Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management (Proceedings of Symposium HS2004 at IUGG2007, Perugia, July 2007). IAHS Publ. 313, 2007.

Experimental hydrometeorological and hydrological ensemble forecasts and their verification in the US National Weather Service

JULIE DEMARGNE[1,2], LIMIN WU[1,3], DONG-JUN SEO[1,2] & JOHN SCHAAKE[1,4]

1 *Office of Hydrologic Development, National Weather Service, National Oceanic and Atmospheric Administration, 1325 East-West Highway, Silver Spring, Maryland 20910, USA* julie.demargne@noaa.gov

2 *University Corporation for Atmospheric Research, PO Box 3000, Boulder, Colorado 80307, USA*

3 *RS Information Systems, 1651 Old Meadow Road, McLean, Virginia 22102, USA*

4 *Consultant, Annapolis, USA*

Abstract An ensemble preprocessor is being developed by the Office of Hydrologic Development, NOAA/National Weather Service (NWS), USA, to produce reliable short-term hydrometeorological ensemble forecasts from single-value forecasts of precipitation and temperature. These hydrometeorological ensemble forecasts are then ingested into the NWS Ensemble Streamflow Prediction system to produce probabilistic hydrological forecasts that reflect the hydrometeorological uncertainty. The preprocessor methodology attempts to remove biases in single-value forecasts, and capture the skill and uncertainty therein, while preserving the space–time statistical properties of the hydrometeorological variables. The ensemble preprocessor currently operates experimentally at four NOAA/NWS River Forecast Centers in the USA. The verification results presented in this paper show that the precipitation ensembles generated from the ensemble preprocessor produce highly reliable probability estimates and improve the streamflow ensemble forecast performance. Further work is needed to reduce and fully account for hydrological uncertainties in order to improve the quality of streamflow ensemble forecasts.

Key words ensemble forecasting; probabilistic verification; uncertainty

INTRODUCTION

One of the main challenges for hydrological prediction is to quantify the uncertainties associated with a forecast. In order to quantify forecast uncertainty and improve forecast accuracy in an operational setting, ensemble methods, which are essentially a Monte Carlo approach, are currently being implemented by the NOAA/National Weather Service (NWS). There are three primary sources of uncertainty in hydrological forecasting: future hydrometeorological inputs, hydrological initial conditions, and hydrological model(s) (model parameters and structure). To account for input uncertainty, the Ensemble Streamflow Prediction (ESP) component of the National Weather Service River Forecast System (NWSRFS), USA, generates ensembles of hydrological forecasts (streamflow and stage) using historical atmospheric forcing inputs as a guide to future conditions (the climate being considered stationary) (Day, 1985). The initial conditions are estimated from a preliminary run of the hydrological models up to the forecast time. A second run with the historical precipitation and temperature time

series produces an ensemble of hydrological forecasts (e.g. streamflow or stage ensembles) from which probabilistic statements are issued. This component produces long-term ensemble forecasts that reflect the uncertainty in the future forcing inputs as far as the historical climate is representative of future conditions. However, it does not use information about future meteorological conditions obtained from short-term forecasts. These include forecasts produced from the Numerical Weather Prediction (NWP) models by the Hydrometeorological Prediction Center (HPC) of the National Centers for Environmental Prediction (NCEP), and the regional River Forecast Centers (RFCs). In order to assimilate such short-term and typically single-value forecasts of precipitation and temperature, an ensemble processor has been developed to generate ensembles by removing biases in the single-value forecasts, and capturing the skill and uncertainty therein while preserving the space–time statistical properties of the hydrometeorological variables. Since this processor operates before the hydrological models are invoked, it is known as the ensemble preprocessor.

This ensemble preprocessor is being tested at four NOAA/NWS River Forecast Centers (RFCs) in the USA by using the operational deterministic forecasts for lead times of one to five days. This paper presents the verification results obtained so far for precipitation and streamflow for five basins located in Arkansas-Red Basin RFC. For flow forecasts, two sets of forcing input ensembles are ingested into ESP: the climatological ensembles and the ensembles generated by the ensemble preprocessor. Ensemble streamflow forecasts were evaluated using observed flows, as well as the simulated flows produced by ESP using observed precipitation and temperature. The forecast verification study carried out in this work aims at two different goals: first, to compare the performance of flow forecasts generated by ESP with and without the ensemble preprocessor (i.e. preprocessed ensembles *vs* climatological ensembles being ingested by ESP), and second, to separate the effects of input and hydrological uncertainties (the latter coming from both initial conditions and hydrological models).

