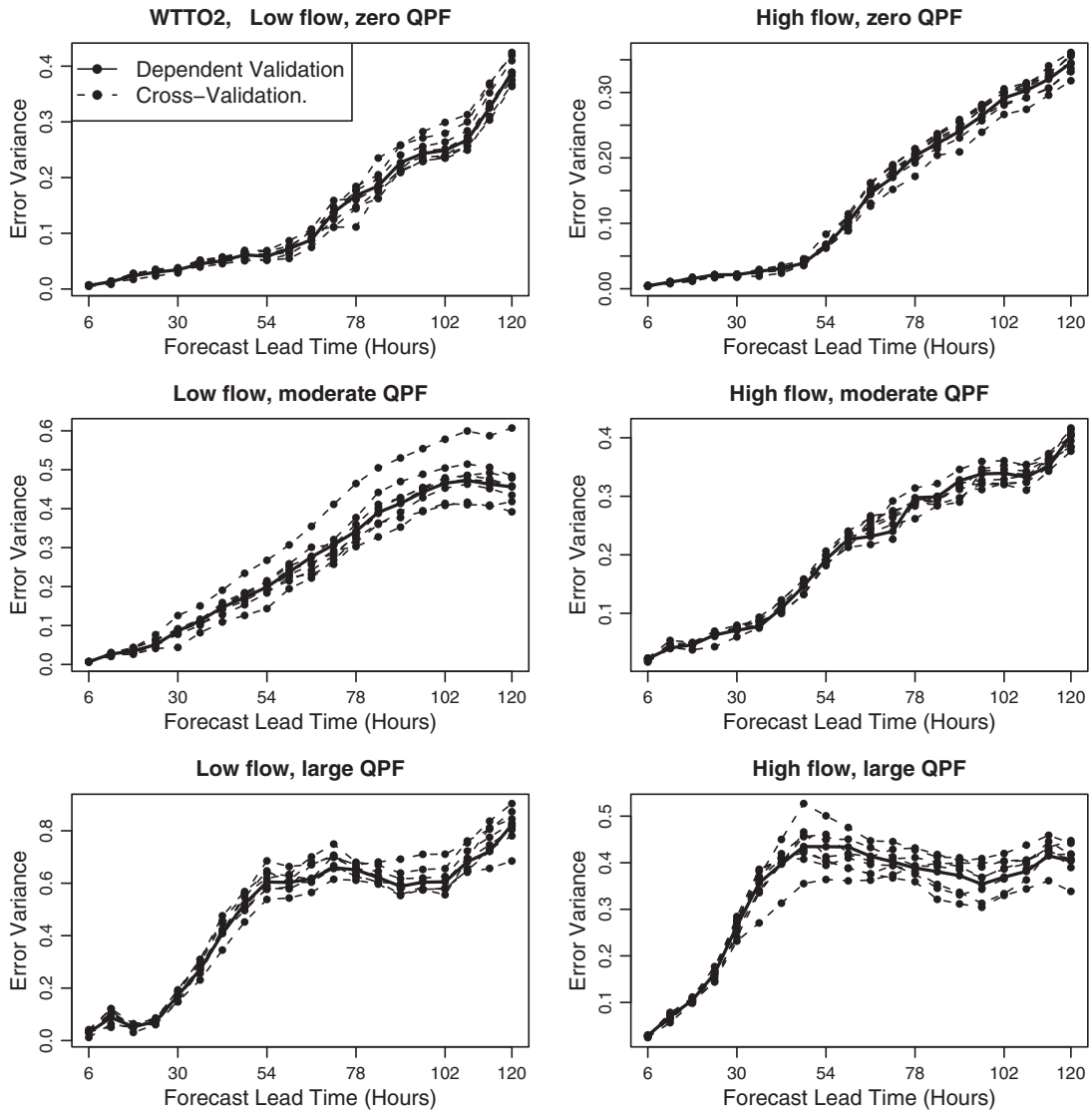

Fig. 5. Error variance estimated in cross validation mode for all six different flow regimes, for the entire forecast horizon for the WTTO2 basin; each curve corresponds to a new data set formed via dropping a period of data, whereas the solid line corresponds to entire data set.

in the b_{k+1} values for the forecasts in the high flow conditions with large QPF (lower right) is due to the lack of skill in the operationally-produced single-valued streamflow forecast at those lead times where the zero QPF, assumed beyond the first 12–24 h, takes effect. For the high flow conditions with moderate QPF (middle right), the b_{k+1} values remain large and steady for the entire forecast horizon, an indication that, for this category, the operationally-produced single-valued forecast provides larger skill than the observed flow at the forecast time except for the smallest basin, BLUO2.

4.2.2. Sample error variance, $\sigma^2_{\tilde{E}_{k+1}}$

Fig. 3 shows the error variance, $\sigma^2_{\tilde{E}_{k+1}}$, in the normal space for all basins and categories. It reflects the normalized uncertainty associated with the predictand. Note that, due to sampling uncertainty, the curves in Fig. 3 are not strictly monotonically non-decreasing. As one would expect, the zero QPF categories show the largest predictability; the $\sigma^2_{\tilde{E}_{k+1}}$ values continue to increase with increasing lead time. As the QPF increases, however, the predictability decreases (middle and bottom panels). For the large QPF categories, the $\sigma^2_{\tilde{E}_{k+1}}$ values plateau well before the nominal maximum forecast lead time of 5 days is reached. Fig. 3 hence provides categorical assessment of the effective lead time of the operationally-produced single-valued forecast. Fig. 3 also shows clearly that predictability increases with basin size and hence the memory of the initial conditions and the relative skill in the QPF. For flood forecasting, the category of high flow and large QPF is the most important; the lower right panel shows that the effective lead times of the operationally-produced single-valued forecast for this category are approximately 1, 1.5 and 2 days for BLUO2, WTTO2 and TALO2, respectively.