ENSEMBLE FORECASTING AND VERIFICATION SYSTEMS

Ensemble preprocessor

The operational ESP system that currently operates at the RFCs uses only the historical climatological time series of mean areal precipitation and temperature (as well as potential evaporation for some basins) as forcing inputs. The ensemble preprocessor aims to integrate the skill of the single-value forecasts produced from NWP models by the HPC and then with value added by human forecasters at the RFCs. The procedure constructs, for each hydrometeorological variable, the joint distribution of forecasts and observations from an archive of historical pairs. The joint distribution for a given day in a year is estimated by using data within a window centred on the given day. This data pooling process increases the sample size to estimate the statistical distribution parameters more reliably. The probability distribution function of the future precipitation or temperature that may occur given a particular single-value forecast is the conditional distribution of observed precipitation or temperature given the forecast. To generate ensemble members for each lead time and each location, the historical observations are first sorted and then replaced with the values sampled from

the conditional distribution. The replacement procedure matches the rank of the generated values with that of the historical, e.g. the largest generated value is assigned to the largest historically observed. This procedure, called the Schaake Shuffle (Clark *et al.*, 2004), is applied independently at each lead time and for each forecast location. By rescaling historical values in this way, it creates ensemble forecasts that preserve the space–time statistical properties between any two hydrometeorological variables (e.g. precipitation and temperature).

The ensemble preprocessor described above (see Schaake *et al.* (2006) for further details) has been developed for precipitation and temperature ensemble prediction using the operational deterministic forecasts produced by the RFCs. The availability of RFC forecasts, which are necessary for statistical modelling by the ensemble preprocessor, depends on the RFC archiving process. It is usually limited to the most recent years and only for the first two to five lead days. When no single-value forecast is available, the ensemble preprocessor estimates the climatological distribution from historical observations by using the data pooling process. The Schaake Shuffle technique is then applied to rescale the historical values with the values sampled from the climatological distribution. The resulting ensembles are called resampled climatological ensembles. The ensemble preprocessor blends the ensemble forecasts generated from single-valued forecasts for the first few lead days, with the resampled climatological ensembles beyond that.

Ensemble streamflow prediction

The ESP component of NWSRFS operates with various conceptual hydrological models. First, the initial conditions (defined as the model state variables) are generated by running the hydrological models using an existing set of initial conditions up to the forecast time with observed forcing inputs. Second, ESP produces hydrological ensemble forecasts by ingesting forcing input ensembles into the hydrological models based on the initial conditions obtained from the first phase of modelling. To evaluate the ensemble preprocessor, streamflow ensembles were produced from precipitation and temperature ensembles generated by the ensemble preprocessor. These streamflow ensembles, referred to as QPF-based streamflow ensembles, were then compared to the climatology-based streamflow ensembles.

Hydrologic Ensemble Hindcaster

To evaluate the quality of the ensemble forecasts, forcing input ensembles and hydrological ensembles were verified by using retrospective forecasts, or hindcasts, generated by the Hydrologic Ensemble Hindcaster for an extended time period. The hindcasting process (Franz *et al.*, 2003; Hamill *et al.*, 2004; Welles, 2005; and references therein) is used to compute verification metrics based on a large sample of hydrometeorological and hydrological forecasts. The Hydrologic Ensemble Hindcaster operates in three stages.

First the ensemble preprocessor is run in the hindcasting mode to generate retrospective precipitation and temperature ensembles from archived RFC determin-

istic forecasts for the verification time period. For each date of the verification time period, these precipitation and temperature hindcasts correspond to the ensembles based on the RFC deterministic forecasts for the first few lead days, blended with re-sampled climatological ensembles for longer lead times.

Second, the historical initial conditions for the hydrological models are produced retrospectively from a set of existing initial conditions (for a given date prior to the verification period) and the historical observed precipitation and temperature time series. The initial conditions were generated and stored for each hindcast time in the verification period. These retrospective initial conditions may not correspond exactly to the initial conditions used in real-time, operational, forecasting, which are frequently modified by the forecasters based on their expertise or by some data assimilation techniques. However, the above process supports an assessment of the impact of the forcing input ensembles on the quality of hydrological forecasts without introducing additional complexities in the forecasting process.

Finally, hydrological hindcasts are produced by ESP for each forecast time in the verification period based on the retrospective initial conditions and the precipitation and temperature ensemble hindcasts. In this work, the streamflow hindcasts were generated from two sets of forcing inputs for comparison: the QPF-based precipitation and temperature hindcasts generated by the ensemble preprocessor, and the historical time series of observed precipitation and temperature to reflect the operational ESP. The differences between the resulting two sets of streamflow hindcasts are due solely to differences in precipitation and temperature ensembles. The resulting error in streamflow hindcasts includes both the input error and the hydrological error.

Ensemble verification system

A prototype Ensemble Verification System (EVS) was developed to verify the input and output ensemble hindcasts and compare them with reference forecasts. EVS consists of procedures for pairing the forecasts with the observations, computing various verification metrics to describe different aspects of the forecast quality, and generating graphics. The probabilistic verification statistics were computed by using several percentile threshold values (computed from the observations) that span a wide range of magnitude for the variable of interest. Two reference forecasts were used for comparative evaluation: climatology (defined as monthly climatological mean) and persistence forecast (defined as the current observation at the forecast time).

Ensemble verification carried out in this study aims to compare the performance of the streamflow ensemble forecasts, with and without the ensemble preprocessor, and to analyse how the ensemble forecast performance may vary with lead times and magnitude of the forecast variable. For flow, the uncertainty in the forecast may be decomposed into the input uncertainty (from the meteorological forcings) and the hydrological uncertainty (from initial conditions, model parameters and model structure). In order to separate the effects of these two uncertainties, streamflow ensembles are compared with two reference flows: observed flows to evaluate both the input and hydrological uncertainties, and simulated flows computed from driving the hydrological models with observed precipitation and temperature values to evaluate only the input uncertainty. The comparison of the verification results obtained with

these two reference flows provides some insight into the impact of meteorological and hydrological uncertainties on the overall quality of streamflow forecasts as the basis for targeted improvements of the ensemble forecasting system.

Because the archive of RFC operational deterministic forecasts is quite limited, the sample size of ensemble forecasts for individual test basins is usually too small to produce verification statistics reliably. To reduce sampling uncertainty, the verification statistics were aggregated for a group of basins by weighted-averaging. The weights were determined from the number of events observed in each basin. For verification statistics based on percentiles, the flow thresholds were estimated individually using the same set of percentile values. Further work is being carried out to quantify the sampling uncertainty in the estimated verification statistics.

TEST BASINS AND DATA

The verification study was performed on five basins (with areas ranging from 1129 to 2357 km^2) and located on the Spring River (in Missouri and Oklahoma), the Shoal Creek (in Missouri), and the Elk River (in Missouri) within the Arkansas-Red Basin RFC's service area. The ensemble forecasts of precipitation, temperature and streamflow were produced at a 6-hour time step at 12:00 GMT, to mimic the actual forecasting process, for forecast lead time up to day 14.

For 6-hour mean areal precipitation, the observations come from the NEXRAD rainfall estimates and gauge measurements (Seo & Breidenbach, 2002). The precipitation forecasts are produced by the RFC from the NCEP/HPC Quantitative Precipitation Forecast (QPF) guidance. These QPFs correspond to precipitation amounts expected to fall over the basin for each of the future 6-hour time periods and are ingested by the hydrological models in the deterministic forecasting process. An archive of these QPFs is available for lead day 1 from 1 April 2000 to 12 August 2005 and for lead day 2 from 6 March 2003 to 12 August 2005. The ensemble preprocessor was calibrated by using the archived QPFs and corresponding observations from these two periods, as well as historical time series from 1961 to 1998 (for the Schaake Shuffle). The ensemble preprocessor produced QPF-based ensembles for lead days 1 and 2 and resampled climatological ensembles for lead days 3 to 14 with 38 members (corresponding to the number of years in the historical time series). The forecast verification was carried out for the period of 6 March 2003 to 12 August 2005. This dependent validation allows assessment of the goodness of fit of the ensemble preprocessor, but is not a substitute for independent validation, which is also in progress.