4.2.3. Parametric uncertainty

Fig. 4 shows the b_{k+1} estimates for WTTO2 obtained from the cross validation experiment in which different periods are withheld (see Table 2 for the number of days withheld). The figure is useful in assessing parametric uncertainty in b_{k+1} due to sampling. For the most part, the b_{k+1} estimates from the entire data set are enveloped by the cluster of estimates from cross validation. At certain lead times, however, the b_{k+1} estimates from the cross validation experiment depart greatly from the cluster. As noted above, these large departures in b_{k+1} result from rather small variations in the objective function value and have little impact on the quality of the ensembles. Fig. 5 shows the error variance, $\sigma^2_{\tilde{E}_{k+1}}$, values asso-

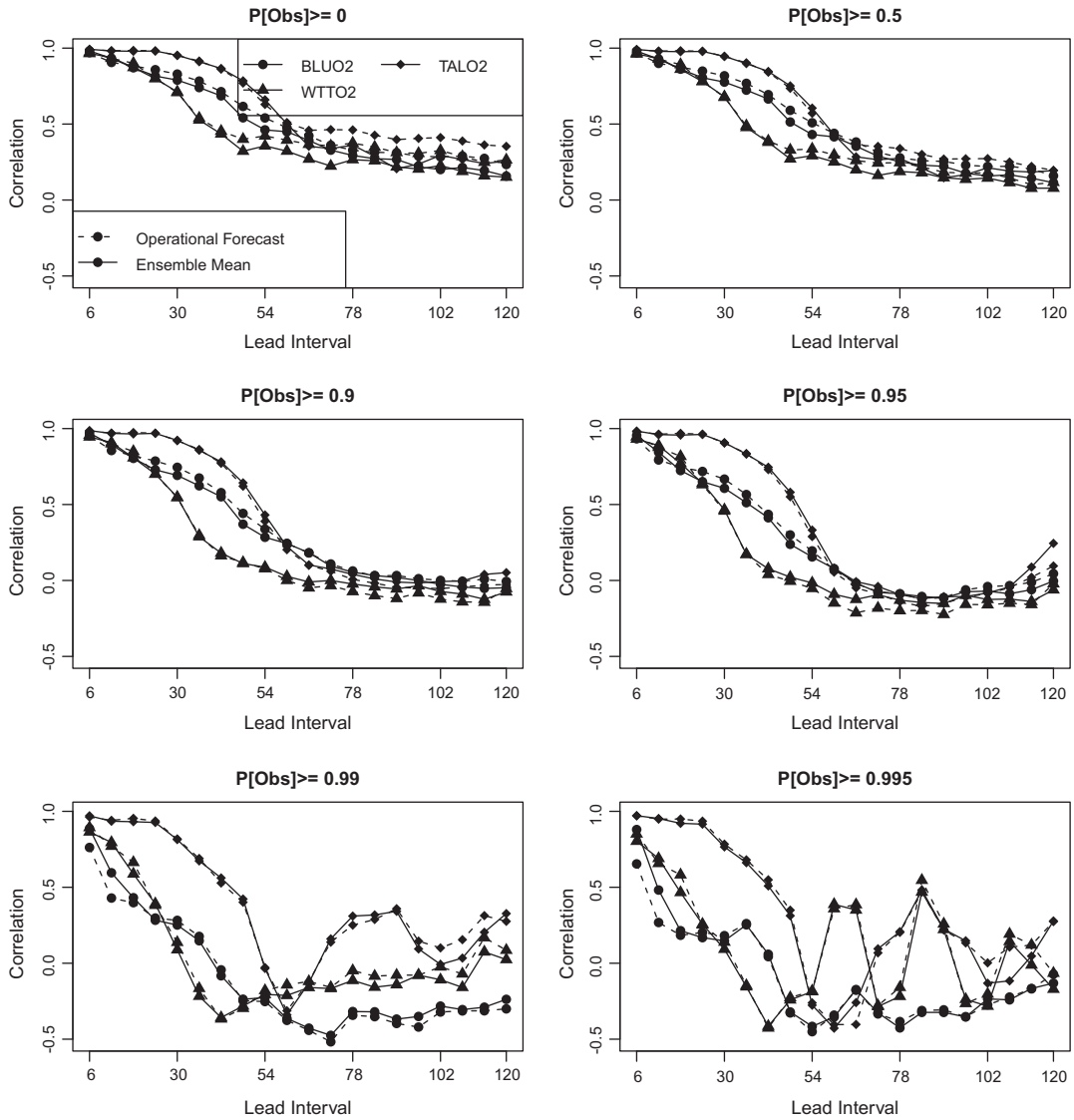

Fig. 6. Correlation calculated for the ensemble mean (calculated from ensemble hindcasts produced in the cross validation mode) and the single-valued operational forecast. Each marker corresponds to a location. Solid line corresponds to ensemble mean and dashed line corresponds to single-valued operational forecast.

ciated with the b_{k+1} estimates shown in Fig. 4. Note that, compared to the b_{k+1} estimates, the error variance values are much more stable, but that, for certain categories (middle left and lower right, in particular), significant sampling effects exist in the magnitude of the uncertainty modeled. The above results suggest that the procedure can benefit from a larger sample size.

5. Performance evaluation

To evaluate the performance of the model, two types of experiments, dependent validation and cross validation, were carried out. The first experiment used all available data for parameter estimation and ensemble generation. In the second experiment, part of the data was withheld for parameter estimation and ensembles were generated only for the withheld period (see Table 2 for the number of days withheld). This was repeated until the union of all withheld periods exhausted the entire period of record. The experiment resulted in 7 cross-validated periods for BLUO2 and TALO2, and 8 for WTTO2 (see Table 2). The number of ensemble members generated is 1000 throughout this work; the choice of the large number is to reduce sampling uncertainties. Ensemble forecasts are also referred to herein as ensemble hindcasts because the ensemble forecasts are generated in a hindcasting mode. The model performance was evaluated by assessing different aspects of quality of ensemble hindcasts that generated in both dependent and cross-validation modes. Various aspects of forecast quality are discussed via calculating both, single-valued and ensemble forecast based verification metrics and visual examination of hydrographs of ensemble forecasts for different hydrologic conditions.

5.1. Single-valued forecast verification

As mentioned in Section 1, one of the objectives of HMOS is to produce streamflow ensemble forecasts that, in addition to providing probabilistic forecasts, improve over the operationally-produced single-valued forecasts in some single-valued sense as well. To that end, we compare the correlation coefficient (CORR) and mean error (ME) of the HMOS ensemble mean hindcasts to those of the operationally-produced single-valued forecasts for all lead times by conditioning on the verifying observed flow of varying magnitude, expressed as different percentiles of the climatological observed flows. The ME is calculated as the mean of the difference between the forecast and the observation. Therefore, a