For 6-hour mean areal temperature, the observations are computed from the stations using a distance and elevation weighting process. No deterministic temperature forecasts were used in the period from 6 March 2003 to 12 August 2005 because of a lack of available forecast archive. The ensemble preprocessor produced resampled climatological temperature ensembles with 38 members using the historical time series from 1961 to 1998.

Streamflow measurements at each basin outlet were obtained from the United States Geological Survey (USGS). For these basins, streamflows are simulated with the Sacramento Soil Moisture Accounting model (SAC-SMA) (Burnash, 1995), the

SNOW-17 model for snow ablation (Anderson, 1973), and the Unit Hydrograph for flow routing, all as implemented in NWSRFS (NWS, 2005). Two sets of streamflow ensemble hindcasts (with 38 members) were produced from 6 March 2003 to 12 August 2005: first, QPF-based streamflow ensembles by using QPF-based precipitation and temperature ensembles produced by the ensemble preprocessor, and second, climatology-based streamflow ensembles by using the precipitation and temperature historical time series.

RESULTS AND DISCUSSION

The verification metrics were computed for precipitation and streamflow ensemble forecasts for the following percentile threshold values: 10%, 25%, 50%, 75%, 85%, 90%, 95%, and 97.5%. These eight threshold values cover a wide range of categories to describe the forecast quality from very small events (10%) to large events (97.5%). For precipitation, the amounts corresponding to the 10% percentile are below 0.15 mm (the intermittency of precipitation being defined by 0.254 mm) and below 46.28 mm for the 97.5% percentile for the five test basins. For streamflow, the verification period considered in this study does not contain many events exceeding the flood stage levels: for the five basins, the flood stage levels correspond approximately to the 99.5% percentile. Efforts are under way to extend the sample for verification of extreme events.

Below we present the results for the following metrics: the Brier Score and its decomposition, the Brier Skill Score, the reliability diagram, and the Relative Operating Characteristic (ROC) (Wilks, 1995; Jolliffe & Stephenson, 2003). These metrics are for 24-hour ensemble forecasts, derived from the original 6-hour ensembles, and are computed for each lead day from 1 to 14. The results are aggregated for the five study basins. We present the verification results for the QPF-based precipitation ensembles, the QPF-based streamflow ensembles, and the climatology-based streamflow ensembles, the streamflow ensembles being verified against first the simulated flows and then the observed flows. In this study, simulated flows were generated using slightly different initial conditions than the ones used for ensemble forecasts due to software constraints; it could lead to a small degradation of the verification results, especially for large events.

The Brier Score (BS) measures the mean squared probability error with respect to the given threshold value. It varies from 0 for an entirely correct forecast to 1 for an entirely incorrect forecast. It may be decomposed into three components, namely: BS = Reliability – Resolution + Uncertainty. The Brier Skill Score (BSS) is defined by: $BSS = 1 - (BS_{forecast} / BS_{reference})$. It measures the improvement in the Brier Score of the studied forecast over the reference forecast, which is climatology in this case. A positive BSS indicates that the forecast is better than climatology whereas a negative BSS indicates that the forecast is worse than climatology. BSS values are plotted for the lead days 1 to 14 and for all percentiles; contour lines are overlaid to show the BSS values.

Regarding the BS results for precipitation (plotted in Fig. 1(a) and (b) for the 10th and 85th percentiles respectively), the BS values increase with lead time and decrease with percentiles, due mainly to the variations of the uncertainty values. All the

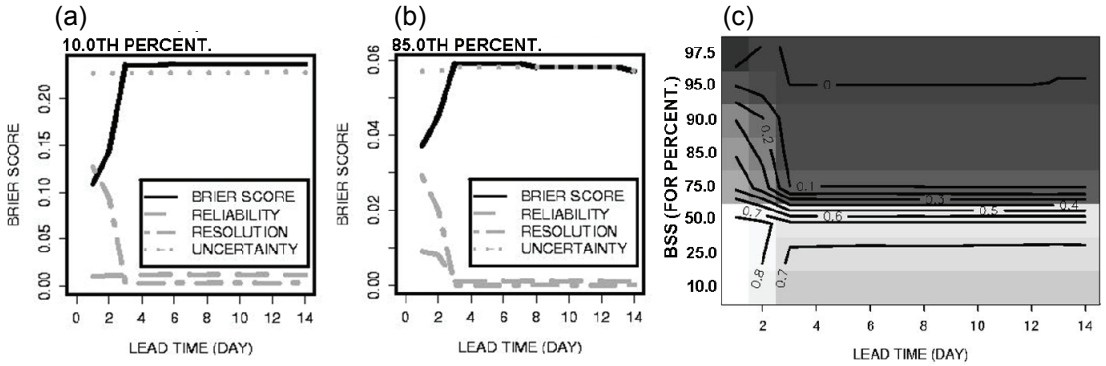

Fig. 1 Brier Score (BS) aggregated statistics for QPF-based precipitation ensembles: (a) BS and its decomposition for the 10th percentile; (b) BS and its decomposition for the 85th percentile; (c) BSS relative to climatology for the eight percentiles.

percentiles show similar patterns with better BS for the QPF-based ensembles (lead days 1 and 2) compared to the resampled climatological ensembles (lead days 3 to 14), especially for the lower percentiles. The better BS values are due essentially to a better resolution of the QPF-based ensembles since the improvement in reliability is smaller. For the BSS results (plotted in Fig. 1(c) for all the percentiles), the QPF-based ensembles perform better than climatology with very high scores up to the 50th percentile; however, the margin of improvement decreases for larger percentiles. Compared to QPF-based ensembles, the resampled climatological ensembles show lower BSS values with similar patterns. The BSS values for large precipitation events (95th and 97.5th percentiles) are close to zero showing that the relative skill of the QPF-based precipitation ensembles is smaller for these events.

For streamflow, the BS for the QPF-based streamflow ensembles is plotted for the 10th percentile in Fig. 2(a) and for the 85th percentile in Fig. 2(b). The BS values increase with lead times with a larger degradation for the 25th to 85th percentiles. The BS and its reliability and resolution components vary similarly between the QPF- and climatology-based streamflow ensembles, with slightly better results for the former above the 25th percentile. The BSS is plotted for all percentiles for the QPF-based streamflow ensembles in Fig. 2(c) and for the climatology-based ensembles in Fig. 2(d). The QPF-based streamflow ensembles have a lower BSS for the 10th and 25th percentiles and higher BSS for larger percentiles, especially around lead days 3 and 4. For example, the performance gain in terms of BSS is about 1 day for the 75th percentile and more than 2 days for the 95th and 97.5th percentiles.

Comparison of the BS of streamflow ensembles verified against observed and simulated flows shows the large adverse effects of hydrological uncertainty, especially for low flows. For example, at all lead days, the BSS based on observed flows is only positive above the 50th percentile. This underlines the need to reduce and accurately account for hydrological uncertainty.

To evaluate the reliability of the ensemble forecasts, the reliability diagram, which measures the agreement between the forecast probability and the mean observed frequency (strictly it plots the agreement between the predicted and observed frequency, since the former is based on Monte Carlo), is produced for each threshold and is plotted for the 85th percentile in Fig. 3 (left). The deviation from the diagonal indicates