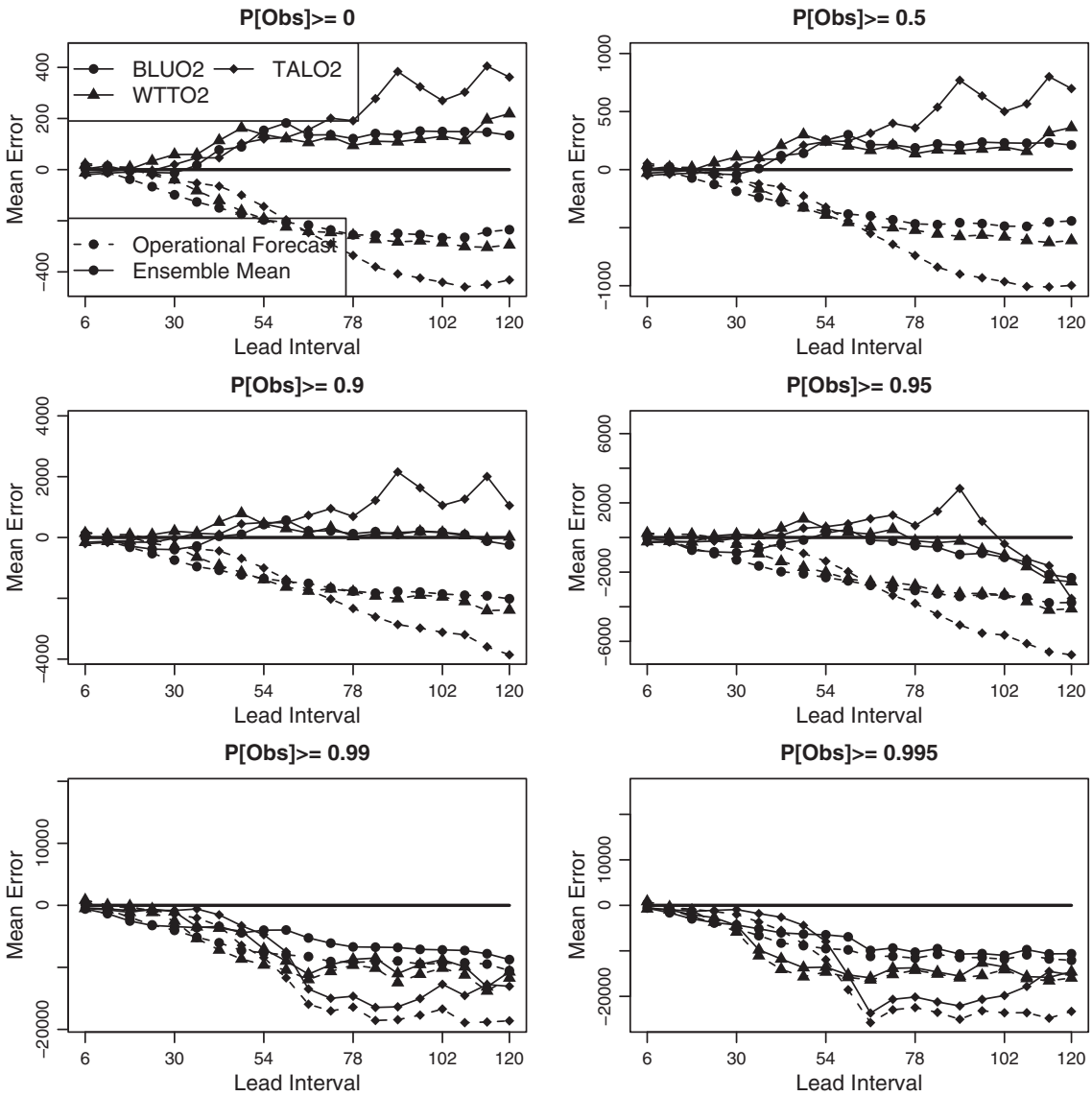

Fig. 7. Mean error for the ensemble mean (calculated from ensemble hindcasts produced in the cross validation mode) and the single-valued operational forecast. Each marker corresponds to a location. Solid line corresponds to ensemble mean and dashed line corresponds to single-valued operational forecast.

positive or negative ME indicates over- or under-forecasting in the mean sense, respectively. Figs. 6 and 7 show the CORR and ME, respectively, for all basins for different ranges of observed flow. The statistics are calculated for all hindcasts from the cross-validation experiment conditioned on the verifying observed flow exceeding the 50th, 90th and 95th percentiles of the climatological observed flow. Fig. 6 shows that the CORR values for the HMOS-generated ensemble mean forecast is comparable to those of the operationally-produced single-valued forecast for all basins and for all categories. We may summarize Fig. 7 as follows. The negative bias in the operational forecast is small for the first few lead times, but more pronounced at long lead times for large observed flow. The very large low bias in the operationally-produced single-valued forecast at long lead times is due to the zero-QPF assumption beyond the first 12–24 h of lead time. The bias in the HMOS ensemble mean is substantially smaller, particularly for large flows. The HMOS-generated ensemble mean forecasts, however, tend to have a high bias unconditionally (upper left) and for all verifying observed flows exceeding the climatological median (upper right). Note also that, while the ensemble mean forecasts substantially reduce the low bias for high flows (bottom panels), they are still biased low significantly. It illustrates the difficulty of correcting large biases completely.

5.2. Ensemble forecast verification

The quality of ensemble hindcasts is assessed by calculating a variety of ensemble verification metrics using the Ensemble Verification System (EVS) developed by the NWS Office of Hydrologic Development (OHD) (Brown et al., 2010). Detailed description of the metrics may be found in Jolliffe and Stephenson (2003) and Wilks (2011). Here we present only the mean CRPS, reliability diagram and Relative Operating Characteristic (ROC). The mean CRPS is a summary performance measure that considers the entire distribution of an ensemble forecast (Hersbach, 2000). It reduces to the mean absolute error (MAE) for deterministic forecasts, and hence allows comparison of the HMOS-generated ensemble hindcasts with the operationally-produced single-valued forecasts. Fig. 8 shows the mean CRPS of the ensemble hindcasts and the MAE of the operationally-produced single-valued forecast. Increasing mean CRPS and MAE with increasing lead time indicates decreasing skill with increasing lead time. Note that the mean CRPS