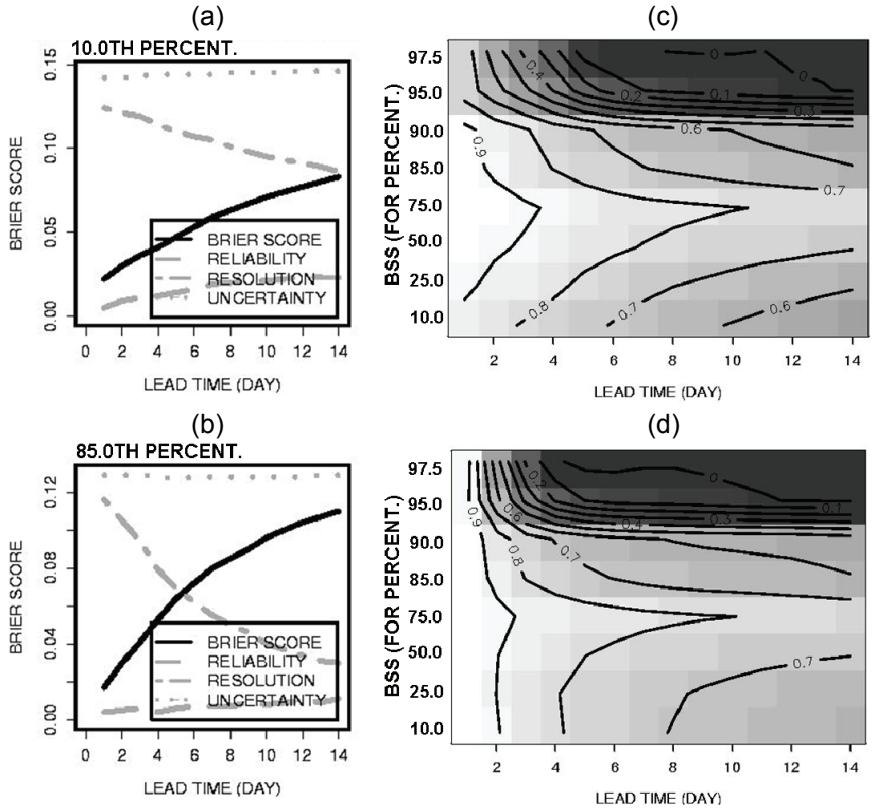

Fig. 2 Brier Score (BS) aggregated results for streamflow ensembles verified against simulated flows: (a) BS statistics for the 10th percentile for QPF-based ensembles; (b) BS statistics for the 85th percentile for QPF-based ensembles; (c) BSS for all percentiles for QPF-based ensembles; (d) BSS for all percentiles for climatology-based ensembles.

conditional bias in forecast probability. The range of forecast probabilities is divided into five bins, the first bin corresponding to the zero-probability event and the last bin to the event with a probability of one. The histogram gives the frequency of forecast probability for the first lead day. It represents the sharpness of the forecast and is expected to be U- or L-shaped for sharp forecasts (the forecast probability being more frequently assigned to the extreme categories). Also the histogram occasionally shows very small sample sizes in bins, which results in erratic estimates of reliability.

For precipitation (Fig. 3(a) left for the 85th percentile), the reliability diagram shows that the QPF-based ensembles for the first two lead days are very reliable up to the 50th percentile and then start to over-forecast for the larger percentiles. The conditional bias is larger for lead day 2 than lead day 1 for the 75th and 85th percentiles. The resampled climatology ensembles have no resolution (with a horizontal line for the first two bins in the reliability diagram) since they are based on climatology.

To assess the impact of input error on reliability of streamflow ensembles, the reliability diagram was computed for the QPF- and climatology-based ensembles and is shown in Fig. 3(b) (left) and (c) (left) respectively, for the 85th percentile. The QPF-based streamflow ensembles are slightly less reliable than the climatology-based ensembles for the very low flows, especially for the first lead day, whereas the QPF-based ensembles perform better for the other percentiles. For the 25th to 75th

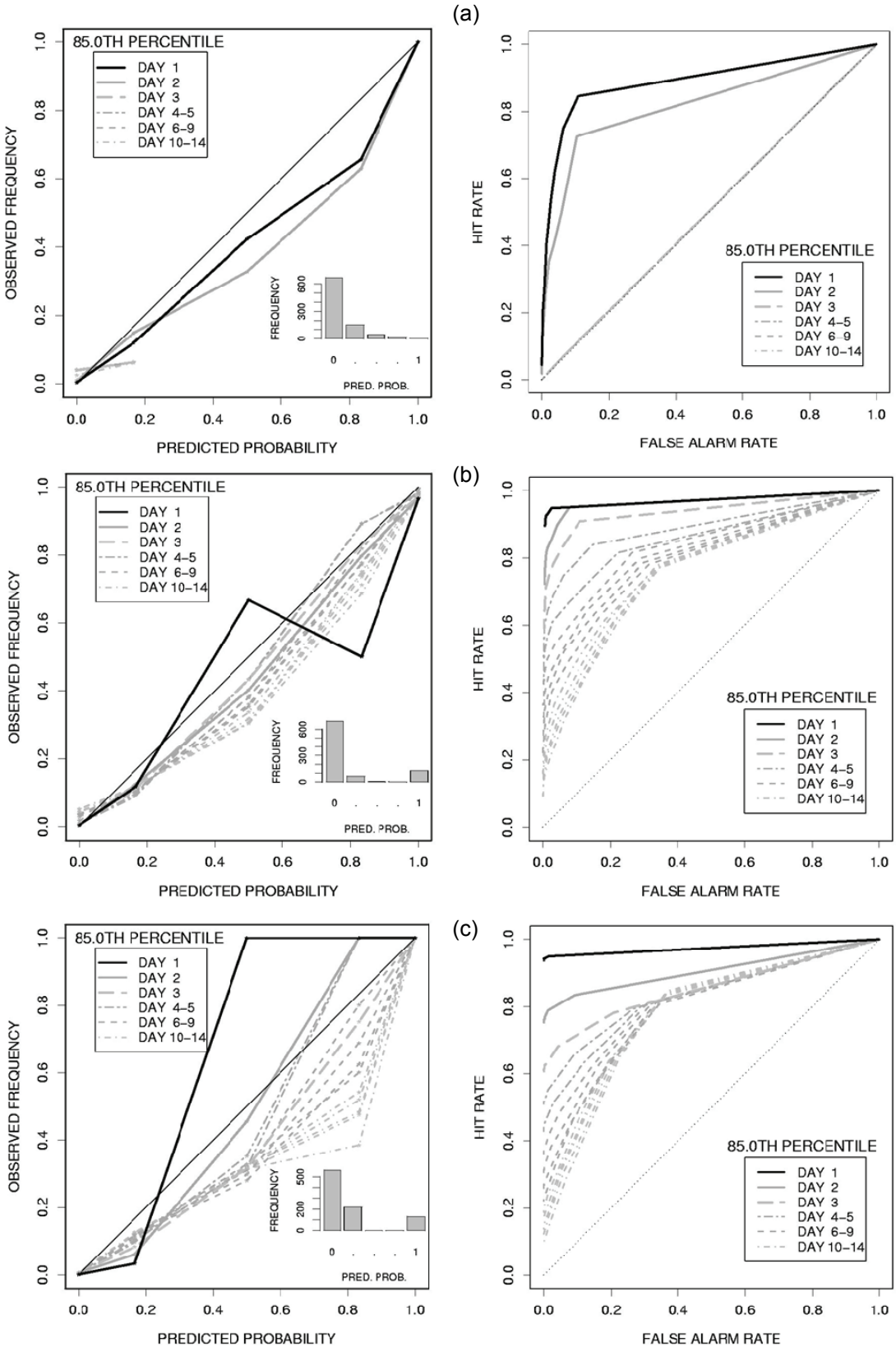

Fig. 3 Reliability diagrams (left) and ROC diagrams (right) for the 85th percentile obtained for the five test basins for three sets of ensembles: (a) QPF-based precipitation; (b) QPF-based streamflow; (c) climatology-based streamflow.

percentiles the QPF-based ensembles are very reliable for all lead days; for the 85th and 90th percentiles, they tend to over-forecast. Also the sharpness is better for the QPF-based ensembles than the climatology-based ensembles.

To assess the combined impact of input and hydrological uncertainties on the reliability of the streamflow ensembles, the reliability diagram was also computed for the QPF-based ensembles using the observed streamflow. When both input and hydrological uncertainties are considered, reliability of streamflow ensembles is significantly reduced: there is a large under-forecast bias from the 10th to the 50th percentiles (i.e., in the low flow regime) and a more pronounced over-forecast bias for the 85th and 90th percentiles for longer lead times.

The forecast resolution was studied with the Relative Operating Characteristic (ROC), which measures the ability of the forecast to discriminate between events and non-events. ROC diagrams were generated with ten probability thresholds used to make the yes/no decision and are plotted for the 85th percentile in Fig. 3 (right). The ROC diagram consists in plotting the Hit Rate against the False Alarm Rate. The Hit Rate (HR) measures the fraction of events that were correctly forecast to occur and is defined by: HR = hits/(hits + misses). The False Alarm Rate measures the fraction of "yes" forecasts that were incorrect and is defined by: FAR = false alarms/(false alarms + correct rejections). The diagonal line corresponds to a forecast with no skill; a forecast with a good resolution has its points near the upper left corner, the perfect forecast corresponding to HR = 1 and FAR = 0.