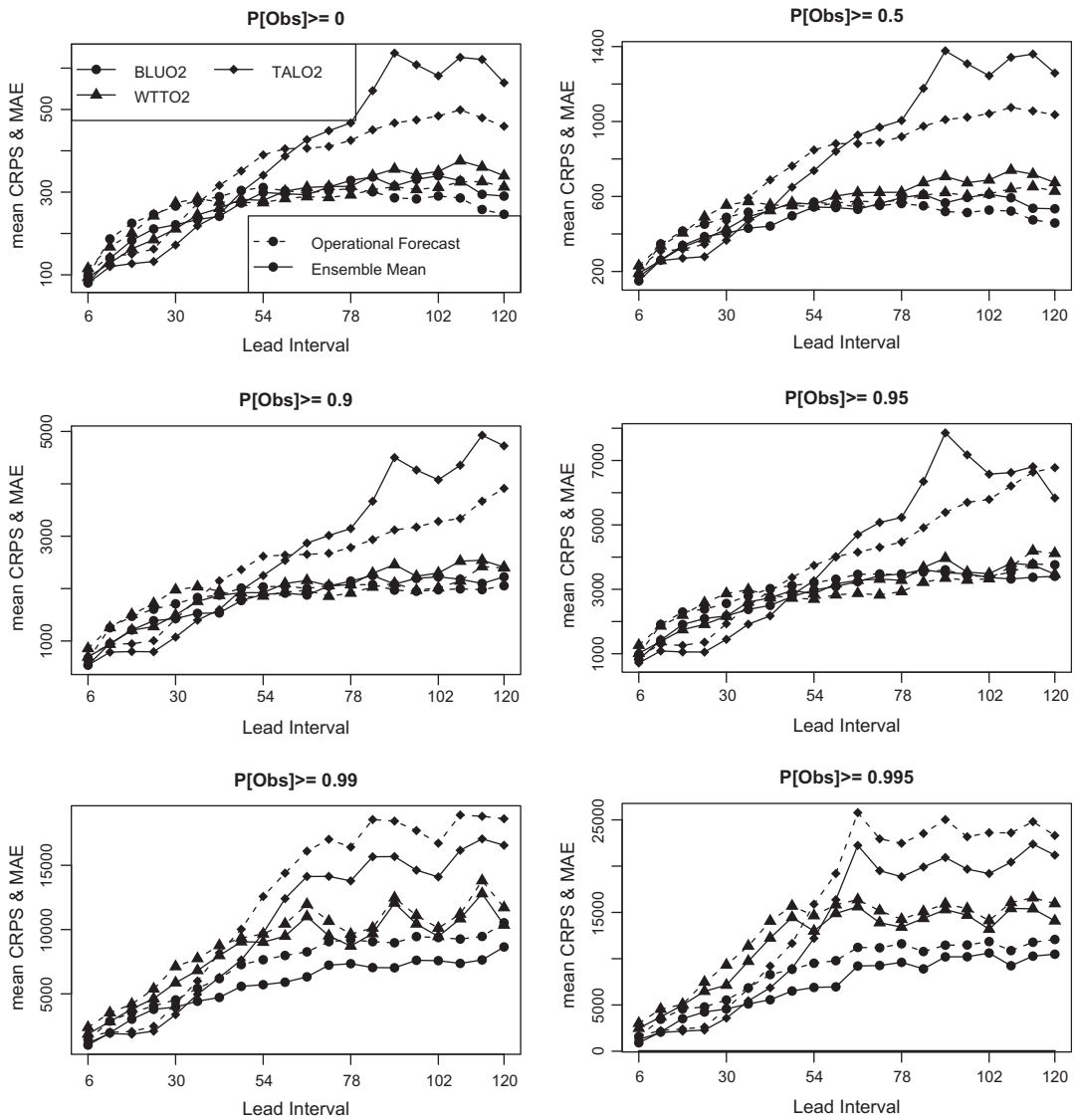

Fig. 8. Mean CRPS for the ensemble hindcasts produced in the cross validation mode (solid line), and Mean Absolute Error for the single-valued operational forecast (dashed line). Each marker corresponds to a location.

of the HMOS-generated ensemble hindcasts is generally smaller than the MAE of the operationally-produced single-valued forecasts at short lead times. At large lead times, however, the mean CRPS of the ensemble hindcasts is generally larger than the MAE of the single-valued forecasts unconditionally and for verifying observations greater than the 95th percentile of the climatological observed flow. For verifying observations of 99th percentile and greater, on other hand, the mean CRPS of the ensemble hindcasts is generally smaller than the MAE of the single-valued forecasts at large lead times. It may seem counterintuitive that the mean CRPS of the ensemble hindcasts is not consistently smaller than the MAE of the single-valued forecasts. The above result arises because reducing the ME tends to increase the mean square error (MSE). For large flows, however, the single-valued forecasts are very much biased on the low side and hence reducing the ME reduces the MSE as well.

Fig. 9 shows the reliability diagrams for ensemble hindcasts from dependent (upper panel) and cross (lower panel) validation for BLUO2 for thresholds of 50th (left panel) and 95th (right panel) percentiles of climatological observed flow. Reliability diagrams measure conditional unbiasedness of the forecast probabilities relative to the observed frequencies (Wilks, 2011; Jolliffe and Stephenson, 2003). The diagrams indicate that ensemble hindcasts are reliable at the 50th-percentile threshold for all 5 days into the future. The reliability diagrams at the 95th-percentile threshold indicate that the ensembles hindcasts are generally reliable out to Day 3. Lack of high forecast probabilities >0.8 beyond Day 3 indicate forecasts with insufficiently high probabilities. This is because probabilities lose sharpness with forecast lead time and mimic climatology. Note that, at high thresholds, the reliability diagrams are more likely to be subject to significant sampling errors due to small sample size. For both thresholds, the histograms of predicted probability indicate sharp ensemble forecasts (i.e., 'L' or 'U' shaped). The above suggest that the probabilistic forecasts are not very reliable at large lead times and for high thresholds due to very large biases in the single-valued forecasts. Recall that, beyond the first 48 h, no precipitation was assumed for QPFs. The resulting bias in streamflow forecasts is very large, which makes statistical correction based on linear regression in the normal space very difficult. Note that, at these long lead times, not only the single-valued forecasts are severely biased but they are also without skill, which adds to the difficulty. Similar reliability diagrams were observed between