For precipitation (Fig. 3(a) right for the 85th percentile), the QPF-based ensembles show good resolution with similar scores up to the 85th percentile and decreasing performance for the larger percentiles. When assessing the impact of input uncertainty on resolution of streamflow ensembles, the QPF-based streamflow ensembles (Fig. 3(b) right for the 85th percentile) show a slightly better resolution than the QPF-based precipitation ensembles. For the shorter lead times there is a slowly decreasing skill when the threshold is increased, whereas the decrease is more significant for longer lead times, especially above the 75th percentile. The ensembles for lead times longer than 5 days show little resolution for the 97th percentile. The QPF-based streamflow ensembles have more resolution than the climatology-based ones (Fig. 3(c) right for the 85th percentile), especially from the 75th to the 97.5th percentiles, with a performance gain of 1 to 3 days for shorter lead time.

With both input and hydrological uncertainty considered, the resolution of streamflow ensembles significantly decreases as expected. Up to the 75th percentile, the forecast resolution with both uncertainties considered for lead day 1 is worse than that with only the input uncertainty for lead day 14. For the larger percentiles, the performance loss is equivalent to 4 to 6 days for shorter lead time.

These verification results show that the QPF-based precipitation ensembles produced by the ensemble preprocessor for the first lead days perform better than climatology for all precipitation amounts, with very good reliability and resolution up to the 85th percentile. This may be improved by using a longer archive of QPF forecasts to better calibrate the ensemble preprocessor, especially for rare events. The streamflow ensembles generated from the ensemble preprocessor outputs perform better than the climatology-based ensembles, except for the very low flows. One of the reasons is that the hydrological models are generally calibrated to perform best for

high flows for flood forecasting purposes. Precipitation, temperature and streamflow ensemble forecasts are being verified for other test basins and using additional single-value forecasts; additional verification results will be reported in the near future. The results also show that the hydrological uncertainty has a large negative impact on the streamflow forecast performance. Further work is also under way to reduce and fully account for hydrological uncertainties with data assimilation and postprocessing toward improving the quality of streamflow ensemble forecasts.

Acknowledgements The support of this work by the Advanced Hydrologic Prediction Service (AHPS) and Climate Prediction Program for the Americas (CPPA) programmes of the National Oceanic and Atmospheric Administration (NOAA) is gratefully acknowledged. The authors would like to thank Bill Lawrence and Gregory Stanley of the NWS Arkansas-Red Basin River Forecast Center (ABRFC), Tulsa, Oklahoma, for providing the hydrometeorological and hydrological data used in this work.

REFERENCES

Anderson, E. A. (1973) National Weather Service River Forecast System – Snow accumulation and ablation model. *NOAA Technical Memorandum NWS HYDRO-17*. US Dept of Commerce, Silver Spring, Maryland, USA.

Burnash, R. J. C. (1995) The NWS river forecast system – catchment modeling. In: *Computer Models of Watershed Hydrology*, (ed. by V. P. Singh), 311–366. Water Resources Publications, Littleton, Colorado, USA.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. & Wilby, R. (2004) The Schaake Shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydromet.* **5**, 243–262.

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

Franz, K. J., Hartmann, H. C., Sorooshian, S. & Bales, R. (2003) Verification of National Weather Service Ensemble Predictions for water supply forecasting in the Colorado River Basin. *J. Hydromet.* **4**, 1105–1118.

Hamill, T., Whitaker, J. S. & Wei, X. (2004) Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Weath. Rev.* **132**, 14341447.

Jolliffe, I. T. & Stephenson, D. B. (2003) *Forecast Verification, A Practitioner's Guide in Atmospheric Science*. Wiley & Sons, Hoboken, New Jersey, USA.

NWS (2005) National Weather Service River Forecast System (NWSRFS) User Manual Documentation. *National Weather Service documentation*, Silver Spring, Maryland, USA.

Schaake, J., Demargne, J., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X. & Seo, D. J. (2006) Precipitation and temperature ensemble forecasts from single-value forecasts. *HESS*, Special Issue "Hydrological Prediction Uncertainty" (accepted for publication at http://www.hydrol-earth-syst-sci.net/special_issue75.html).

Seo, D. J. & Breidenbach, J. P. (2002) Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. *J. Hydromet.* **3**, 93–111.

Welles, E. (2005) Verification of river stage forecasts. PhD Thesis, University of Arizona, Tucson, Arizona, USA.

Wilks, D. S. (1995) *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, San Diego, California, USA.