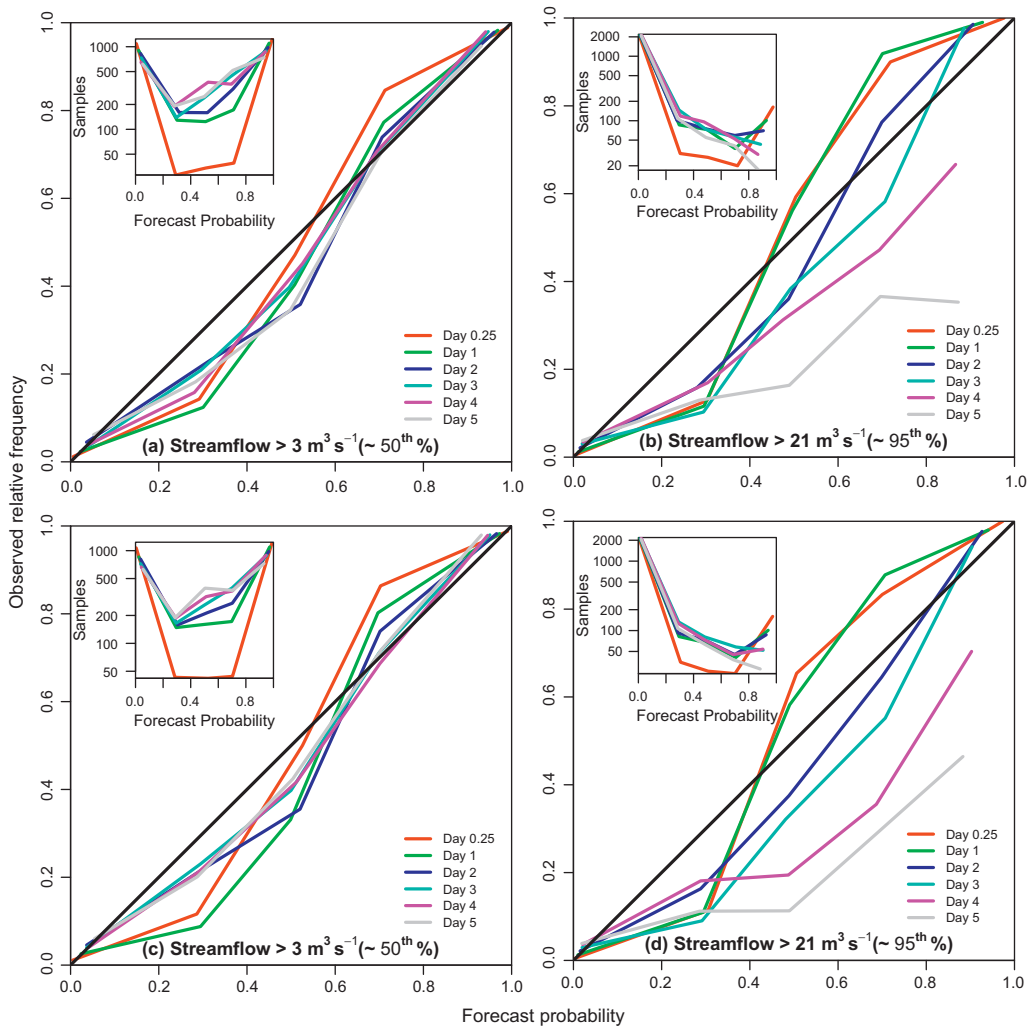

Fig. 9. Reliability plots for 50th and 95th percents of observed flows, for 6-, 24-, 48-, 96- and 120-h forecast lead time, for BLUO2. Top and bottom row corresponds to dependent and cross validation periods, respectively.

the two different validation modes, an indication that the period of record used in this work is reasonable.

The mild "S shapedness" of the diagrams shown in Fig. 9 is due to the assumption in the regression model used in this work that the mean and variance of the residual in the normal space are the same over all ranges of the residual. In reality, however, these statistics are dependent on the magnitude of the residual. By using conditional mean and variance, rather than unconditional mean and variance, it is possible to remove the S shapedness. To reliably estimate the conditional statistics, however, one needs a much larger data set which, at least in is not currently available in NWS.

The ROC connects the (false alarm rate [FAR], hit rate [HR]) combinations associated with different levels of forecast probability for the user-defined event. The ROC curve measures discriminatory skill, i.e., the forecast's ability to discriminate between two different observed outcomes. In the context of short-term hydrologic forecasting, the ability to distinguish flood events from non-flood events, and critical stage levels such as action stage from regular stages are of the greatest interest. As such, we define two events of the observed flow exceeding the 95th and the 99.5th percentiles of the climatological observed flow for the ROC analysis. We also calculate the HR and FAR for the single-valued forecast for both events for comparison. Because, by definition, the discharge levels do not approach the above two thresholds most of the times, the FAR of almost any forecast would be very small, thereby rendering comparative evaluation difficult. To reduce the dominant effect of correctly forecasting low flows, we calculated the above statistics considering only those data points for which the verifying observed flow exceeds the 50th percentile of the climatological observed flow. The resulting conditional statistics of ROC, HR and FAR then assess discrimination under a tougher criterion. Fig. 10 shows that, overall, the ensemble forecasts exhibit good discrimination out to Day 2 even for large events. Comparison of the (HR, FAR) positions of the single-valued forecast with those of the ROC curves suggests better or comparable discrimination by the ensemble forecast in that the ROC curves encompass or match the corresponding (HR, FAR) positions for all lead times and for both thresholds. The ensemble hindcasts from cross validation (lower panel) exhibit slightly and noticeably decreased event discrimination than those from calibration (upper panel) for the 95th and 99.5th thresholds, respectively, an indication that the performance may be improved by increasing the sample size.

5.3. Visual examination of hydrographs

The operations concept for the HMOS forecast is to visualize it alongside the single-valued forecast to provide a measure of the magnitude and direction of the bias that may be present in the single-valued forecast and a sense of the uncertainty bound in the bias-corrected forecast. Fig. 11 illustrates such a concept. Examples

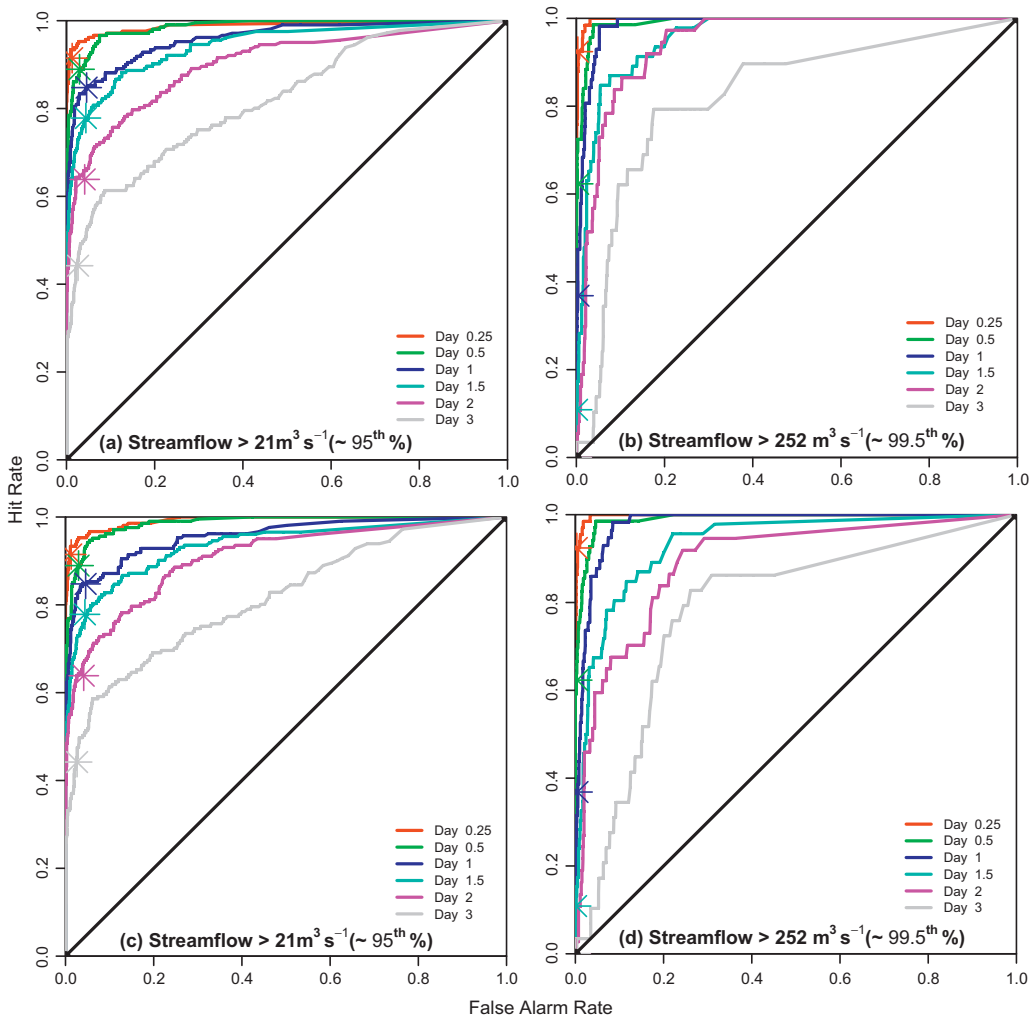

Fig. 10. Relative Operating Characteristic (ROC) curves for 95th and 99.5th percents of observed flows, for 6-, 12-, 24-, 36-, 48- and 72-h forecast lead time, for BLUO2. Symbols marked with '*' correspond to the single-valued operational forecast. The ROC curves calculated using the data of which observed flows are greater than 50th percentile of observed flows. Top and bottom row corresponds to dependent and cross validation periods, respectively.

of HMOS-produced ensemble forecasts for different hydrologic conditions are presented in Fig. 11. The single-valued operational forecast (red) and corresponding verifying observed flow (green), ensemble members (black), and ensemble mean (blue) are plotted on the same panels. The future streamflow and precipitation conditions and the corresponding category (see Table 1) are denoted for each lead time in color-filled circles at the top of each panel. In general, the ensemble members encompass the verifying observation reasonably tightly and are able to reproduce the temporal patterns of variability present in the observed flows. They indicate that the modeling of the temporal dependence of the forecast error and the approximations in sequential ensemble generation are reasonable. The ensemble members are tight for the first few lead times and relatively wider for long forecast lead times, a reflection of the increased forecast uncertainty with increasing lead time. The figure also shows dependence of the ensemble spread on the future conditions. Note, for example, that the ensemble members for low flow forecasts have a smaller spread than those for high flow forecasts.

6. Conclusions

While the source-specific approach should generally be preferred in hydrologic ensemble forecasting, it requires modeling of all significant sources of uncertainty and hence is a rather expensive undertaking. Many agencies for operational hydrologic forecasting have invested greatly in single-valued forecast systems, for which considerable institutional experience and expertise may exist. As a transitional approach, bridging the above two, one may consider producing probabilistic forecasts from single-valued forecasts. To that end, we present a statistical procedure for generating short-term ensemble streamflow forecasts from single-valued, or deterministic, forecasts produced operationally by the U.S. National Weather Service (NWS) River Forecast Centers (RFC). The procedure, referred to as Hydrologic Model Output Statistics (HMOS), generates ensemble traces of streamflow from a parsimonious approximation of the conditional multivariate probability distribution of future streamflow given the single-valued streamflow forecast, the quantitative precipitation forecast (QPF), and the most recent streamflow observation. The resulting ensemble forecast provides an estimate of the predictive uncertainty associated with the single-valued forecast to support risk-based decision making by the forecasters and by the users of the forecast products, such as emergency managers. As a by-product, the technique provides quantitative assessment of the effective lead time of the single-valued forecast as a function of the magnitude of the forecast flow and that of the QPF.

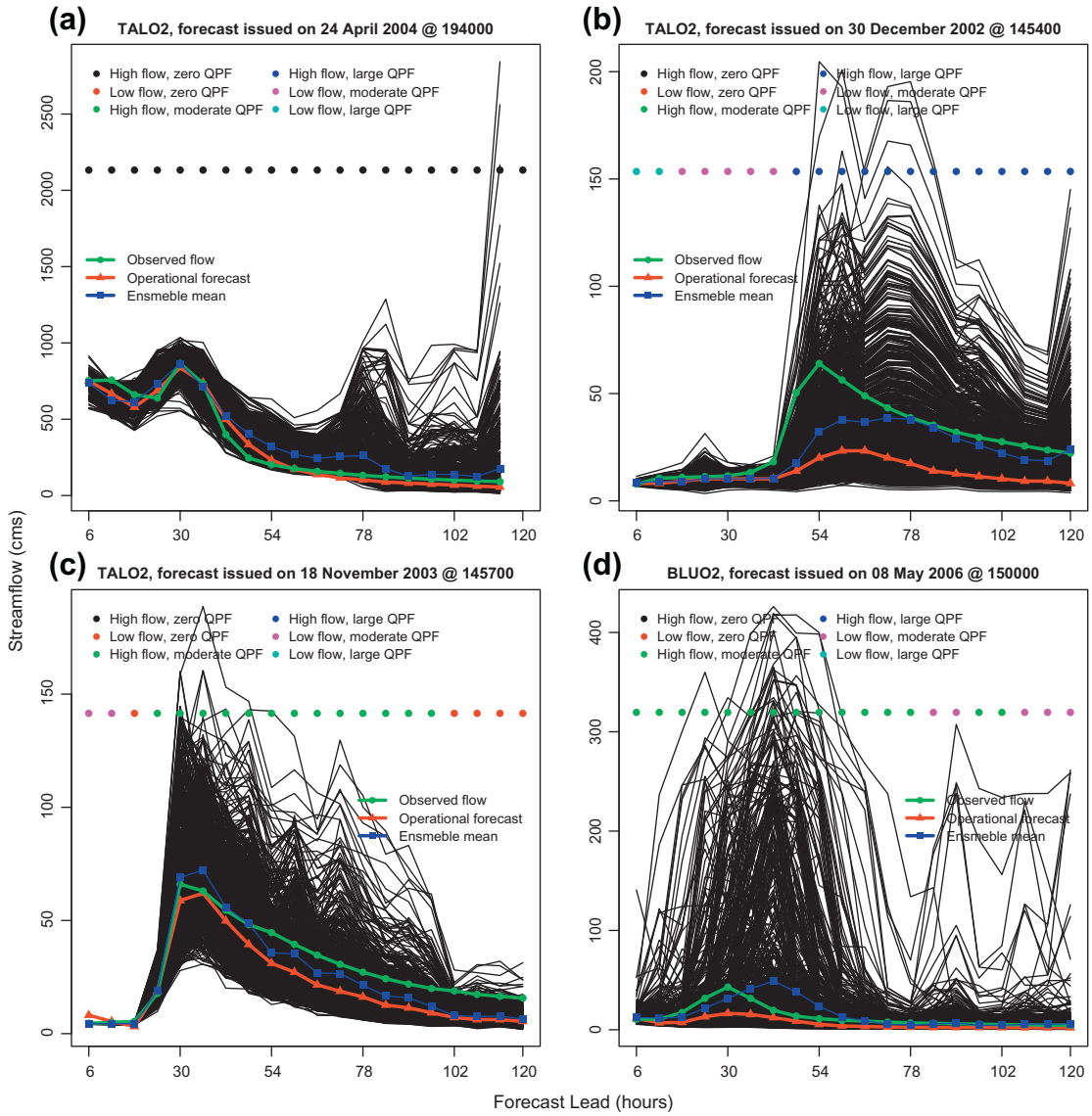

Fig. 11. Hydrographs of probabilistic forecast for different hydrologic conditions: (a) forecast issued time 24 April 2004 for TALO2; (b) forecast issued time 30 December 2002 for TALO2; (c) forecast issued time for 18 November 2003 TALO2; (d) forecast issued time 8 May 2006 for BLUO2. The single-valued operational forecast (red) and corresponding verifying observed flow (green), ensemble members (black), and ensemble mean (blue) are plotted on the same figure. The future streamflow and precipitation conditions and corresponding category for each lead time are denoted in colored filled circles at the top of each figure.

For parameter estimation and evaluation of the procedure, we used a multiyear archive of single-valued river stage forecasts for several basins in Oklahoma, produced operationally by the NWS Arkansas-Red River Basin River Forecast Center (ABRFC) in Tulsa, Oklahoma. To evaluate the procedure, we carried out hindcasting experiments in both dependent and cross-validation modes. The results indicate that the short-term streamflow ensemble forecasts generated from the proposed procedure are generally reliable within the effective lead time of the single-valued forecast and capture the skill in the single-valued forecast very well. For smaller basins, however, the effective lead time for flood forecasting is rather limited due to shorter basin memory and reduced skill in the single-valued QPF. Recommendations for future research include improving classification of data, dealing with heteroscedasticity (e.g., Montanari and Brath, 2004; Montanari and Grossi, 2008; Coccia and Todini, 2011; Bogner and Pappenberger, 2011), and tail modeling (Bogner et al., 2012). It is acknowledged, however, that limited sample size, which is the norm with archived operational forecasts, may inhibit some of these objectives.

As implemented here, the HMOS method uses three predictors, i.e., the single-valued operational forecast, the QPF, and the most recently observed streamflow. However, other skillful predictors might be considered, depending on the basin characteristics. For example, basins in cold regions that are driven by snowmelt may be augmented by QTF. The proposed method is a generic statistical method insofar as additional predictors may be considered in other cases. However, any additional predictors should be selected carefully, because they increase the number of parameters to estimate and hence the data requirements. The conditional distribution of observed flows can be estimated either by developing a first-order autoregressive model with multiple exogenous inputs, or by applying ARX(1,1) model on categories that are formed based on a new set of predictors. If the relation between streamflow and the additional variables is complex and nonlinear then, for the same reasons that QPF is not directly included in the regression, the conditioning variables may be categorical rather than continuous.

The current version of the HMOS prototype is a self-contained FORTRAN executable wrapped with a JAVA model adapter to oper-

ate within the Community Hydrologic Prediction System (CHPS). The software is coded specifically to the variables and classification schemes used in this study. However, a degree of modularity should allow the software to be integrated into another system, and also provides flexibility to modify the code to incorporate additional variables. The software generates ensemble of streamflows provided the parameters and time series of QPF and streamflow in the prescribed format. So far, the prototype software has been successfully piloted at the ABRFC in CHPS.

Acknowledgements

This work is supported by the Advanced Hydrologic Prediction Service (AHPS) program of the National Weather Service (NWS) and by the Climate Predictions Program for the Americas (CPPA) of the Climate Program Office (CPO), both of the National Oceanic and Atmospheric Administration (NOAA). The support for the second author was provided in part by the agreement between LEN Technologies, Inc., and the University of Texas at Arlington under an order for services (Contract Number GS35F0575T) from the NOAA/National Weather Service Hydrologic Operations and Services. These supports are gratefully acknowledged. The first author thanks the National Research Council fellowship for the part of the support, James Paul at the ABRFC for the support in the data acquisition and Mary Mullusky NWS Hydrologic Services Division for inception of the HMOS approach. The authors acknowledge two anonymous reviewers for their comments.

Appendix A

The dependence between the error and recent observed values in the normal space is accounted as below mentioned.

$$\sigma^2_{Z_{o,k+1}} = (1 - b_{k+1})^2 \sigma^2_{Z_{o,k}} + f_c \sigma^2_{E_{k+1}} \tag{A1}$$

where

$$f_c = 1 + 2(1 - b_{k+1})\rho(Z_{o,k}, E_{k+1}) \frac{\sigma_{Z_{o,k}}}{\sigma_{E_{k+1}}} \tag{A2}$$

Note that, in terms of variance propagation, Eq. (A1) is equivalent to:

$$Z_{o,k+1} = (1 - b_{k+1})Z_{o,k} + b_{k+1}Z_{f,k+1} + \sqrt{f_c}E_{k+1}, f_c \geqslant 0 \tag{A3}$$

where $Z_{o,k}$ and E_{k+1} are assumed to be independent. If f_c is negative, we assume that there is no white noise being added at the current time step, and that all uncertainty at the current time step is propagated from the previous time step. This is achieved by adjusting b_{k+1} in Eq. (A1) by δb_{k+1} so that we have for Eq. (A1):

$$Z_{o,k+1} = (1 - b_{k+1} - \delta b_{k+1})Z_{o,k} + (b_{k+1} - \delta b_{k+1})Z_{f,k+1} \tag{A4}$$

$$\sigma^2_{Z_{o,k+1}} = (1 - b_{k+1} - \delta b_{k+1})^2 \sigma^2_{Z_{o,k}} \tag{A5}$$

The adjustment δb_{k+1} may be calculated by equating the right-hand side (RHS) of Eq. (A5) with the RHS of Eq. (A1), which yields for δb_{k+1}:

$$\delta b_{k+1} = 1 - b_{k+1} \pm \sqrt{(1 - b_{k+1})^2 + f_c \frac{\sigma^2_{E_{k+1}}}{\sigma^2_{Z_{o,k}}}} \tag{A6}$$

Of the two solutions in Eq. (A6), we chose the one that produces a smaller δb_{k+1}. Note that Eqs. (A1)–(A6) are only an approximation in that, while they guarantee the accuracy of $\sigma^2_{Zo,k+1}$, they do not capture the correlation between $Z_{o,k}$ and E_{k+1}. The motivation for the above approximation over other techniques such as state augmentation, is that, for HMOS, accurate modeling of $\sigma^2_{Zo,k+1}$ is considered more important than capturing the correlation between $Z_{o,k}$ and E_{k+1}.

References

Bogner, K., Pappenberger, F., 2011. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. Water Resour. Res. 47, W07524. http://dx.doi.org/10.1029/2010WR009137.

Bogner, K., Pappenberger, F., Cloke, H.L., 2012. Technical note: the normal quantile transformation and its application in a flood forecasting system. Hydrol. Earth Syst. Sci. 16, 1085–1094. http://dx.doi.org/10.5194/hess-16-1085-2012.

Boucher, M.-A., Tremblay, D., Delorme, L., Perreault, L., Anctil, F., 2012. Hydro-economic assessment of hydrological forecasting systems. J. Hydrol. 416–417, 133–144.

Box, G., Jenkins, G., 1976. Time Series Analysis: Forecasting and Control. Holden Day, San Francisco, USA.

Brown, J.D., Seo, D.-J., 2012. Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions. Hydrol. Process.. http://dx.doi.org/10.1002/hyp.9263.

Brown, J., Demargne, J., Seo, D.-J., Liu, Y., 2010. The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ. Model. Softw. 25, 854–872.

Chen, S.-T., Yu, P.-S.h., 2007. Real-time probabilistic forecasting of flood stages. J. Hydrol. 340, 63–77.

Coccia, G., Todini, E., 2011. Recent developments in predictive uncertainty assessment based on the model conditional processor approach. Hydrol. Earth Syst. Sci. 15, 3253–3274. http://dx.doi.org/10.5194/hess-15-3253-2011.

Dale, M., Wicks, J., Mylne, K., Pappenberger, F., Laeger, S., Taylor, S., 2012. Probabilistic flood forecasting and decision making: an innovative risk-based approach. Nat. Hazard.. http://dx.doi.org/10.1007/s11069-012-0483- (ISSN: 1573-0840, Springer Netherlands).

Demargne, J., Wu, L., Regonda, S., Brown, J., Lee, H., He, M., Seo, D.-J., Hartman, R., Fresch, M., Zhu, Y., 2013. The science of NOAA's operational hydrologic ensemble forecast service. Bulletin of American Meteorological Society. http://dx.doi.org/10.1175/BAMS-D-12-00081.1.

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York, 350pp.

Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteorol. 11, 1203–1211.

Gneiting, T., Raftery, A.E., Westveld, A., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133, 1098–1118.

Hantush, M.M., Kalin, L., 2008. Stochastic residual-error analysis for estimating hydrologic model predictive uncertainty. J. Hydrol. Eng. 13, 585–596.

Hashino, T., Bradley, A.A., Schwartz, S.S., 2007. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. Hydrol. Earth Syst. Sci. 11, 939–950. http://dx.doi.org/10.5194/hess-11-939-2007.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast. 15, 559–570.

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley and Sons, p. 240.

Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press London.

Krzysztofowicz, R., 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resour. Res. 35, 2739–2750.

Krzysztofowicz, R., 2001. The case for probabilistic forecasting in hydrology. J. Hydrol. 249, 2–9.

Krzysztofowicz, R., 2002. Bayesian system for probabilistic river stage forecasting. J. Hydrol. 268, 16–40.

McCollor, D., Stull, R., 2008. Hydrometeorological short-range ensemble forecasts in complex terrain. Part II: Economic evaluation. Weather Forecast. 23, 557–574.

Montanari, A., Brath, A., 2004. A stochastic approach for assessing the uncertainty of rainfall–runoff simulations. Water Resour. Res. 40, W01106. http://dx.doi.org/10.1029/2003WR002540.

Montanari, A., Grossi, G., 2008. Estimating the uncertainty of hydrological forecasts: a statistical approach. Water Resour. Res. 44, W00B08. http://dx.doi.org/10.1029/2008WR006897.

Muluye, G.Y., 2011. Implications of medium-range numerical weather model output in hydrologic applications: assessment of skill and economic value. J. Hydrol. 400, 448–464.

National Research Council (NRC), 2004. Confronting the Nation's Water Problems: The Role of Research. National Academies Press, Washington, DC.

National Research Council (NRC), 2006a. Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts. National Academies Press, Washington, DC.

National Research Council (NRC), 2006b. Toward a New Advanced Hydrologic Prediction Service. National Academies Press, Washington, DC.

Pappenberger, F., Beven, K.J., 2006. Ignorance is bliss: or seven reasons not to use uncertainty analysis. Water Resour. Res. 42, W05302. http://dx.doi.org/10.1029/2005WR004820.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1988. Numerical Recipes in C: The Art of Scientific Computing. Cambridge Univ. Press, New York.

Ramos, M.H., van Andel, S.J., Pappenberger, F., 2012. Do probabilistic forecasts lead to better decisions? Hydrol. Earth Syst. Sci. Discuss. 9, 13569–13607. http://dx.doi.org/10.5194/hessd-9-13569-2012.

Reggiani, P., Weerts, A.H., 2008. A Bayesian approach to decision-making under uncertainty: An application to real-time forecasting in the river Rhine. Journal of Hydrology 356, 56–69. http://dx.doi.org/10.1016/j.jhydrol.2008.03.027.

Roulin, E., 2007. Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrol. Earth Syst. Sci. 11, 725–737. http://dx.doi.org/10.5194/hess-11-725-2007.

Schweppe, F.C., 1973. Uncertain Dynamic Systems. Prentice-Hall, Englewood Cliffs, New Jersey, p. 563.

Seo, D.-J., Herr, H.D., Schaake, J.C., 2006. A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. Hydrol. Earth Syst. Sci. Discuss. 3, 1987–2035.

Seo, D.-J., Demargne, J., Wu, L., Liu, Y., Brown, J.D., Regonda, S., Lee, H., 2010. Hydrologic ensemble prediction for risk-based water resources management and hazard mitigation. In: 3rd Federal Interagency Hydrologic Modeling Conference, Las Vegas, NV, Jul.

Smith, P.J., Beven, K.J., Weerts, A.H., Leedal, D., 2012. Adaptive correction of deterministic models to produce accurate probabilistic forecasts. Hydrol. Earth Syst. Sci. 16, 2783–2799. http://dx.doi.org/10.5194/hess-16-2783-2012.

Verkade, J.S., Werner, M.G.F., 2011. Estimating the benefits of single value and probability forecasting for flood warning. Hydrol. Earth Syst. Sci. 15, 3751–3765. http://dx.doi.org/10.5194/hess-15-3751-2011.

Weerts, A.H., Winsemius, H.C., Verkade, J.S., 2011. Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). Hydrol. Earth Syst. Sci. 15, 255–265. http://dx.doi.org/10.5194/hess-15-255-2011.

Wilks D.S. 2011. Statistical Methods in the Atmospheric Sciences, Academic Press, p. 676.

Zhao, L., Duan, Q., Schaake, J., Ye, A., Xia, J., 2011. A hydrologic post-processor for ensemble streamflow predictions. Adv. Geosci. 29, 51–59. http://dx.doi.org/10.5194/adgeo-29-51-2011.